Understanding, Building, and Evaluating Models for Context Aware Conditional Natural Language Generation

David Chan

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-15
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-15.html

April 16, 2024

Understanding, Building, and Evaluating Models for Context Aware Conditional Natural Language Generation

by

David McCloud Chan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Canny, Chair
Professor Trevor Darrell
Professor Alison Gopnik

Spring 2024

Understanding, Building, and Evaluating Models for Context Aware Conditional Natural Language Generation

Abstract

Understanding, Building, and Evaluating Models for Context Aware Conditional Natural Language Generation

by

David McCloud Chan

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor John Canny, Chair

If you ask a human to describe an image, they might do so in a thousand different ways. Each of these descriptions depends not only on the image but also on a rich tapestry of contextual hints and clues surrounding the image (up to and including the person doing the describing themselves). Until now, the field of conditional natural language generation has focused almost solely on the perception component of the task: how do we perceive what is in the stimulus – be it audio, visual, or textual – and relay it to the user? In this dissertation, we argue that models that focus solely on the stimulus (and not the associated context) suffer significant shortcomings in their ability to generate language that aligns well with human judgments of quality and content while decreasing their overall utility for downstream tasks. This dissertation focuses on three core objectives in the pursuit of building a context-aware conditional natural language generation (CNLG) model: (1) capturing and understanding the information *within, among, and between* generated conditional texts, (2) developing multimodal models that better integrate contextual information, and (3) designing CNLG evaluation methodologies that better align with human judgment. Through these objectives, we demonstrate the power of context in natural language generation and help to answer the question: "How can we understand, build, and evaluate context-aware models for conditional natural language generation?"

Contents

List of Figures

List of Tables

Acknowledgments

Alone we can do so little, together we can do so much.

Helen Keller

Looking back on the years that I have spent working on my Ph.D., it is hard to imagine where I would be without the people who have guided and supported me along the way. Without them, I would not be the person or researcher I am today, and I would like to take this opportunity to express my deepest appreciation to those who helped me along this journey.

First, I am amazingly grateful to my research advisor, Professor John Canny, who turned me from a student interested in research into a scientist. I cannot thank him enough for his willingness to believe in me, and his endless support of my research and academic career. Throughout my Ph.D. he has always been there to offer guidance, turn my half-baked ideas into fully baked ideas, and tell me when I was completely and utterly wrong. He has taught me a systematic and methodical approach to research, programming, and problem-solving that I will keep with me for the rest of my life.

I am also deeply thankful to my second advisor Professor Trevor Darrell. Trevor has always been there to be excited about the ideas that I bring to him, and to be my champion when building connections to industry and beyond. In much the way John taught me to be a scientist, Trevor has shown me how to be a Professor, and shown me the ins and outs of running a research lab, obtaining grant funding, and building a community around a research idea.

My experience at Berkeley would not have been what it was without the many professors and collaborators I've worked with over the years. Thank you to Professor Allison Gopnik, Professor Avideh Zakhor (who along with John and Trevor formed the core of my thesis committee) and Professor Joseph E Gonzalez, Professor Gopala Anumanchipalli, Professor Adrian Aguilera, and Professor Anca Dragan, for your support and advice in research, and willingness to listen to some truly crazy research ideas. Thank you as well to my co-authors, collaborators, and friends at UC Berkeley, Tsung-Han Wu, Giscard Biamby, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Suzanne Petryk, Kehan Wang, Jasmine Collins, Deepak Pathak, Pulkit Agrawal, Colorado Reed, Eliza Kosoy, Andrew Head, Nate Weinman, Caroline Figueroa, Tiffany Luo, Andrea Jacobo, Alan Munoz and Minx Manuel. Every one of you was instrumental in the research that we did together, and I cannot imagine what my Ph.D. would have been like without the conversations that we had, and the

Over the course of my Ph.D. I have had the opportunity to work closely with some amazing researchers at both Google and Amazon through the BAIR Commons program. At Google, I'd like to thank David Ross who gave me (an un-proven Ph.D. student) the opportunity to work on some amazingly interesting and challenging problems, Austin Myers,

who was always there to provide research insights and thoughtful directions to take our work, Sudheendra Vijayanarasimhan, who was always willing to lend a hand in reviewing rebuttals and reading my (all too awful) paper drafts, and Bryan Seybold who helped to place all of our work in perspective.

At Amazon, I'm especially grateful to my mentor and manager Shalini Ghosh, who has always stood by me through the time that I spent at Amazon. She has been an amazing friend, an avid supporter of my work, a reliable research collaborator, and a valuable conversational partner in helping me to discuss novel ideas and guide my projects to success. I'm also thankful to my collaborators at Amazon, Björn Hoffmeister, and Ariya Rastrow, for being there to validate our ideas, and secure project resources, and Hitesh Tulsiani, Debmalya Chakrabarty, Aparna Khare, Pranav Dheeram, Chao-Han Huck Yang, Jasha Droppo, Andreas Stolcke, Yi Gu, Grant Strimel, Philip Mayer, Venkatesh Ravichandran, and Phani Nidadavolu for being there as friends, people to discuss ideas with, and supporters during my time at Amazon.

I have also been lucky to work with some amazing folks in other parts of the industry over the years. First, at NASA JPL, thank you to Ali Agha-Mohammadi who showed me how to manage a research lab, and balance a seemingly infinite number of projects simultaneously. At Dropbox, thank you to Chris Lesniewski and Thomas Berg who taught me what it means to integrate machine learning and artificial intelligence into real-world production pipelines, and Daniel Reiter Horn, who taught me that programming is about being a hacker and that with a "it's possible" attitude, even the impossible can be possible. At Deepmind, I'd like to thank Jessica Hamrick, Sandy Huang, and Nan Rosemary Ke for showing me a greater world of reinforcement learning, and always being there to talk about the weird things that children and agents seem to do the same.

To the members of John's lab (the CannyLab?) who I have had the pleasure to work with over the last few years, Forrest Huang, Roshan Rao, Biye Jiang, Daniel Seita, Philippe Laban, Suhong Moon, Marwa Abdulhai, Hellina Hailu Nigatu, Xinlei Pan, and Jinkyu Kim, you have always helped me to gain new perspectives on the field of AI research and helped to shape the ways I think about AI. To all of my mentees during the Ph.D., Yash Jain, Anirudh Sundar, Vladislav Lialin, Dipika Khullar, Kevin Cai, Chonghua Liu, Adrian Liu, Dhruv Jhamb, Andy Fang, Illian Herzi, Aatif Jiwani, Shubhrakanti Ganguly, Bofan Xue, Yiming Ni, Oliver Bryniarski, Austin Kao, Karen Lu, and Tianrui Chen, thanks for putting up with me as an advisor, and I can only hope that you learned as much working with me that I learned working with you.

This Ph.D. would not have been possible without the support of BAIR and EECS, and I am deeply thankful to the administrators in BAIR and EECS, Angie Abbatecola, Ami Karagiri, Roxana Infante, Jean Nguyen, Shirley Salanio, and Patrick Hernan, for keeping us all on track, providing the spaces we need to work, and making my (and so many others) Ph.D. possible.

I'd also like to thank those who turned me into the researcher I am today. Thank you to Eric Benson for inspiring my love of science and learning, Charlie Jackson for teaching me my first programming language, Stephen Lacks for introducing me to statistics, Randy

Shepler for believing in and supporting my passion for technology, Mohammad Mahoor and Nathan Sturtevant for teaching me how to do research, and believing in me when I said that I could write a paper as an undergraduate, and Mario Lopez for showing me that math does indeed have an answer for everything.

Beyond those who helped guide my way in an academic sense, my Ph.D. would not have been possible without the support of my friends and family. Zach, thanks for being there for the last twenty years, it's been amazing to share this journey with you. Roshan, Chandan, and Forrest, I can't imagine my Ph.D. experience without those mid-week snack shack runs, and the times that we spent together. Rasmus, thanks for putting up with those crazy ideas, and all the kvetching I needed to do as I finished out my Ph.D. To David and Bofan thanks for keeping me sane over the years by hanging out with me in whatever free time I had.

Finally, to my family, Stephanie, Andrew, Paul, and Alesia. Without your unwavering support as I've pursued my dreams this whole thing would not have been possible. You have been with me through the ups and downs of this journey, and this dissertation is as much a product of your support, love, and championship as it is my own effort.

Chapter 1

Introduction

In late February 2015, an image of a dress was posted on the internet. This, in and of itself is not surprising; according to some estimates, over three billion images are posted to the internet every day (Thomson et al., 2022). But what made this particular image special? Why did this image inspire 840,000 views and 11,000 tweets per minute, while helping to set a new record for concurrent users at BuzzFeed (673,000)? Well, let me ask you a simple question: "What color is the dress?" If you said blue and black, well, you would agree with me (and according to Lafer-Sousa et al. (2015), 57% of people). But if you didn't, that's not surprising either - you might have seen white and gold, or blue and brown, or agree with the 2% of people who said something else entirely. This disagreement was one of the largest questions the internet faced in 2015, all caused by a surprising phenomenon: every person sees color differently. While the exact reason is still debated, according to Wallisch (2017), the potential difference can be ascribed to assumptions that each person viewing the image makes about the illumination of the image; people who assumed the illumination was in natural light were more likely to see the dress as white and gold, while those who assumed the illumination was artificial were more likely to describe it as blue and black.

Figure 1.1: A picture of a dress. Photo by Cecilia Bleasdale.

So, if we were to ask an artificial agent to describe the image, what would it say? Asking a state-of-the-art image description model (Li et al., 2023b) gives the result "A blue and black dress on a mannequin." This description is great for 57% percent of the people who naturally agree with the system, but for the other 43%, the description model fails to generate an answer that they would agree with (is it a coincidence that the model agrees with the majority? See chapter 6). This result belies a greater problem in image description, and in fact, in conditional language generation as a whole: traditional approaches for conditional natural language generation largely ignore the greater contexts that they lie within. Indeed, it is clear that language never exists in a vacuum: surrounding each generated utterance is a rich tapestry of contextual clues, hints, and conditions that impact the final result for the generative model.

Figure 1.2: A quaint European street. Photo by Zhang Kaiyv.

One might argue that this is only an issue with subjective images, such as the image of the dress above, but that's hardly the case. Take for instance the image in Figure 1.2. If the image here is used in an article about a public bicycle hire scheme named Hire-a-bike, then the bike is the focus, and a good caption/description of the image could be "A woman rides a Hire-a-bike along a city road". However, if it's used in an article about a dispute between the café and the restaurant, a better description might be "The storefronts of the 'Café Bar Hotel' and 'Alpen Hotel Restaurant"'[1]

Until now, the field of conditional natural language generation has focused almost solely on the perception component: how do we perceive what is in the stimulus (be it audio, visual, or textual), and relay that to the user? In the case of image description, this has meant a focus on understanding what is in the image, and largely ignoring any contextual clues. In the case of automatic speech recognition, this means focusing on the audio itself and ignoring the context in which that audio occurs. In many cases, however, such a context is either helpful or necessary, for the output of the model. Thus, to address these challenges, we must pivot towards a more nuanced understanding of conditional natural language generation by recognizing that effective communication and information exchange hinge not solely on the literal interpretation of the stimulus but additionally on an intricate web of contextual cues and environmental factors.

In this dissertation, we ask the overall question: *How can we understand, build, and*

[1]Thanks to Jake Archibald, and his amazing blog post on alt-text, for this example: `https://jakearchibald.com/2021/great-alt-text/`.

evaluate context-aware models for conditional natural language generation? To explore this problem we delve into several domains (see section 1.1), to explore ways in which the context surrounding a piece of text can influence its generation, and how we can leverage contextual clues (from some surprising sources) to understand, evaluate, and build more powerful multimodal models. Overall, this dissertation is broadly organized into three core sections, each treating some aspect of the context-aware conditional natural language generation (CNLG) problem:

- **Understanding information within, among, and between generated samples**: First, we dive deeply into understanding joint distributions of images/video and text, and information that can be captured within the generated texts (i.e. the distribution of language in a dataset) and among/between the generated texts (among, referring to the the information present in multiple samples for a single image, and between, referring to the larger distribution of language that can be imputed from the sample set). In chapter 3, we look at the behavior of image/text joint distributions and uncover interesting details about how the datasets we use for images and videos are structured, and in chapter 4, we explore how such implicit characteristics in the dataset can be used to select a small/efficient set of samples during training.
- **Building multimodal models for CNLG**: Next, we introduce several methods for building models for CNLG across several domains. In the image captioning domain, we discuss in chapter 6 how the full distribution learned by CNLG models can be leveraged effectively to produce single high-quality captions. In the ASR domain, we explore how different types of context including videos (chapter 7), text catalogs (chapter 8) and dialogues (chapter 9) can be leveraged to improve the quality of generated natural language.
- **Evaluating CNLG models**: Finally, we introduce two new methodologies for evaluating models that are capable of CNLG. The first, introduced in chapter 11, evaluates models by looking at the full learned distribution, instead of only a single best sample for the model. The second, introduced in chapter 12, leverages the implicit distribution of human preference learned by large language models to improve the evaluation of generated text.

1.1 Application Domains

Conditional natural language generation is a broad field, encompassing many different domains and applications. In this dissertation, we focus on two application domains: Image/Video description and Automatic Speech Recognition. Here, we broadly define the task domains.

1.1.1 Visual (Image/Video) Description

Visual description is the task of translating an input image/video into a natural language description of that image/video. Here, the key stimulus is the image/video, and the resulting generated text is a description of what is present in the video. Such a task has several key applications including:

(1) Alt-Text Generation / Accessibility: One of the most important applications of visual description is the automated generation of alt-text (Yoon et al., 2019). "alt-text", short for "alternative text" is a short, concise description of the input visual media designed to replace the media for people who cannot see the image (either because they are visually impaired or using a screen reader, or because the image does not load on a webpage).

(2) Summarization: Another exciting application of visual description is video and image summarization (Zhang et al., 2016). Short natural language summaries of what is happening in a visual scene may be more time-efficient for a user to consume, rather than watching the full video.

(3) Indexing: Following summarization, another application of visual descriptions is for indexing and search. While methods do exist for text-image search/recall (Radford et al., 2021a; Miech et al., 2019), many search engines are optimized for performing text-text queries, a process which is enabled for images by hand-written metadata or image descriptions. Instead of relying on metadata, images/videos can alternatively be encoded using a visual description, which serves as a proxy during the search process.

An example of a good visual description

Figure 1.3: Helen Keller stands in the right two-thirds of a vertical frame. She is wearing a loose, satin-like white robe with a dusky gray border that wraps around her neck and downward. Her dark brown hair is parted in the center and gathered neatly, billowing in tight curls above her ears. She stares intently at the camera, her thin lips flat, expressionless. Her hands come together at her waist as she gently holds a white rose whose long stem starts out of frame, proceeds upwards as deep green leaves spread forth, and ends in a trio of white roses by her sternum. In the foreground, is a white rose bush occupying the left third of the frame, deep green leaves below, and delicate white flowers above. A gradient grey background completes the intense, solemn portrait. Photo by alperyesiltas.

is shown in Figure 1.3. Here, the visual description goes into detail about the image, focusing on the foreground, and background, and a detailed discussion of the feeling and emotion conveyed in the work. Unfortunately, such "good" descriptions are relatively uncommon. Examples of more common descriptions, along with some of the nuances of existing image description tasks, are discussed in chapter 3, where we delve into how the actual data distribution impacts image/text models.

In this work, we use the terms "description" and "captioning" interchangeably, however, it is worth noting that there is a fine-grained distinction between the two in practice. Visual "description" is the process of describing what's in the image/video for somebody who can't see the image (i.e. as a natural language substitute for the image). Image "captions" on the the other hand may not closely mirror what the image shows (for example, the caption might include the date the photo was taken, who took the photo, where the photo was taken, etc.).

1.1.2 Automatic Speech Recognition (ASR)

The second domain that we explore in detail in this dissertation is automatic speech recognition (ASR). This is the task of taking a spoken utterance and converting that utterance into a natural language rendering of that utterance. There are several common applications for ASR tools including:

1. **Live captioning and transcription:** ASR is used to automatically transcribe audio and video content into text, allowing for the creation of subtitles, meeting minutes, transcriptions of lectures/interviews, and other applications.

2. **Virtual assistants, chatbots, and smart speakers:** Devices such as the Amazon Echo, Google Home, and Apple's Siri use ASR tooling to interpret voice commands and provide responses, control smart devices, and more. In this dissertation, many of our experiments are performed in the context of virtual assistant dialogues (with Amazon's Alexa assistant, see chapters 7,8, and 9).

3. **Voice commands, dictation, and accessibility:** Many ASR tools are used for assisting users with disabilities by converting speech into text or commands enabling easier/hands-free interaction with devices and software.

Beyond such applications, ASR tools can be used in domains such as automotive systems, healthcare, language learning, customer service/call centers, security and authentication, journalism, and law (among others).

While ASR applications are conditional natural language generation, it may not be immediately obvious that they suffer from the same contextual problems that image description suffers from: the language that is spoken is not dependent on each person, is it? It turns out that in many cases, particularly in other languages, the transcription of an utterance heavily depends on contextual hints. Those hints may come from other parts of the conversation,

or they may have to be inferred from the visual clues. One example is in Japanese, where the utterance あついい (atsui) could be transcribed as 熱い, 厚い, 篤い, 暑い, depending on if the user meant that something was hot (in a temperature sense), thick (as in a thick book), warm (in terms of hospitality), or hot (for the weather). We explore these ideas more in part II, where we explore ways to integrate context clues into ASR models, drawing context from several different sources.

Chapter 2

Statement of Prior Publications and Authorship

This dissertation is composed of several projects, each of which contains core ideas and research previously published at various venues, listed in Table 2.1. I am the first author in these papers, but this research would not have been possible without the assistance of my collaborators, including David Ross, Austin Myers, Sudheendra Vijayanarasimhan and Bryan Seybold at Google, Shalini Ghosh, Hitesh Tulsiani, Debmalya Chakrabarty, Ariya Rastrow and Björn Hoffmeister at Amazon, and Suzanne Petryk, Yiming Ni, Joseph E Gonzalez, Trevor Darrell and John Canny at UC Berkeley. In chapters 3, 4, 6, 7, 8, 9, 11, and 12, I reflect their support by using 'we'. While all of this work is my own, this dissertation makes no novel claims to the intellectual property contained within.

Table 2.1: List of prior work included in this dissertation, along with publication venues, coauthors, and in which chapter they appear in this dissertation.

Title	**Author List**	**Venue**	**Chapter**
Task Oriented Dialogue as a Catalyst for Self-Supervised Automatic Speech Recognition	David M Chan, Shalini Ghosh, Hitesh Tulsiani, Ariya Rastrow, Björn Hoffmeister	2024 IEEE International Conference on Acoustics, Speech and Signal Processing	chapter 9
CLAIR: Evaluating image captions with large language models	David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, John Canny	Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing	chapter 12
IC3: Image Captioning by Committee Consensus	David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, John Canny	Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing	chapter 6
Domain Adaptation with External Off-Policy Acoustic Catalogs for Scalable Contextual End-to-End Automated Speech Recognition	David M Chan, Shalini Ghosh, Ariya Rastrow, Björn Hoffmeister	2023 IEEE International Conference on Acoustics, Speech and Signal Processing	chapter 8
Distribution Aware Metrics for Conditional Natural Language Generation	David M Chan, Yiming Ni, David A Ross, Sudheendra Vijayanarasimhan, Austin Myers, John Canny	Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation	chapter 11
Multi-modal pre-training for automated speech recognition	David M Chan, Shalini Ghosh, Debmalya Chakrabarty, Björn Hoffmeister	2022 IEEE International Conference on Acoustics, Speech and Signal Processing	chapter 7
What's in a Caption? Dataset-Specific Linguistic Diversity and Its Effect on Visual Description Models and Metrics	David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, Bryan Seybold, John F Canny	Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)	chapter 3
Active learning for video description with cluster-regularized ensemble ranking	David M Chan, Sudheendra Vijayanarasimhan, David A Ross, John F Canny	Proceedings of the Asian Conference on Computer Vision (2020)	chapter 4

Part I

Understanding

> On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.
>
> *Charles Babbage*

When building models for conditional generation, it is important to understand the data upon which the model is built. Unfortunately, it has been shown repeatedly that models are willing to exploit patterns that are implicit in the data to achieve strong metric performance on benchmark tasks, without actually solving the target problem (for a good discussion of this, see Skalse et al. (2022)). Indeed, models trained with one dataset often fail in a new dataset or scenario. Thus, it makes sense to begin our investigation by looking closely at the data itself: what patterns are present in the data? Can we take lessons from these patterns during our modeling process? Can we exploit or correct those patterns efficiently and effectively?

In this section of the dissertation, we focus on several aspects of the data distribution unique to conditional natural language generation: the distribution of information *within* the generated texts, and the distribution of information *among and between* the generated texts. Information *within* the generated texts is what we consider our standard language distribution: looking at all of the generated texts in a dataset, how do they behave regardless of the implicit partitioning ascribed by the stimulus. More interesting (and less explored), however, is the information among samples generated given a single condition. Multiple samples given the same image can capture further information than a single sample on it's own. Even further, it is interesting to understand the information found "between" these samples. Given that we have a sample set, we can impute a larger conditional distribution of text given the image – what can we learn from the imputed distribution? and how can we exploit this distribution to build better models?

To explore these directions, in chapter 3 we look at how linguistic diversity present in the dataset itself can impact models for conditional natural language generation in the image and video description domains. We examine several popular visual description datasets, and capture, analyze, and understand the dataset-specific linguistic patterns that models exploit to achieve strong performance. For example, at the token level, sample level, and dataset level, we find that caption diversity is a major driving factor behind the generation of generic and uninformative image and video descriptions. We further show that state-of-the-art models even outperform held-out ground truth captions on modern metrics and that this effect is an

artifact of linguistic diversity in datasets. Understanding that this linguistic diversity is key to building strong captioning models, we recommend several methods and approaches for maintaining diversity in the collection of new data and dealing with the consequences of limited diversity when using current models and metrics.

Next, in chapter 4, we turn our attention to the collection of data for the video captioning task: how can we exploit the inherent structure in the visual-linguistic data to reduce the amount of training data we need to collect. Here, we explore various active learning approaches for automatic video captioning and show that a novel method based on cluster-regularized ensembling of models provides the best active learning approach to efficiently gather training sets for video captioning. We evaluate our approaches on the MSR-VTT and LSMDC datasets using both transformer and LSTM-based captioning models and show that our novel strategy can achieve high performance while using up to 60% fewer training data than the strong state-of-the-art baselines.

The lessons that we learn in this section are crucial for understanding the failure cases of our models, and to understanding how we can build models that are robust to the artificial domain shift introduced by a shifting underlying dataset. We can see the impact of the work from this chapter in later sections, perhaps most notably chapter 6, where we directly exploit the lessons learned in chapter 3 to produce high-quality caption models without additional training, and chapter 11, where we develop improved measures to test the deficiencies found in this chapter. Ultimately, it is necessary to focus first on the data, as without a strong grasp of the choices made in the underlying distributions, it is impossible to design models that effectively and efficiently transfer to interactive and intelligent systems.

Previously Published Works Appearing In This Section:

1. Chan, David M., et al. "What's in a Caption? Dataset-Specific Linguistic Diversity and Its Effect on Visual Description Models and Metrics." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

2. Chan, David M., et al. "Active learning for video description with cluster-regularized ensemble ranking." Proceedings of the Asian Conference on Computer Vision. 2020.

Chapter 3

Dataset-Specific Linguistic Diversity and Its Effect on Visual Description Models and Metrics

As discussed in chapter 1, automated visual description is an emergent field in computer vision, aiming to generate natural language descriptions of visual information. Unfortunately, despite recent improvements in model architectures (Liu et al., 2021a; Perez-Martin et al., 2021), metrics (Jiang et al., 2019; Wang et al., 2021b), and datasets (Monfort et al., 2021; Wang et al., 2019), automated visual description has been plagued by issues of poor generalization and description quality (Stefanini et al., 2021; Aafaq et al., 2019; Smeaton et al., 2019; Yang et al., 2020a). Models consistently perform poorly on novel data, generate nonsense descriptions, or produce descriptions that are too vague to be of use to visually impaired users (MacLeod et al., 2017). It remains an open question in visual description to understand the source of these generalization issues.

This work is motivated by both the fact that often state-of-the-art methods outperform leave-one-out experiments with ground truth sample data (explored in 3.2) as well as results demonstrating poor cross-dataset generalization in video captioning from Smeaton *et al.*(Smeaton et al., 2019) and Yang *et al.*(Yang et al., 2021b). We find that one major issue in current datasets—description linguistic diversity—explains a great deal about model evaluations.

Our work, consisting of analyses on several popular visual description datasets, contains several primary contributions:

1. We demonstrate that a lack of linguistic diversity at a token and n-gram level can bias models to generate descriptions lacking in semantic detail (section 3.3).

2. We show that diversity among ground truths for a single visual context presents a catch-22: low within-sample linguistic diversity leads to generic captions, as information is repetitive; on the other hand, high within-sample diversity leads to a breakdown

of single-sample metrics, causing inconsistencies in model evaluation and inaccurate understanding of model performance (section 3.4).

3. We detail how a lack of semantic diversity at the dataset level can encourage models to generate generic descriptions through classification, instead of learning to understand and relay visual phenomena at various levels of detail (section 3.5).

4. We discuss our findings demonstrating the need for future research in visual description datasets, methods, and metrics, present recommendations on possible solutions to current linguistic diversity, and introduce a new toolkit for dataset evaluation and split generation focused on linguistic diversity (section 3.6).

3.1 Experimental Design

In this work, we explore the field of visual description data through the lens of some of the most popular visual description datasets. While there are a large number of visual description datasets to choose from, we decided to focus on some of the most common datasets for video description, and an additional dataset for image description: [1] MSR-VTT (Xu et al., 2016), VATEX (Wang et al., 2019), MSVD (Chen and Dolan, 2011) and MS-COCO (Lin et al., 2014) (for full details, see Appendix A).

All of these datasets collect multiple ground truth descriptions per visual context, and the ground truth descriptions that they do collect are generated by human annotators (via Amazon Mechanical Turk for these datasets). Unfortunately, very large benchmark datasets such as Conceptual Captions (Sharma et al., 2018) and HowTo-100M (Miech et al., 2019) often contain only a single description per image/video, of questionable quality as the datasets are not annotated by hand. While datasets like S-MiT (Monfort et al., 2021) contain human-annotated ground truths, they post-process spoken language with automated speech recognition tools, making the dataset difficult to analyze from an n-gram metric angle. Both ActivityNet Captions (Krishna et al., 2017) and YouCook (Zhou et al., 2018a) are dense video description datasets that contain high-quality descriptions, however only contain a single ground truth per video.

Given the datasets, we will contextualize our experiments through the lens of several standard metrics for visual description. The BLEU (or BLEU@N) (Papineni et al., 2002) score is a measure of n-gram precision, the ROUGE-L (Lin, 2004) score is a measure of longest common sub-sequence recall, the METEOR (Banerjee and Lavie, 2005) score is a F1-oriented alignment-based metric, and the CIDEr (Vedantam et al., 2015) score is a TF-IDF weighted similarity metric. For more details of the individual metrics, see Aafaq et al. (2019). Recently, metrics which focus more on including visual content directly such as TIGEr (Jiang et al., 2019) and FAIEr (Wang et al., 2021b) have shown improvements in human-judgement

[1] As described in section 3.6, we make the tools available for this analysis public, so any additional datasets can be analyzed.

correlation and scores such as CLIP-score (Radford et al., 2021b), BERT-score (Zhang et al., 2020d), and SMURF (Feinglass and Yang, 2021) have been shown to closer approximate semantic content. While improving the metrics is an extremely important area of research, we also believe that analyzing both why current metrics are failing and what patterns models exploit to optimize these metrics, can give essential insight into model improvements.

We selected a set of recent works from the field as representing the state of the art. For visual description on MSR-VTT and MSVD, we refer to SemSynAN (Perez-Martin et al., 2021), a recent work that uses semantic embeddings based on POS tagging to achieve strong results. SemSynAN was not evaluated on the VATEX dataset, so for VATEX, we refer to the performance of MGRMP (Motion Guided Region Message Passing) (Chen and Jiang, 2021), a recent method for visual description which leverages message passing between object regions. For MS-COCO, we refer specifically to Vin-VL (Zhang et al., 2021b), a method that uses object-level attention and vision and language pre-training for visual description.

3.2 How Can Models Outperform Humans?

Recently, there has been a strong contrast between the metrics-based evaluation of methods for generating visual descriptions on data sets and whether those methods generalize to real-world use cases (Stefanini et al., 2021). The goal of our analysis is thus to understand some of the core reasons why models are failing to generalize and to make recommendations for the future design of datasets, models, and metrics, in an attempt to avoid further generalization shortcomings.

A core indicator of the difficulty of using standard metrics to improve generalization is that the “leave-one-out" performance of the ground truths for each dataset is typically poor. Because we investigate datasets that have more than a single ground truth sample per visual context, we can measure the metric scores between a randomly sampled ground truth, and the remaining ground truths for that visual context. When averaged over many trials, this stochastic approach generates an estimate of human performance on the dataset (see Appendix A for details).

Our results are summarized in Figure 3.1. We can see that SOTA methods significantly outperform this estimate of human performance on the MSVD, MSR-VTT, and MS-COCO datasets. This result is not only counter-intuitive, but detrimental to progress in the field of video description, as it draws into question the usefulness of standard metrics as an indicator of model performance and generalization. These results motivate questions of understanding: “Why, and how, do models exploit the current metrics to achieve strong performance?” and “How can we limit the the exploitation of N-Gram centric metrics”. The goal of the next several sections is to explore these questions through the lens of data diversity. Through analysis of single-token, n-gram, within-sample, and cross-sample diversities, we demonstrate how linguistic patterns affect models and metrics and explore how we can mitigate these effects.

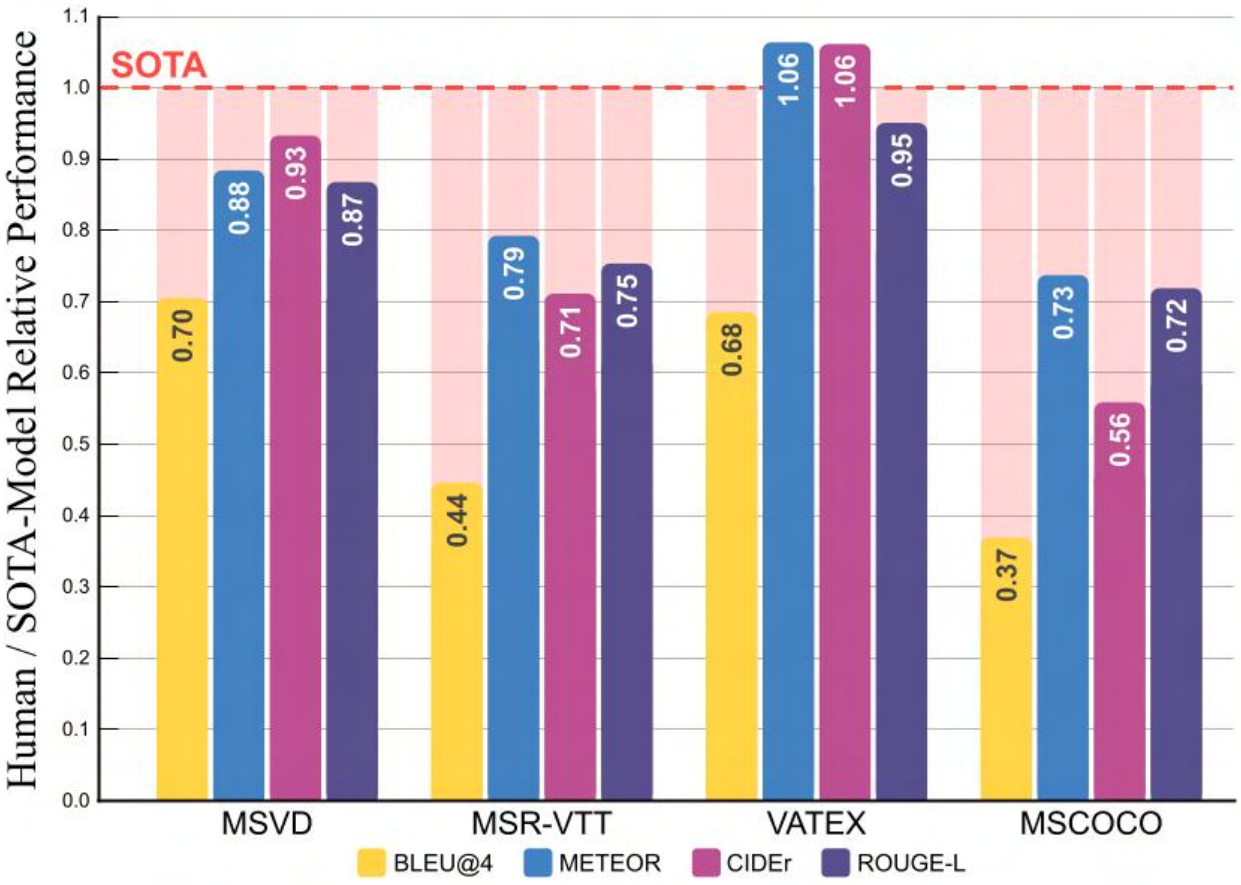


Figure 3.1: Captions generated by state-of-the-art (SOTA) models outperform held-out ground truth captions written by humans on common visual description datasets and metrics. Despite being far from human-level, SOTA models appear to outperform humans on most datasets and metrics, with the exception of VATEX, a relatively new dataset (and not even on all metrics). This discrepancy begs the question, "What causes these effects?" and "Are these effects indicative of a more serious issue with visual description datasets or model evaluation methods?" The figure above shows metric performance normalized to a recent SOTA model across several visual description datasets.

Dataset	Unique	BS-Unique	WS-Unique	Head
MSVD	9455	1.21%	11.8%	944
MSR-VTT	22780	0.76%	21.55%	1636
VATEX	31364	0.33 %	24.87%	1363
MS-COCO	35341	0.22%	33.76%	824

Table 3.1: Vocabulary metrics for each of the datasets. Unique: The number of unique tokens. BS-Unique: Average percent of tokens per description that are unique. WS-Unique: Average percent of of tokens that are unique within a sample. Head: The number of unique tokens comprising 90% of the total tokens.

3.3 Single Sample Diversity

3.3.1 Token-Level Diversity

Table 3.1 provides a token-level analysis of each of the datasets. In addition to reporting the number of unique tokens in the dataset, we also introduce three new measures of diversity. Why are such measures necessary? It is well known that text follows a power-law distribution (Zipf, 2016) in their vocabulary size, thus, the entropy of the dataset is approximated by the natural log of the vocabulary size. While measuring the entropy of the dataset is important, that only provides a general idea of the potential spread of the data, it doesn't say much at all about how easy the dataset is to solve with simple captioning model, capable only of classification. Ideally, in addition to the entropy of the data, we would also like to measure something like the mutual information between captions of the same image, and more generally, the conditional mutual information between captions across multiple images in the dataset (i.e. how easy it is to disambiguate the captions, given the image). To address this, we introduce the following measures:

- **"Between-Sample Uniqueness"** measures the percentage of tokens in each caption that are unique in the dataset. Between-Sample Uniqueness assesses the proportion of tokens in each caption that are unique across the dataset. This measure directly relates to the concept of conditional entropy, which quantifies the amount of information needed to describe the outcome of a random variable given that the outcome of another variable is known. In the context of image captioning, conditional entropy can be thought of as the amount of information required to predict the caption of an image, given the dataset. High Between-Sample Uniqueness suggests that many tokens used in captions are unique to specific images or contexts within the dataset, thereby increasing the conditional entropy. This indicates that predicting a caption based on the dataset's distribution requires more information, highlighting the diversity and specificity of the dataset. It challenges models to learn fine-grained distinctions between samples, as a large number of unique tokens implies that simple classification strategies may not

suffice for accurate caption generation.

- **"Within-Sample uniqueness"** measures the percentage of tokens that are unique within a particular image/video. This measure provides insight into the internal diversity of captions, reflecting the variability of language used to describe a single image or video. High Within-Sample Uniqueness indicates that captions contain a wide variety of tokens even within a single description, suggesting that the language used is rich and varied. This directly impacts the model's need to understand and generate nuanced, detailed descriptions, as opposed to relying on repetitive or generic language. In terms of the Zipf distribution, where a small number of words are extremely common while the majority are rare, this measure suggests an encouragement of utilizing the "long tail" of the vocabulary, thus enriching the model's capacity for detailed and specific caption generation.

- **"Vocab-Head"** measures the number of tokens making up 90% of the tokens in the dataset. The vocab-head provides a direct link to the Zipf distribution by highlighting the concentration of vocabulary usage within the dataset, indicating how reliant the dataset is on a core vocabulary. In datasets where the Vocab-Head is small, a significant portion of the content can be described using a limited vocabulary, which might simplify the task for models but also limit their ability to deal with rare or unique descriptions. This measure inversely relates to the concept of entropy, as a smaller Vocab-Head implies lower entropy and potentially lower conditional entropy, making the task of caption prediction more about recognizing common patterns than understanding nuanced or diverse descriptions. It encourages models to learn beyond the most frequent tokens to improve their performance on more complex, diverse datasets.

Together, these measures encourage a deeper examination of datasets beyond simple entropy calculations, and optimizing these measures compels models to engage with the data's complexity, diversity, and specificity, aligning with the goal of enhancing the mutual information between captions and images. This aligns with the broader objective of improving model performance on tasks that require nuanced understanding and generation of language.

From Table 3.1, we can see that while the number of tokens in the dataset can be diverse, the captions themselves are relatively lacking in diversity. Between sample diversity is relatively low, meaning that tokens are often re-used, suggesting a lack of unique vocabulary (which can often benefit models). Within-sample diversity ranges between 11% and 35%, suggesting that within samples, the descriptions are relatively varied. We discuss the impact of within-sample diversity in section 3.4. Particularly surprising is the size of the head of the token distribution for each of the datasets. As expected, a small fraction of tokens represent 90% of the occurrence in most of the datasets. In MS-COCO, 2% of the tokens represent 90% of the occurrences, while at the other extreme 10% of the tokens are required for MSVD. This begs the question: how does the effective vocab size impact performance?

To validate how effective vocab size impacts performance, we used the same setup as in section 3.2 to compute the performance of the ground truths, however, replaced tokens in the

Table 3.2: Effective vocab size (EVS), number of tokens per caption (TPC) and Effective Decision (ED@N). The EVS-n is the percentage of n-grams that do not act like 1-grams in the dataset. A large EVS-n means that language is more diverse, while a small EVS-n means that there are very few combinations of possible n-grams. The ED@N is the expected number of decision that a model has to make when generating captions of length N. WT-103 is WikiText-103 (Merity et al., 2017), a common natural language dataset.

Dataset	**TPC**	**EVS-2**	**EVS-3**	**EVS-4**	**ED@10**
MSVD	7.03	47.83%	25.29%	14.67%	2.90
MSR-VTT	9.32	52.96%	26.44%	13.68%	2.88
VATEX	15.29	54.84%	32.60%	18.86%	3.38
MS-COCO	11.33	53.91 %	32.59%	20.56%	3.51
WT-103	87.04	95.19 %	34.49%	17.81%	3.72

tail of the token distribution with unique "UNK" tokens. Performance dropped significantly in all cases, with the most dramatic drop for MSVD (drop of 63.87%) and the least for MS-COCO (drop of 51.23%). MSR-VTT experienced a decrease of 58.66% and VATEX experienced a decrease of 56.20%). Counter-intuitively, the longer the tail, the less performance decreased. This result, confirmed in classification by Tang et al. (2020), implies that models which generate from a limited vocabulary are advantaged (in terms of n-gram performance) when the head is relatively small, leading to undesirable generation behavior.

Following Wang et al. (2019), we analyze the datasets at the level of the parts of speech in the dataset (See Appendix A for details). VATEX has more than 2 verbs per caption on average (by design, see (Wang et al., 2019)) while the other datasets have at most 1.3 verbs. While VATEX is the most linguistically complex, the distribution has significantly different base statistics, likely explaining poor cross-dataset generalization to VATEX from MSR-VTT and MSVD trained models. MSR-VTT is the most diverse from an object perspective (1512 nouns representing 90% of the noun mass), which lends additional support to the observations by Zhang et al. (2020g), who find that a strong object detector and good object features are necessary for strong MSR-VTT performance. Notably, MS-COCO has a very high within-sample noun diversity, suggesting that many of the captions in MS-COCO focus on different objects in each sample, and supporting hypotheses introduced in Anderson et al. (2018) based on multiple-object attention for this dataset.

3.3.2 N-Gram-Level Diversity

From tokens, we can move on to exploring how the tokens fit together. One of the major issues in overall dataset diversity is a tendency for language models to accentuate a lack of n-gram diversity, leading to domination of common n-grams over visually likely n-grams (Hendricks et al., 2018). A standard metric reported by Wang et al. (2019) in VATEX is the

number of unique n-grams in the dataset, however, we find that alone, the number of unique n-grams does not allow for strong comparison between datasets, both because the number is not normalized, and the number of n-grams says little about the overall distribution of those n-grams.

Instead of only looking at the number of n-grams, in order to measure the amount of n-gram diversity that is introduced into a dataset, we introduce the N-Gram Effective Vocab Size metric (EVS-N), which measures the percentage of n-grams that do not act like 1-grams in practice. Formally, EVS-N is the percentage of tokens for which an N-gram language model has zero conditional variance (i.e. the percentage of tokens for which an n-gram language model does not assign 100% probability to a single next token). This metric can be thought of as a language-generation complexity metric — a higher EVS means that it will be more difficult for a model to memorize captions, while a low EVS suggests that models need only determine the first few words in order to generate a high-quality caption. Table 3.2 shows EVS-N performance, and a shocking result. The EVS-2 is approximately 50% for all datasets, suggesting that in the majority of cases, the model is able to make only one decision to generate two tokens, contrasting with WikiText-103 (Merity et al., 2017), where the EVS-2 is 95.19%.

In addition to just understanding the EVS, we can combine the EVS scores with the average number of tokens in the dataset to compute the average number of "decisions" that a model has to make during generation. The ED@N, or expected number of decisions made in a description of length N is also given in Table 3.2. Formally, the ED@N is the expected number of tokens in a description of length N for which an n-gram language model of the dataset has non-zero variance conditioned on the sentence so far. Surprisingly, most of the datasets have very similar ED scores (despite their differing average token lengths), and the number is low: only 3-3.5 decisions have to be made on average to get the desired caption. This low number has major implications in the quality of the captions: the fewer the number of decisions that need to be made at training, the less diverse the captions will be during test time, and the less likely models trained on the low-ED data will be able to generalize to fine-grained differences between samples. Further, this means that the number of captions models will be able to generate is restricted to V^{ED}, where V is the size of the vocab, a notably smaller number than expected with large vocab sizes, and long captions. We believe that this is one of the reasons that non-auto-regressive approaches such as those in Liu et al. (2021a) and Yang et al. (2021a) are able to perform so well on these datasets: they can focus on the visual information, and don't have to worry about the syntactic structure as it is similar for all descriptions.

3.4 Within Sample Diversity

While we have seen that token-level diversity is important for the generation of high quality captions, we also want to understand how within-sample diversity (i.e. diversity within a collection of ground truths for a single visual context) impacts the performance of

visual description models.

To define how much within sample diversity there is in a dataset, there are several methods that we can use. One metric, common to many papers, is an analysis of how many captions in each sample are novel. VATEX (100%) and MS-COCO (99.9%) have high caption novelty, while MSR-VTT (92.66%) and MSVD (85.3%) contain somewhat less exact novelty. Further, we could look at within-sample token diversity (shown in Table 3.1), which suggests that within a sample, diversity is actually relatively high, with 11% to 33% of tokens being unique within a sample. Further, the within sample verb (15% to 56%) and noun (13% to 35%) uniqueness is relatively high as well, suggesting that individually, captions discuss unique parts of a visual context (Full results are given in Appendix A). This is demonstrated qualitatively in Figure 3.4.

The issue with these measures of novelty is that they account only for novelty at the caption or token level by exact matching, but do not directly target the semantic novelty of the captions. In order to look closer at within-sample diversity, we compute the pairwise semantic distance between each description and all other unique descriptions in the sample using the cosine distance between MP-Net embeddings (Song et al., 2020) trained for sentence similarity. Figure 3.2 shows the minimum of the inter-sample cosine distances, a metric we call sample redundancy. Notably, almost 10% of the samples in MSVD have a very close semantic match, suggesting that MSVD has more semantically redundant information than other description datasets.

Sample redundancy is both a blessing and a curse. Datasets that have very high sample redundancy will tend to have high performance on leave-one-out ground truth metrics, as most of the ground truth captions will share large amounts of information. This means that pair-wise metrics such as the standard n-gram metrics will often perform well, as any generated sample should also lie close to at least one ground truth sample. Unfortunately, as we increase the number of diverse ground truths (increase the sample variance), the minimum distance between samples increases (See Appendix A for a figure). Because of this increase in distance, the leave-one-out performance of ground truths decreases, as shown in Figure 3.3, leading to a breakdown of the n-gram metrics (and all metrics that rely on a single-sample pairwise comparison to the set of ground truths). This effect is what causes SOTA models to outperform leave-one-out samples as demonstrated in section 3.2. While ideally, metrics should be independent of the variance in the ground truth data, for the datasets we analyze in the paper it is clear the sample variance is sufficient that this is not the case. Interestingly, the leave-one-out fall-off occurs at different rates for the different datasets, suggesting that some datasets are more-redundant to semantic variance than others: while we hypothesize that this is due to the choice of tokens and distribution of semantic structure, it is interesting future work to confirm this hypothesis.

Why are SOTA models immune to the effects of sample variance? It's important to note that when evaluating models, we only look at *a single sample from the model distribution*. We hypothesize that instead of attempting to approximate the full distribution of captions, models are picking up on trends between samples in the data, such as a wealth of descriptions that contain simple semantic structures (as described in section 3.3) or individually strong

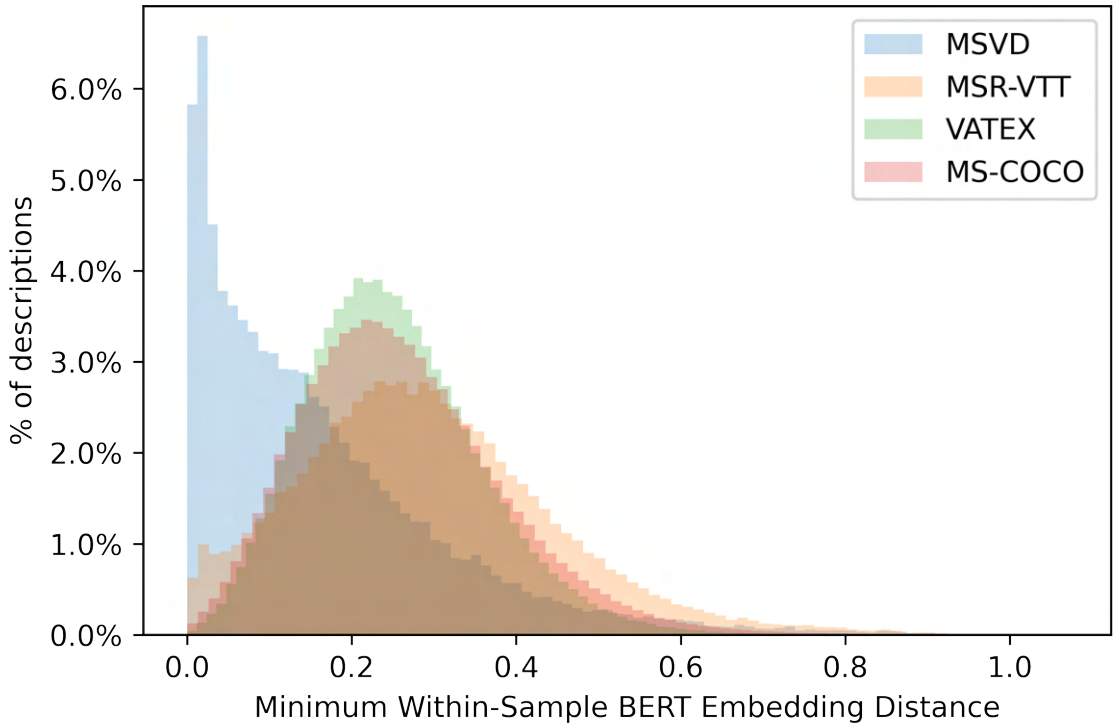


Figure 3.2: Histogram of within-sample minimum distances under the MP-Net (Song et al., 2020) BERT-style embeddings. MSVD and MSR-VTT both have a high number of descriptions which have zero within-sample minimum distance, while MS-COCO and VATEX have a higher within-sample diversity.

training descriptions (which we will discuss in section 3.5) which allow the model to reduce the effective variance of the ground truth dataset during the evaluation phase by ignoring most of the ground truth captions, and only focusing on a specific subset of descriptions. While these trends are likely model-specific, we believe it is important future work to quantify and understand the kinds of descriptions that models learn to approximate, and more closely monitor the effects of over-fitting to a small subset of captions to reduce the effects of ground-truth sample variance.

The effect of reducing semantic variance appears in practice via a training trick exploited by both Perez-Martin et al. (2021) and Liu et al. (2021a) who find that *decreasing* the number of reference captions during training leads to improved evaluation performance on n-gram metrics. By artificially restricting the semantic variance of the training dataset, models are able to over-fit to a smaller subset of semantically redundant captions, and exploit current pairwise metrics.

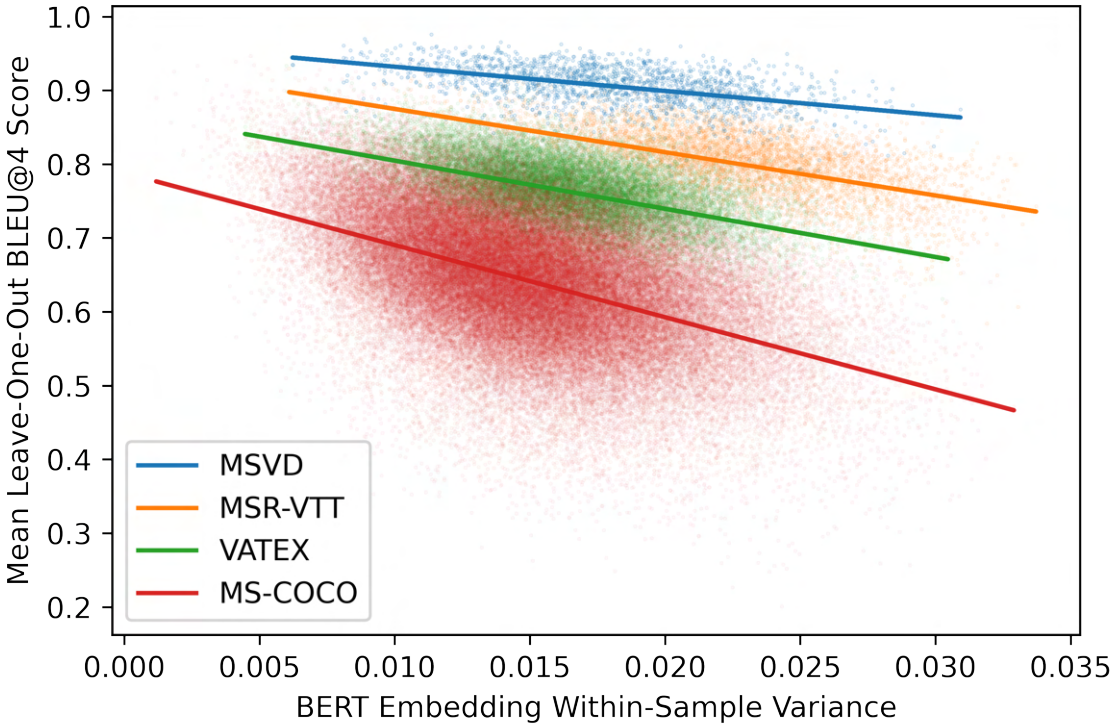


Figure 3.3: Plot showing the relationship between semantic variance and the performance of leave-one-out ground truth estimates of human performance on the BLEU@4 metrics. As we increase semantic variance, the average minimum distance between ground truth samples increases, and metric performance falls.

Thus, we are stuck in a catch-22 when it comes to adding more captions per sample. If we increase the number of captions, we decrease our metrics' ability to accurately discern caption quality, however if we reduce the number of captions, we can improve the accuracy of current metrics, and obtain models that achieve higher metric scores, at the cost of bland and generic captions.

3.5 Dataset Level Diversity

Not only do sample level diversity and within-sample diversity have important impacts on models and metrics, but dataset-level conceptual diversity matters as well. A common criticism of captioning models is that they are not generative, but instead, reproduce captions from the training set based on a set of global criteria. In general, we hypothesize that a lack of diversity in the dataset, both in the lack of overall visual concept diversity, and the

Dataset: MSR-VTT (Video 6641)
Inter-Sample BERT-Embedding Distance [Min, Avg., Max]: [0.41,0.58,0.66]
Mean Leave One Out BLEU@4 Score: 0.4028

- there is a woman is talking about a simple recipe (BLEU@4: 0.920)
- there is a woman is serving a dish on the table
- a plate of food is displayed with two chopsticks next to it
- a woman is making a singapore style noodle dish in a pot
- a noodle dish with sprouts and shrimp sits on a countertop
- a pair of black chopsticks with decorative bands rests on a bowl of noodles with shrimp white bean sprouts and leafy green garnish
- some singapore noodles of egg is kept on the plate and the stick is placed on it
- there is a woman is talking about noodles
- the narrator tells about making singapore style fried noodles with prawns in a prawn-based broth
- noodles kept on the plate along with the sticks to eat it
- a woman describing a noodle dish in a white bowl (BLEU@4: 0.978)
- a noodles is been placed in the plate and a girl is describing about it
- delicious noodles are in a bowl with chopsticks
- a lady is explaining how to make a singapore dish which consists of fried noodles and prawns
- a woman describes an elegant recipe consisting of fried noodles with prawns in a prawn-based broth garnished with fresh lime wedges
- a bowl of vegtables prawns and noodles is shown chop sticks rest on the edge of the bowl while a woman's voice talks about the dish
- a lady is presenting fried noodles with prawns in a prawn broth
- there is a bowl on the table with a singapore style dish of fried noodles with prawn in a prawn broth

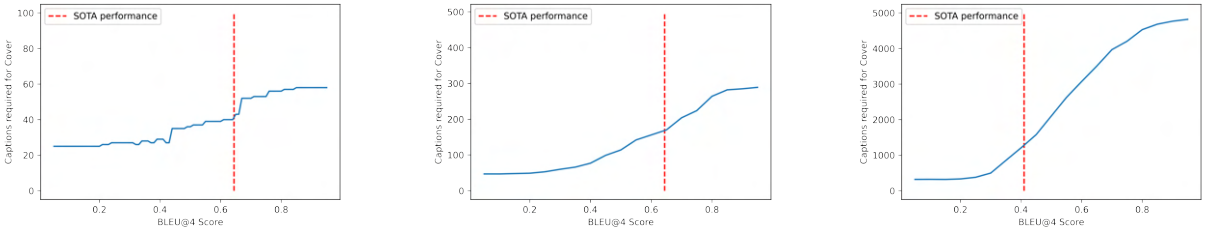
O2NA Baseline: there is a man is talking about a dish (BLEU@4: 0.467)

Figure 3.4: A qualitative example from MSR-VTT demonstrating several diversity effects. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Notably, both captions are much more generic than the other captions in the data, a trend which is consistent across all samples. We can see that the variance within this sample is high, however the tokens themselves are similar (annotators select similar tokens for the same sample). Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

exact distribution of that diversity in the dataset itself leaves models vulnerable to choosing classification over generation. We further hypothesize that a lack of conceptual diversity leads models to produce a few generic captions based on high-level visual features, instead of generating semantically detailed captions. In order to support this hypothesis, we attempt to answer two questions: "how much performance can we achieve with classification alone?" and "how much does the explicit selection of visual samples encourage models towards classification over generation?"

3.5.1 How many captions make up a dataset?

One interesting question to ask is, how many captions do you reasonably need to use in order to solve a dataset to a particular score? This metric is a reasonable proxy for concept-level diversity, and can more globally measure the performance of a model. To answer this question, we used a greedy approximation algorithm for optimal set cover to approximate the minimum number of captions from the training set that need to be chosen for MSR-VTT and MSVD in order to achieve a particular BLEU@4 score on the validation set. We don't compute this number for VATEX/MSCOCO or metrics beyond BLEU due to the computational cost of computing a full matrix of caption distances. Figure 3.5 demonstrates the results of this experiment. We can see here that to achieve SOTA BLEU@4 performance, we need only to select optimally from a set of 43 captions in the case of MSVD, and 156 captions in the case of MSR-VTT. Even further, it's interesting to see that with only 58 captions in MSVD and 289 captions in MSR-VTT, we can achieve almost optimal BLEU



(a) MSVD Dataset (b) MSR-VTT Dataset (c) MS-COCO Dataset

Figure 3.5: For several datasets, how many captions from the training dataset are required to achieve a particular BLEU@4 score on the test set. We can see that in the optimal case, only a few (58 for MSVD, 197 for MSR-VTT, 1578 for MS-COCO) captions are required to achieve SOTA performance on the dataset. Notably, MS-COCO uniquely requires a unique description for each image.

Table 3.3: Percentage of samples in the visual description datasets which contain at least one description that has a sub-string matching a label from the pre-training dataset.

Dataset	**ImageNet**	**Kinetics**	**COCO**	**Places**
MSVD	98.27%	38.88%	89.03%	55.68%
MSR-VTT	68.88%	23.51%	59.82%	46.44%
VATEX	98.60%	40.12%	76.86%	60.55%
MS-COCO	93.22%	8.83%	91.70%	60.49%

scores.

This particular result, combined with the fact that models only need to make a few token-level decisions when generating language (See subsection 3.3.2) appears to be a real cause for models producing generic captions. Not only do models not have to make many decisions, but overall, they don't have to select from many visual concepts either.

3.5.2 Does the feature set matter?

Caption models are limited not only by a classification effect but also by the concept-level diversity of the feature extractors that they use. When models rely on particular feature extraction methods, we expect pre-initialized features to bias models towards classification over generation, particularly classification among the concepts present in the pre-training data. Recently, Srinivasan and Bisk (2022) showed that these biases can compound - so it seems natural to ask the question: how much do we expect biases in our datasets to compound with feature extractor bias?

Table 3.4: Performance on BLEU@4 score when using the best core-set ground truth from overlapping categories. Performance remains surprisingly high when using shared captions, implying that models are able to leverage template captions instead of scene understanding. GT: random within-sample leave-one-out ground truth performance.

Dataset	**GT**	**ImageNet**	**Kinetics**	**COCO**	**Places**
MSVD	0.453	0.652	0.442	0.634	0.470
MSR-VTT	0.210	0.678	0.467	0.650	0.521
VATEX	0.234	0.576	0.460	0.547	0.485
COCO	0.152	0.680	0.515	0.704	0.292

In order to measure how much particular datasets are biased towards particular feature extractors, we compute a concept-level "overlap" between several popular feature datasets (Deng et al., 2009; Carreira et al., 2018; Lin et al., 2014; Zhou et al., 2017), and the visual description datasets. Table 3.3 demonstrates the percentage of samples in the visual description datasets which contain at least one description that has a sub-string matching a label from the pre-training dataset. While exact overlap from labels to descriptions may exclude some cases (for example the label "playing baseball" does not overlap with any description which has only the word "baseball"), we found that fuzzy matching induced significant numbers of false-positives. This metric thus, represents a lower-bound on the overlap (as can be seen in the case of MS-COCO, where only 91% of the descriptions contain an object from the official label set).

We can see that in datasets except for MSR-VTT, the dataset overlap with ImageNet is relatively high, likely leading to models which achieve performance based solely on the use of ImageNet features, as the classification effect detailed in both subsection 3.5.1 and subsection 3.3.2 can be exaggerated. Similarly, for datasets besides MSR-VTT, adding object detection features is likely to exaggerate the classification effect, as the model will be pre-disposed to split samples into object-category bins.

To explore exactly how much classification performance can be achieved splitting only along feature extractor boundaries, we generate sets of captions that match (using exact matching) a particular label in the feature extractor pre-training dataset. For each sample, we generate a hypothesis using a randomly sampled caption from the union of the matching concepts and compute the metric score of that hypothesis (See Appendix A for a detailed discussion). The results of this experiment are given in Table 3.4, and we can see that without sufficient conceptual diversity, models can achieve strong performance by segmenting samples among higher-order labels instead of leveraging visual understanding.

3.6 Recommendations & Limitations

Our aim in this work is to demonstrate that there are three unique levels of diversity that need to be maintained when collecting a dataset: Token-level diversity, within-sample diversity, and dataset conceptual diversity.

In section 3.3 we showed that a lack of token diversity diversity can lead to simple captions from a core data level: few decisions need to be made to generate captions, and a large number of the tokens responsible for this generation are relatively common, opening the door for potential limits to model diversity. Token-level diversity is primarily controlled during the labeling phase of dataset collection, so we believe that both when researchers collect novel data, and when they are building splits for current datasets, they should focus on token diversity. Primarily, to encourage models to generate from a diverse set of captions, we recommend maximizing the ED@N score from section 3.3, along with increasing token EVS by improving the diversity of collected captions. Prompts encouraging crowd-source workers to include higher semantic detail and limits on sentence complexity (such as those introduced in VATEX (Wang et al., 2019) and Barbosa and Chen (2019)) could help prevent token-diversity effects from appearing in downstream models.

On the other hand, collecting too many ground truths, as discussed in section 3.4 presents a model training issue. Currently, models are trained to reduce semantic variance, which can lead to captions which are less complex than we expect. We believe that it is essential future research to explore how to account for the fact that variance in ground truth video descriptions is signal and not noise. Methods for managing multi-modal conditional distributions such as Slade and Gedeon (1993) or multi-label learning such as Tsoumakas and Zhang (2009) may represent step towards such methods. Further, metrics that we use reinforce semantic variance effects by computing maximums with single samples. We believe that investigating metrics which focus on comparing multiple model samples to the full set of ground truth samples represents a possible solution. By forcing models to approximate the entire ground truth distribution we may avoid creating models which optimize away variance in the data.

Finally in section 3.5, we discussed how a lack of diversity at a concept level can impact the performance of models. When metrics have fewer global concepts, or high overlap with feature extraction methods, they are more likely to trend towards classification over generation. In order to remedy this effect, we recommend the creation of datasets through sampling independent from the label sets of feature models. We additionally recommend that when creating training, validation, and testing splits in the dataset, the concept-level diversity is monitored to avoid introducing potential feature or concept biases with respect to popular feature extraction methods.

Visual Description Toolkit Alongside this work, we released a new toolkit[2] for visual description dataset evaluation, which is designed to analyze the performance of models (or ground truths) across the axes explored in this work. We hope that by making tools for

[2]Toolkit available at `https://github.com/CannyLab/vdtk`

evaluating visual description datasets easily accessible, we can encourage the field to deeply investigate the sample diversity in their data and predictions. We hope that such methods for evaluation can help uncover the deviations of the model from the ground truth data, and paint a more complete picture of our descriptive models beyond n-gram scores.

Limitations While we have demonstrated how diversity at several levels directly impacts the performance of downstream models, we believe that additional research is required to further understand how the problem of visual description differs from classification and natural language processing. In section 3.4, we use several proxies for caption complexity, however it is not immediately clear that such proxies are good measures for the semantic complexity of a caption. As far as we are aware, no such measure of the "usefulness" of a caption to a visually impaired user exists, that we can use to evaluate our current caption data. Figure 3.4 and the additional qualitative examples in Appendix A) demonstrate some correlations between caption complexity, and the mean caption, however we believe that deeper analysis is necessary.

Our methods are also limited by the choice of metrics used in this work. Explorations of recent metrics such as FAIer (Wang et al., 2021b) may indicate that they alleviate diversity effects by focusing on visual information over textual information, and leveraging pre-trained grounding models. While novel metrics may solve some of the problems, the training effects observed in section 3.4 remain common between all models, and the diversity in section 3.3 and section 3.5 are local to the datasets, and will remain regardless of the metric used.

3.7 Related Work

This is not the first work to analyze video description data from a dataset and metric perspective, however, we believe that it is the first to focus on how dataset diversity and metric choices directly affect caption generalization. Hendricks et al. (2018), Bhargava and Forsyth (2019), Tang et al. (2021) and Zhao et al. (2021) have all demonstrated that visual description data is often biased with respect to protected attributes (such as race, gender or religion), and introduced new methods for handling specific biases - however, they do not discuss the impact of general biases on model performance. Both Smeaton et al. (2019) and Yang et al. (2021b) demonstrate poor cross-dataset generalization in visual description, and demonstrate that the choice of dataset directly affects model generalization ability, as well as introduce additional model-centric methods for mitigating the impact of dataset effects. These works complement our own, and they support our core hypotheses that we discuss in section 3.6.

Outside of visual description, the evaluation of how linguistic data and metrics affects the performance of downstream vision and language models is prevalent. Cadène et al. (2019) demonstrate unimodal language biases in visual question answering and Choi et al. (2019) do the same for action recognition. While many papers (Yang et al., 2020a; Shah et al., 2020; Li and Vasconcelos, 2019; Singh et al., 2020; Clark et al., 2020; Joo and Kärkkäinen, 2020) make

recommendations for reducing linguistic bias based on the modeling framework, these works do not focus on the quality of generation, and instead, focus on the equally important trend of models relying heavily on language priors to solve tasks. Barbosa and Chen (2019) introduce methods for dataset collection which attempt to reduce linguistic bias, which represents a great leap forward from standard Amazon Mechanical Turk (AMT) collection methods, but does not discuss how the diversity impacts the performance of downstream models beyond balancing language priors.

3.8 Conclusion

In this work we have taken a close look at linguistic diversity in common visual description datasets, and detailed how diversity can impact models and metrics. At the token level, we showed that a lack of diversity impacts the ability of metrics to assess the quality of captions, and the ability of models to generate diverse descriptions. At the sample level, we demonstrated that high within-sample diversity is both a blessing and a curse, leaving us with either a failure of metrics to correctly measure performance, or leaving us with correct metrics, but bland and generic captions. Finally, at the dataset level, we demonstrated that even when single sample and within-sample diversity is maintained, a lack of conceptual diversity at the dataset level can bias models towards visual classification over language generation, opening the door for models which can use a few, generic, samples to solve the visual description task instead of generating captions which are rich in semantics.

While this work demonstrates the potential pitfalls of a lack of diversity in visual description datasets, we believe that by introducing new tools for analysis, and additional recommendations for data collection and model evaluation, the field will be able to investigate the sources of poor model generalization more closely, and build models which are both robust to visual diversity and can generate diverse, high quality, and semantically meaningful captions.

Chapter 4

Active learning for video description with cluster-regularized ensemble ranking

As discussed in chapter 1, automatic video description is an emerging area in computer vision research that aims to generate textual descriptions of the visual components of a video. Unfortunately, training models to do video captioning requires manual descriptions of every second of the video from a large corpus of representative videos. One of the largest current single-clip video captioning datasets, MSR-VTT, has only tens of thousands of unique uncorrelated videos whereas solving video captioning will likely require several orders of magnitude more to express the wide diversity of subjects, situations, and relationships possible in video data.

Active learning is a valuable approach in domains where unlabeled and partially labeled examples are readily available but obtaining manual annotations is expensive, such as is the case with automatic video captioning. However, while there has been significant investigation of active learning for computer vision tasks such as object recognition (Collins et al., 2008), object detection (Vijayanarasimhan and Grauman, 2011), video classification (Yan et al., 2003) and video segmentation (Vijayanarasimhan and Grauman, 2012), video captioning has received comparatively little attention. The reason for this is likely rooted in the complexity of the label space. Video captioning requires both sequential input and output, dramatically increasing the complexity of traditional active learning frameworks. To our knowledge, this is one of the first works to define active learning strategies for efficiently collecting training sets for automatic video captioning.

In this work we explore several active learning strategies for sequence to sequence active learning in video captioning, including uncertainty sampling based on label confidence, sequence entropy and query by committee methods. There are several unique challenges to active learning for deep sequence to sequence models: While traditional active learning methods (Settles, 2009) select one example at a time to label, retraining the model in its entirety after each new example selection, this strategy is impractical for training models such as transformer networks and LSTMs (Zhou et al., 2018b; Venugopalan et al., 2015), due to increased training time (hours vs. minutes) and increased inference time (seconds

vs. milliseconds). Thus, it is far more efficient to select a large batch of examples at a time to label when using a crowd-sourced collection process (Xu et al., 2016; Deng et al., 2009). Traditional batch-active learning often uses ranking functions which are intractable in deep sequence to sequence learning (Hoi et al., 2009; Brinker, 2003; Vijayanarasimhan et al., 2010), making active learning for video description a challenging problem, with no tractable solutions for deep neural networks.

We conduct a thorough empirical analysis of various active learning strategies on two recent and standard video captioning datasets, MSR-VTT and LSMDC, using both transformer based and LSTM based captioning models, and describe a novel cluster-regularized method which is both tractable to compute, and provides strong performance in our test scenario. Our key contributions are:

1. Demonstrating that traditional uncertainty sampling techniques do not significantly outperform random sampling, likely because of the difficulty of estimating the sequence entropy.

2. A novel ensemble based ranking method (Cluster-Regularized Ensemble Divergence Active Learning, Section 4.2.1) specifically designed for video description active learning which outperform random sampling by a significant margin.

3. A clustering-based active learning regularization method which can help to increase sample diversity, and when combined with our query-by-committee methods can save as much as 60% of the manual annotation effort while maintaining high performance (Section 4.2.2).

4.1 Related Work

In order to reduce human effort when constructing training sets, various active learning strategies have been proposed for computer vision tasks such as object recognition (Collins et al., 2008; Vijayanarasimhan and Grauman, 2009), detection (Vijayanarasimhan and Grauman, 2011), video classification (Yan et al., 2003) and video segmentation (Vijayanarasimhan and Grauman, 2012). These methods typically select the next example to query for a label based on uncertainty sampling, entropy, or predicting reductions in risk to the underlying model (see (Settles, 2009) for a comprehensive review). However, active learning for sequence labeling tasks such as automatic video captioning has received litle attention.

In the natural language processing literature, active learning methods have been proposed for actively selecting examples based on uncertainty sampling (Culotta and McCallum, 2005; Scheffer et al., 2001) or query by committee approaches (Dagan and Engelson, 1995). In (Settles and Craven, 2008), the authors provide a thorough analysis of various active learning methods for sequence labeling tasks using conditional random field (CRF) models. Current state-of-the-art video captioning models, however, typically utilize neural network based architectures such as transformer networks (Zhou et al., 2018b) or LSTMs (Venugopalan

et al., 2015) and very little research exists on how to successfully apply active learning for complex models — Transformer networks and LSTMs are expensive to train, taking hours to days to converge, compared to shallow linear models or CRFs employed in previous active learning studies (taking only minutes). Therefore querying a single example at a time is inefficient. It is far more efficient to select a large batch of examples at a time to label when using a crowd-sourced collection process as is typically the case (Vijayanarasimhan and Grauman, 2011).

Batch-mode active learning methods have been proposed for vision and other tasks in (Hoi et al., 2009; Brinker, 2003; Vijayanarasimhan et al., 2010). Batch selection requires more than selecting the N-best queries at a given iteration because such a strategy does not account for possible overlap in information. Therefore, the selection functions typically try to balance informativeness with a diversity term to avoid querying overlapping examples (Brinker, 2003). In this work, we take cues from (Brinker, 2003), and develop a batch active-learning method for sequence learning, that is regularized using a measure of information diversity (an idea from (Brinker, 2003)), but is tuned to be computed efficiently over sequence learning tasks, such as those in (Settles, 2009).

In addition to moving to batch sampling, automated video description is unique in that it has multiple possible correct sequence labels. Recent methods are usually based on expected gradient updates (Huang et al., 2016a) or the entropy of a sample distribution (Settles and Craven, 2008), and are unable to account for scenarios where there are multiple correct labels, or there is dynamic underlying label entropy. In addition, these methods often require computing an estimate of expected model updates over the space of possible labels. This estimate can be extremely expensive for sequence learning (which has exponential label space growth), and there's no clear way of sampling from caption spaces without learning a complex language model.

Among recent methods, Coreset active learning (Sener and Savarese, 2018), uses an integer linear program (or a greedy optimization) to find a lambda-cover over the feature set. By operating at a feature level, Coreset takes advantage of the semantic compression of the model to find sets of unlabeled samples that are useful to the model's prediction. We discuss our method compared to Coreset in Section 4.2.3.

Some recent methods including VAAL (Sinha et al., 2019) and ALISE (Deng et al., 2018) have approached active learning from an adversarial perspective. These methods train a discriminator which attempts to determine which samples are labeled and unlabeled, then select the likely unlabeled samples for training. However, they typically require large number of samples to reliably train the discriminator which is unavailable in the beginning of the active learning process. Nonetheless, it would be an interesting future direction to explore adversarial models for active learning on complex latent spaces. Deep Bayesian active learning (Gal et al., 2017) shows some promise, however strong Bayesian networks for multi-modal vision and language problems are still out out of reach for large scale complex datasets.

4.2 Query By Committee Ensemble Active Learning

In this work we introduce a new method for sequence active learning for video description, Query By Committee Ensemble Active Learning, and compare against several baseline algorithms (Those listed below, along with Coreset (Sener and Savarese, 2018) active learning and ALISE (Deng et al., 2018)). Throughout this section, we refer to a video v_i, and its associated set of descriptions $\mathcal{D} = \{c_1(v_i) \dots c_n(v_i)\}$. A set of descriptions generated by a model m_j is referred to by $\{c_{m_j,1}(v_i) \dots c_{m_j,n}(v_i)\}$. Videos may have multiple descriptions either through multiple-sampling of the model generative distribution, or through multiple ground-truth labels of the same video. The probability distribution $P_{m_j}(c_i)$ is the likelihood of a description c_i under the model m_j, and the distribution $\mathcal{P}^{cond}(m_j, c_i^k) = P_{m_j}(c_i^k(v_i)|c_i^{k-1}(v_i), \dots, c_i^0(v_i))$ is the conditional distribution of the next word k under the model given the previous words in the description.

4.2.1 Active Learning Methods

Random Selection Baseline: Our baseline method is to sample new data points from the training set uniformly at random. Random selection is a strong baseline. It directly models the data distribution through sampling, placing emphasis on representative data, but not "novel" data. Trying to sample outside the random distribution is more likely to cause over-sampling of parts of the data (demonstrated in Figures 4.3), leading to poorer overall validation performance.

Maximum Entropy Active Learning: Traditional methods for active learning (Settles, 2009) are often entropy based. As a second strong baseline, we present a maximum entropy active learning method in which we rank samples based on a sample of the entropy of the dataset. Unfortunately, given the exponential number of computations that have to be made in the sequence length, the entropy of the entire output distribution is intractable to compute directly. Thus, to approximate the entropy of the description distribution we compute the mean entropy of the word output distributions at each new word along the generation process of a new description of a sample using our current model. Thus, using a candidate model m, we sample K candidate sentences for each video, and we select samples which maximize the ranking function:

$$R(v_i) = \frac{1}{K} \sum_{k=1}^{K} \sum_{w=1}^{|c_{m,k}(v_i)|} -P_m(c_{m,k}^w(i)) \log P_m(c_{m,k}^w(v_i)) \tag{4.1}$$

where $R(v_i)$ is our approximate estimate of the entropy of any given sample's distribution.

Minimum Likelihood Active Learning: In the minimum likelihood active learning scenario, we select samples where the descriptions that the model generates have the lowest log likelihood under the model distribution. Thus, using a candidate model m, we sample

K candidate sentences for each video, and then choose samples which minimize the ranking function:

$$R(v_i) = \frac{1}{K}\sum_{k=1}^{K}\sum_{w=1}^{|c_{m,k}(v_i)|} \log P_m(c_{m,k}^w(v_i)|c_{m,k}^{w-1}(v_i)\dots c_{m,k}^0(v_i)) \tag{4.2}$$

Empirically, we find that the minimum likelihood active learning method is a stronger method than the entropy for use in video captioning (See Figure 4.2), however this measure of uncertainty suffers from the fact that the model may be very confident about its wrong answers, and will be unable to learn effectively when this is the case. Because these very confident wrong answers are never sampled (or are sampled later in the training process), the model is unable to correct for the initially introduced bias.

Query By Committee Ensemble Agreement Active Learning: To help alleviate the issues with single model uncertainty, we introduce the notion of an ensemble agreement active learning ranking based on query by committee methods for traditional active learning (Dagan and Engelson, 1995). With this method, we sample a set of likely captions from each member of an ensemble of models (using beam search), and compute the mean pairwise likelihood. For an ensemble of L models $\{m_1, \dots, m_L\}$, from each model m_l we sample captions $\{c_{m_i,1}...c_{m_i,K}\}$ for each available unlabeled video. Our ranking criterion is then to minimize:

$$R(v_i) = \frac{1}{L(L-1)}\sum_{p=1}^{L}\sum_{\substack{q=1\\q\neq p}}^{L}\sum_{k=1}^{K}\sum_{w=1}^{|c_{m_p,k}(v_i)|} \frac{\log \mathcal{P}^{cond}(m_q, c_{m_p,k}^w(i))}{K|c_{m_p,k}(v_i)|} \tag{4.3}$$

The idea here is to select samples for labeling which have low agreement, as these are the samples have higher uncertainty under our model/training process. In this scenario, we alleviate many of the concerns with models having high confidence in wrong answers, as this phenomenon tends to be local to particular models, and these highly incorrect answers will have low likelihood under the learned distributions of the other members of the ensemble.

Query By Committee Ensemble Divergence Active Learning (Proposed Method): While entropy/perplexity measures for active learning have been well explored in the literature (Settles, 2009), it is unclear if these measures are correct for the captioning task. Even if the caption distribution for a video has high entropy, meaning there are many possible likely captions (or even many possible correct captions), this high entropy does not mean that the model is unsure of the outcome. Samples that have many possible captions will thus be over-sampled, since any of the generated captions will have fundamentally lower likelihood than a sample with fewer possible captions. In order to avoid this, we present a method, which computes the KL-divergence between the conditional distributions of the ensemble members. Thus, if the models would choose similar words, given similar inputs - we consider the models to be in agreement. Similarly to the above, for an ensemble of L models $\{m_1...m_L\}$, from each model m_l we sample captions $\{c_{m_i,1}...c_{m_i,K}\}$ for each available unlabeled video. We then

choose samples which maximize:

$$R(v_i) = \frac{1}{L(L-1)zhou} \sum_{p=1}^{L} \sum_{\substack{q=1 \\ q \neq p}}^{L} \sum_{k=1}^{K} \frac{D_{KL}\left(P_{m_p}(c_{m_p,k}(v_i)) || P_{m_q}(c_{m_p,k}(v_i))\right)}{K} \tag{4.4}$$

Unfortunately, computing the full joint distribution is prohibitively expensive. Thus, instead we restrict the computation of the divergence to the sum of per-word divergences:

$$R(v_i) = \frac{1}{L(L-1)} \sum_{p=1}^{L} \sum_{\substack{q=1 \\ q \neq p}}^{L} \sum_{k=1}^{K} \frac{D(m_p, m_q, c_{m_p,k}(v_i))}{K} \tag{4.5}$$

where

$$D(m_p, m_q, c_{m_p,k}(v_i)) = \frac{\sum_{w=1}^{|c_{m_p,k}(i)|} D_{KL}\left(\mathcal{P}^{cond}(m_p, c^w_{m_p,k}(v_i)) || \mathcal{P}^{cond}(m_q, c^w_{m_p,k}(v_i)\right)}{|c_{m_p,k}(v_i)|} \tag{4.6}$$

is the per-word KL-divergence along the generation of the description $c_{m_p,k}(v_i)$ in each of the models. Compared to the likelihood method, this model gives a better estimate of the divergence of the distributions learned by the models of the ensemble. This measure is also independent of the sample length, and distribution perplexity, confounding factors when looking only at the likelihood of the samples.

4.2.2 Improving Diversity With Clustering

During the training of the initial active learning models, we noticed through a qualitative investigation that models seemed to be over-sampling parts of the training feature space. This was confirmed by running the experiments shown in Figure 4.3. To help combat this, we enforced a clustering-based diversity criterion. We first performed a k-means clustering of the training data using the mean (across the temporal dimension) of our visual features. We chose $K = N/20$ clusters, where N is the number of training samples in the dataset. See section 4.3 for a justification for this number of clusters. We then force the active learning algorithm to select at most ϕ samples from each cluster, which notably increases diversity. For the experiments in this work, we found $\phi = 3$ to be the best hyper-parameter value, out of $\phi = 1, 2, 3, \ldots 10$.

4.2.3 Comparison with Coreset Active Learning

While our method shares some significant similarities at a glance to Coreset (Sener and Savarese, 2018) (i.e. we both use delta-covers of a space to regularize the sampling), they have some notable differences. The Coreset method uses the distribution of the feature space,

combined with k-centers over the unlabeled data to select a set of samples which should be annotated. This is equivalent to finding a delta cover over the distribution of the data in the unlabeled space. Our proposed method (Ensemble Divergence + Cluster Regularization) uses the uncertainty of the underlying model to compute a score, and then attempts to regularize this score across the data space by enforcing that no two samples are too close together. Our method not only achieves better performance on our sequence learning tasks, but also runs notably quicker than Coreset, which can fail to solve the Integer Linear Program efficiently. It is interesting future work to explore selecting among Coresets using our uncertainty metric. Figure 4.1 directly compares Coreset and Greedy Coreset with our proposed model on the video description problem.

4.2.4 Models

The goal of this work is to explore active learning methods across multiple different model structures. In our experiments we use both a transformer-based model based on Zhou et al. (2018c), and the popular S2VT RNN architecture (Venugopalan et al., 2015) (See supplementary materials for details). Our models are able to achieve performance comparable to state-of-the-art models using vision-only features and without using policy gradients to optimize a downstream metric (Venugopalan et al., 2015; Aafaq et al., 2019). By adding multi-modal features, and direct REINFORCE optimization, you can gain 7-10 CIDEr points over our implementations (Aafaq et al., 2019). However, while there are more complex model pipelines, we chose two very simple architectures to demonstrate the efficacy of active learning, improve iteration time, and decrease the chance of confounding visual effects. We expect the presented methods to transfer to more complex optimization schemes, and powerful architectures given the flexibility of the formulation and our ablation results.

4.2.5 Datasets

We demonstrate the performance of our model on two common video description datasets, MSR-VTT (Xu et al., 2016) and the LSMDC (Rohrbach et al., 2017). While these methods may apply to video datasets generated using Automated Speech Recognition (HowTo-100M (Miech et al., 2019)) or dense captioning tasks (ActivityNet Captions (Krishna et al., 2017)), we focus on pre-clipped videos with high quality descriptive annotations. We refer the reader to the supplementary materials for a description of the datasets in use.

4.2.6 Experimental Setup

4.2.6.1 Feature Extraction and Pre-processing:

To avoid conflating the visual representations of the data with the performance of the captioning model, we follow video-captioning convention and pre-extract features from the videos using a Distill-3D (D3D) (Stroud et al., 2020) model pre-trained on the Kinetics-600

dataset for activity recognition. The videos are resized on the short-edge to 256px, then center-cropped to 256x256. They are down-sampled to 64 frames at 3 frames per second (with cyclic repetition for videos that are too short), and then passed through the D3D model to generate a 7x1024 representational tensor for the video which is used in the captioning process. The text data is tokenized using a sub-word encoding (Kudo and Richardson, 2018) with a vocabulary size of 8192.

4.2.6.2 Training:

Each model is trained in PyTorch (Paszke et al., 2017) with a batch-size of 512 for 80 epochs. We use the ADAM (Kingma and Ba, 2015) optimizer with a warm-up learning rate schedule with 1000 steps of warm-up, ranging from $1e^{-6}$ to $1e^{-3}$, then decaying over 250,000 steps to 0. We run our experiments using 8 Tesla T4 accelerators on Google Cloud Platform, making use of NVIDIA Apex[1] for mixed-precision fp-16 training.

4.2.7 Evaluation:

In all active learning methods, we begin by seeding the method with 5% of the data, chosen randomly. For a fair comparison, this random slice is shared across each of the active learning methods. We then train an initial classifier for use with our active learning method. When the classifier has converged (achieved minimum loss on a validation dataset, or trained for 80 epochs, whichever comes first), we use the classifier, and the proposed ranking methods (Using a cluster-limit $\phi = 3$, and a set of 8 sampled captions) to select an additional 5% of the training data. This happens 19 additional times (20 total evaluation points), until all of the training data has been selected. At each step, we run an evaluation of the model to determine the performance. Exploring the active learning process when using larger and smaller batches is interesting future work — when selecting very few examples, there is more potential benefit, but more computation required, selecting more samples requires less computation, but can be a more difficult task. Exploring ideas in continual learning, where the classifiers are re-initialized with weights from the previous active learning step is also interesting future work, however we found in practice that this does not heavily influence the training process.

During evaluation, we sample 8 candidate sentences with a temperature of 0.8, which are then evaluated using the COCO Captions Evaluation Tools (Chen et al., 2015). For tuning, we use a validation dataset sub-sampled from the training dataset, and we report the results on the official validation dataset (unseen during training/tuning) below. For ensemble-based metrics, we use the mean performance of the ensemble members. For non-ensemble based metrics, we perform multiple runs of active learning, and report the error as a 95% bootstrapped confidence interval. While the 95% is somewhat arbitrary, we present the full trajectories, for readers to explore.

[1] `https://github.com/NVIDIA/apex`

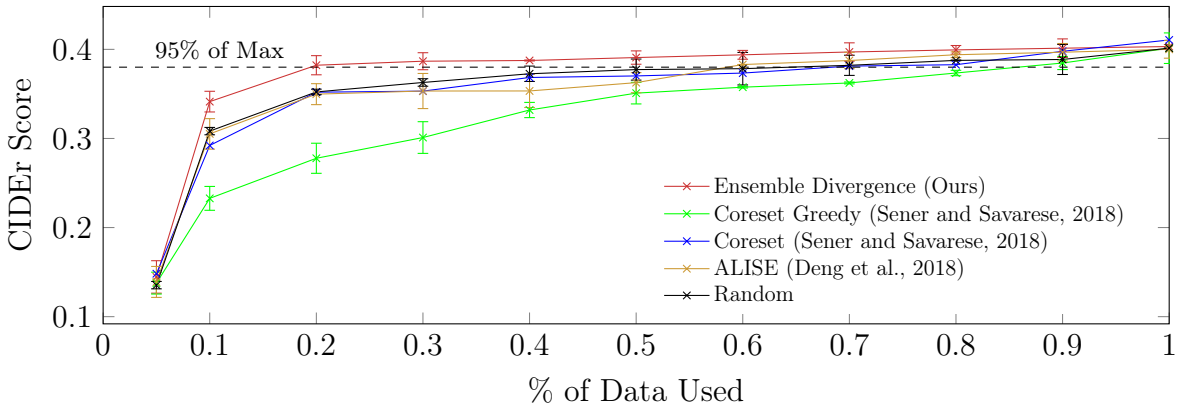


Figure 4.1: Validation performance of active learning methods on the MSR-VTT dataset using the CIDEr metric (Vedantam et al., 2015). Each run represents the mean of a bootstrap sample of ten runs. Our proposed method significantly outperforms all other methods, achieving 95% of the max performance while using only 25% of the data. This figure is measured 10 intervals instead of 20, due to the cost of Coreset's ILP solver.

4.3 Results & Discussion

Our key results using the transformer architecture on the MSR-VTT dataset are presented in Figure 4.1. Clearly, we can see that the clustered-divergence outperforms the benchmark models by a wide margin, using about 25% of the data to achieve a CIDEr score of 0.38 (95% of max). A full set of results is shown in Figure 4.2 for the methods from Section 4.2.1. Some additional qualitative results are presented in the supplementary materials.

Our method is highly prone to over-sampling parts of the distribution. To demonstrate over-sampling by examining the performance of our models across multiple clusters. Figure 4.3 shows that enforcing diversity is key to our approach: If we use no clustering, we actually fail to outperform random performance while adding a few clusters allows us to mitigate this effect and adding sufficient clustering allows for significant performance benefits. We can also see the effect of clustering by examining the mean distance to the validation set over the active learning iterations. We can also see from Figure 4.3 that the agreement method alone is unable to efficiently distribute across the validation set, however random and clustered methods achieve similar distribution effects. It's interesting to note, however, that even without the cluster enforcement the agreement metrics select from more visual diversity than the entropy/likelihood methods - leading to better performance (Table 4.1). The results for a cluster-regularized random selection method are given in Figure 4.2, however it is not significantly different from random alone, since the random method already samples uniformly from the set of input samples. Figure 4.4 shows that as we increase the number of ensemble members, the performance increased, however there are diminishing returns, as the models begin to capture the same data.

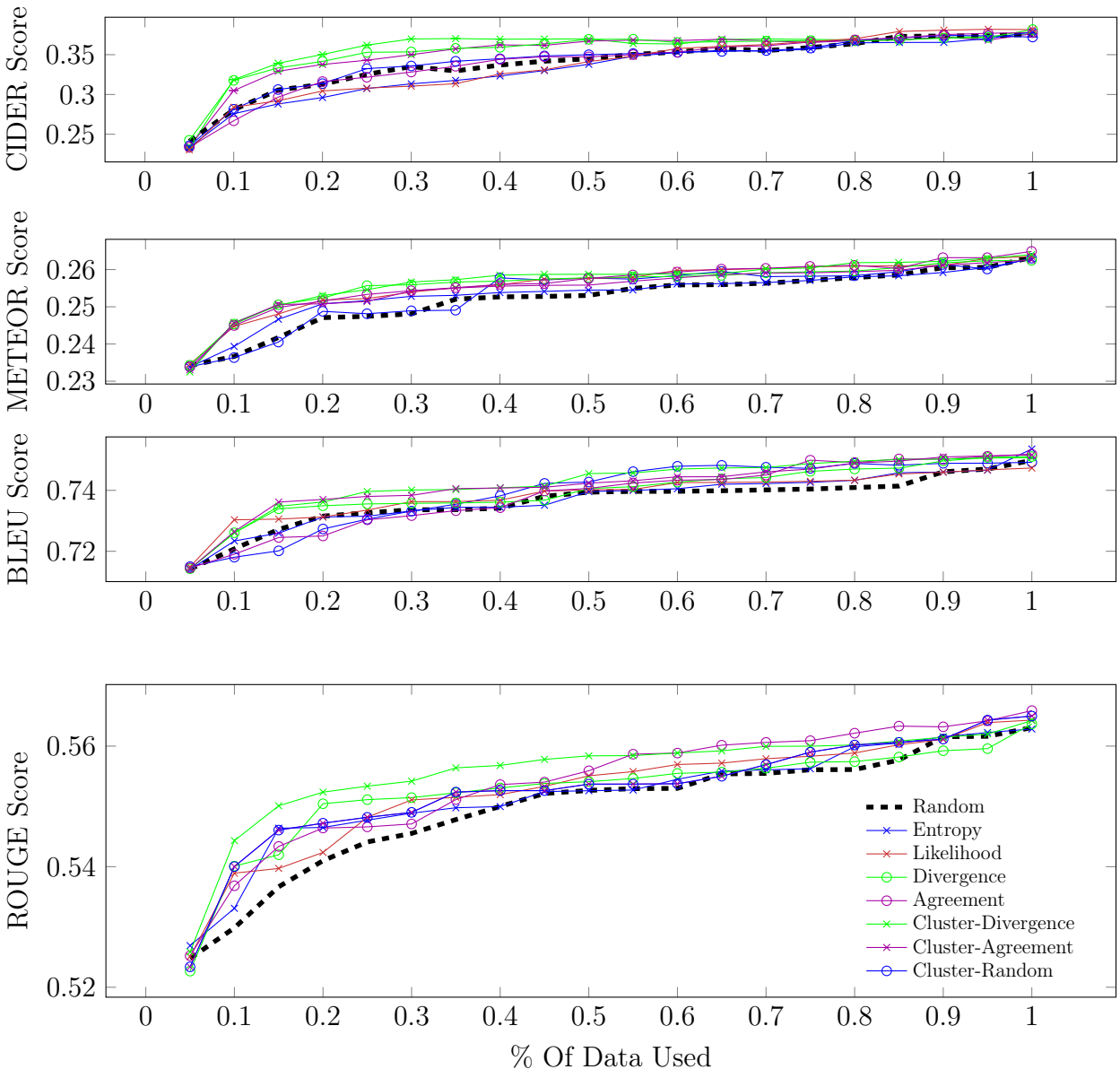


Figure 4.2: Validation performance across many potential active learning methods on the MSR-VTT dataset using the transformer model structure with respect to CIDEr Score (Vedantam et al., 2015), METEOR Score (Agarwal and Lavie, 2008), BLEU Score (Papineni et al., 2002) and ROUGE-L Score (Lin and Hovy, 2002). The curves presented are the means of 3 individual experiments using each method. Error bars are omitted for clarity. ALISE and Coreset are omitted due to computation time costs (However see Figure 4.1 for a comparison on CIDEr).

Table 4.1: Average number of clusters selected per iteration. The random and cluster-normalized methods select from a wider visual variety of samples, while the non-normalized samples select very few clusters on average.

Method	**Mean Number of Clusters Selected/Iteration**
Random Selection	195.47 ± 21.2
Cluster-Regularized Divergence	212.5 ± 14.4
Cluster-Regularized Agreement	202.3 ± 17.6
Cluster-Regularized Entropy	215.7 ± 12.8
Agreement Only	181.00 ± 16.9
Entropy	160.31 ± 16.4
Likelihood	169.25 ± 13.7

We can also see from Figure 4.2 that the ordering of methods can be dependent on the metrics chosen. While our proposed method outperforms all of the baseline methods, it is most helpful under the CIDEr and ROUGE metrics which prefer higher-level descriptions of the scenes. The method helps less for improving metrics that depend on lower-level semantics, such as BLEU and METEOR. We suspect that this is due to the influence of the active learning method on sampling a diverse set of samples - as increasing the sample diversity can help to improve high-level understanding of a scene, while perhaps having detrimental impacts on the language modeling process.

While we have made the case that a strong diversity of samples is required, it is also interesting to look at exactly which samples were selected. Figure 4.5 demonstrates some of the diversity of samples selected by our methods in comparison to the samples selected by the random method. We can see that the active learning method is sampling from a diverse set of elements from each cluster, while the random method is sampling a representative sample, but not necessarily the most relevant or useful videos.

One important thing to note is that because we are sampling from data that is in the initial training data for the two datasets, the results presented in this work may be an optimistic upper bound for the performance of an active learning tool. There is a significant amount of cleaning and curating that goes into these datasets which may or may not impact the final results of the classification, and the effort may be higher when annotating video in the wild. Future techniques may need to be developed for efficiently cleaning data, or curating samples that are relevant to captioning as a whole.

One downside to our experimental method is that our models do not achieve optimal performance in each training step, as the optimal hyper-parameters of the model change as more data is added. To ease this issue we use an adaptive training scheme which trains for more iterations than necessary with early stopping, however it is an interesting direction of future work to explore auto-tuning during the learning process to improve performance.

Our proposed method is not limited to the dataset or model. Figure 4.7 demonstrates the

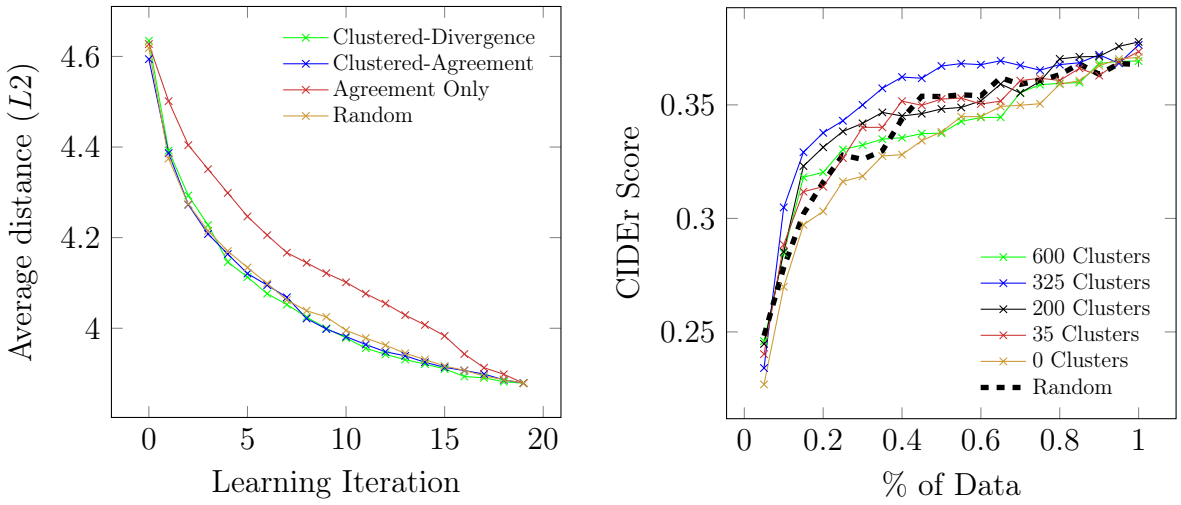


Figure 4.3: (Left) Average distance of validation samples to the nearest training sample over the active learning process. Models with improved diversity improve the distance to the training set more rapidly. We suspect this diversity is why random methods work well vs. non-diversity enforced methods as random methods contain a built-in coverage of the dataset. (Right) Performance of the cluster-divergence active learning method across different numbers of clusters. Performance is greater with greater numbers of clusters, until saturation, where performance regresses to random.

performance of our best method, clustered divergence, on the LSMDC dataset. We can see here that we achieve a CIDEr score of 0.121 (95% of max) with only 50% of data required by random sampling. Thus, we can see that the performance of the active learning method is not just limited to the MSR-VTT dataset. In addition, Figure 4.6 demonstrates that the performance is not limited only to our transformer based model. The S2VT model also improves, achieving a CIDEr score of 0.3219 with only 60% of data required by random selection.

In addition to requiring fewer data, our method can be significantly more efficient than the current state of the art methods. On our test-bench machine, we saw the following ranking times using the MSR-VTT dataset (Samples / Sec): Random: 2012.04, Entropy: 12.41, Cluster-Regularized Ensemble Ranking: 11.08, ALISE: 6.89, Coreset-Optimal: 0.11, and Coreset-Greedy: 11.89.

4.4 Conclusion

In this work, we have presented an initial set of methods aiming to tackle the active learning problem for video description, a challenging task requiring complex modeling where due to the complexity of the output distribution, many active learning methods are unable to

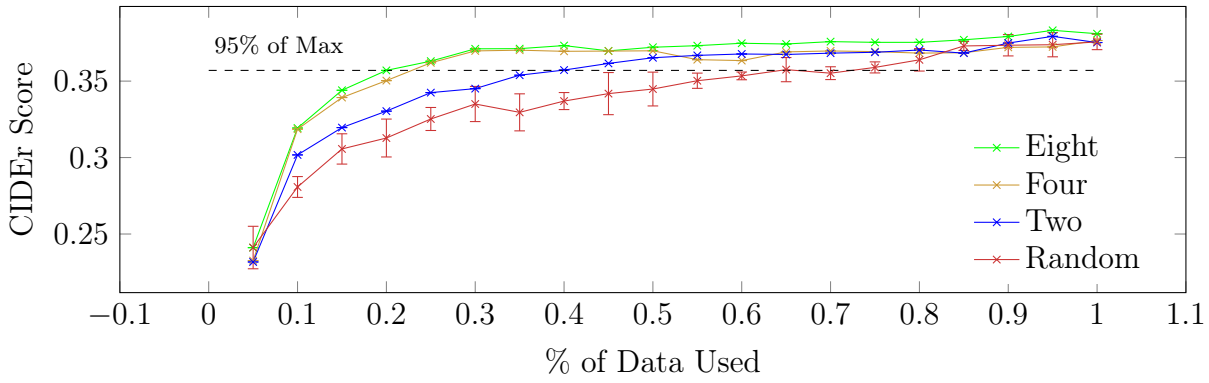


Figure 4.4: Validation performance with differing numbers of ensemble members on the MSR-VTT dataset. We see increasing the number of ensemble members leads to increased performance. We speculate that the diminishing returns are caused by independent models capturing similar information.

function efficiently, or at all. We have shown that we can achieve 95% of the full performance of a trained model with between 25% and 60% of the training data (and thus, manual labeling effort), across varying models and datasets.

While pairwise measures among ensemble members may be a good model of uncertainty, there are many such measures. Expected gradient variance methods such as (Huang et al., 2016a; Settles and Craven, 2008) are good candidates for future exploration. While such methods now do not account for the complexity of multiple correct labels, and dynamic entropy distributions, we may be able to compute high quality estimates. Such gradient methods may work in scenarios where the KL divergence between the final distributions of the models may be relatively low, but the evaluated sample has useful second-order gradient information.

It is also interesting, and likely fruitful, future work to explore different methods for clustering the elements of the training dataset. In many cases, we would like to enforce a subject-level diversity among the different inputs (as show by Figure 4.5), however visual similarity may not necessarily be the best metric to use for clustering. Using additional features to rank the diversity of the samples may provide better results, by increasing the individual diversity of each cohort more than k-means clustering in the visual space.

By exploring the applications of our work in practice, we can build robust active learning methods and collect large and effective datasets for video description. We hope these datasets will be used to improve the performance of downstream description tools in this complex and challenging labeling domain.

Figure 4.5: Visualization of four clusters of videos from the training dataset. Highlighted elements were selected by the cluster-divergence learning method (red), or the random method (yellow) in the first two iterations. In clusters with low visual diversity active learning selects fewer samples (top-left, bottom-left, bottom-right), while selecting more samples in clusters with high visual diversity (top-right), suggesting that the active method is choosing more informative samples.

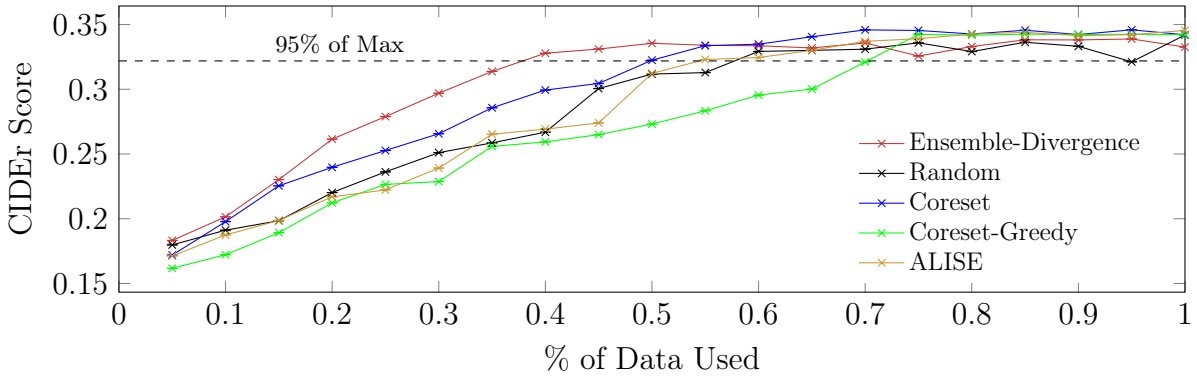


Figure 4.6: Performance using the LSTM model. While overall performance is lower, the clustered-divergence learning method can save more than 20% percent of the data.

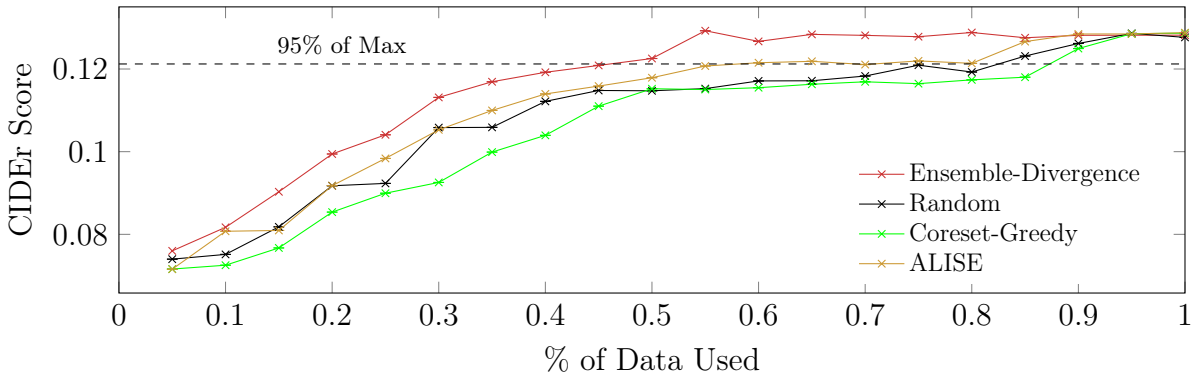


Figure 4.7: Validation performance for the LSMDC dataset. We achieve strong performance using almost 35% less data. We do not include Coreset, as it took > 24 hours per active-learning step to compute.

Chapter 5

Discussion

This section of the dissertation has provided a comprehensive analysis of two crucial elements in the development of conditional natural language generation models for visual descriptions. Initially, we explored the impact of linguistic diversity within datasets on the performance of these models. Through an in-depth examination of popular visual description datasets, we uncovered that the diversity of captions at various levels significantly influences the quality of generated descriptions, often leading to generic outputs. This finding underscores the importance of maintaining linguistic diversity in dataset collection to foster models capable of generating more informative and specific descriptions. Following this, our discussion on active learning strategies for video captioning highlighted an innovative approach to reducing the volume of training data required without compromising the model's performance. The cluster-regularized ensemble strategy, in particular, emerged as a highly effective method for assembling training sets, demonstrating the potential to enhance efficiency in model training processes significantly. Together, these insights not only illuminate the challenges and opportunities in improving natural language generation models but also propose actionable strategies for future research and development in the field, emphasizing the need for thoughtful dataset curation and innovative training methodologies to advance the capabilities of these models.

Emerging from both of these works is a key theme: *it is important to focus on the diversity of the underlying data*. In the work on linguistic diversity, we can see that models that are trained on more diverse datasets can lead to models that are more capable of downstream generalization. In hindsight, such a thought may seem obvious – of course training on more diverse data is important! However, it is important to understand what kinds of diversity are effective. From this work, we realize that we need to focus on not just "more" samples, but samples that cover a wider range of topics, and samples that require more decisions. I strongly believe that by increasing the required number of grounding choices (i.e. increasing the ED@N of the dataset), we can build models that generalize better to downstream tasks. We can already see this in practice in some recent work. The DALL-E 3 training procedure (Betker et al., 2023) works to generate better images by leveraging more complete captions (almost certainly requiring a higher ED@N), leading to better grounding in the model. Similar

work by Zhu et al. (2023b) has shown similar results, and even in our work (Lialin et al., 2023) we have seen that increasing the caption ED@N has led to more grounded, less hallucinatory models.

It further seems that beyond just increasing diversity, increasing diversity in a targeted way is important. In chapter 4, we saw that retaining only one video from a particular cluster is important, as well as only retaining samples that lead to highly diverse outputs (i.e. have large QBC disagreement). This further supports the idea that not just arbitrary diversity, but some notion of "correct" or "useful" diversity is required. It remains interesting for future work to further look into the sources of diversity in captioning data, and determine what kinds of diversity are "useful" for models, and what kinds of diversity lead only to noise and inefficiency in the training procedure.

Going beyond diversity, many other questions remain unanswered about data context in the captioning and ASR domains. One that I find particularly interesting is: *What kinds of data lead to what downstream capabilities in the models?* Recent work has begun to show that certain types of data lead to particular downstream behaviors, for example, code pre-training in LLMs leads to better performance on arithmetic tasks or conversational ability is largely driven by Reddit data (Touvron et al., 2023; Zhou et al., 2024). Diving into this question has the ability to impact how we train our models, and how we collect the data that drives our LLM alignment.

Such a deeper focus on data not only impacts how we train the models, but it also has the ability to start uncovering the sources of hallucination: *why do models assign a high probability to things that aren't present in the context?* Hallucination is one of the biggest challenges driving LLM adoption and one of the most mystifying issues: if there's no particular stimulus for an output, why are such outputs naturally generated? Diving deeper into the datasets will help us to understand this. Perhaps hallucinations are an artifact of spurious correlations in the data, or perhaps they are caused by the outsize influence of language priors. Perhaps they are caused by a lack of data coverage. Without a deep understanding of the connection between datasets, data distributions, and model performance, we will likely never know.

As we navigate the complex landscape of conditional natural language generation models, the critical role of data diversity and innovative training methodologies remains at the forefront of advancing the field. By embracing a nuanced understanding of dataset composition and leveraging strategic data curation, we can pave the way for models that not only generate more accurate and grounded descriptions but also exhibit a deeper understanding of the nuances of human language. This journey towards refining natural language generation models underscores a broader commitment to pushing the boundaries of AI research, fostering models that can interact with the world in increasingly sophisticated and meaningful ways.

Part II

Building

To know an object is here to lead to it through a context which the world supplies.

William James

As we discussed in chapter 1, no task exists in a vacuum. There are always task-specific and even human factors that impact the text that our models generate. In the last section, we discussed the most important contextual clue, the underlying dataset, and the task, however, there are many different kinds of context that can impact the downstream results! Take, for example, the automatic speech recognition task. Many people may argue that all that is necessary for ASR is the waveform itself. But it is clear that when deciphering what audio has been uttered, the probabilities for any particular word depend not only on the waveform itself but the context surrounding that waveform. What did the person say right before they made this utterance? What is on-screen (if the audio is tied to a video)? What does the user's contact list look like? What is the user trying to accomplish? All of these implicit signals can impact the likelihood of an output. Not only this but in some languages, such context is necessary for disambiguation. For example, in Japanese, the utterance あか (AH-kah) can be transcribed as 赤, 紅, 朱, or 緋 depending on the specific shade of red that the person is referring to. Disambiguating this utterance requires additional context: perhaps the person was talking about the color of a pink rose or the color of blood. Not only does the world-state around the user matter, but the user themselves matters. If given the image in Figure 1.2, depending on who it is, a person might describe that image in a thousand different ways, each way capturing a different aspect of interest, or of focus to that person. In this section of the dissertation, we describe several *technical* approaches for incorporating these kinds of context directly into models, focusing on the domains of interest in automatic speech recognition and image/video description.

First, in chapter 6, we discuss how if you ask a human to describe an image, they might do so in a thousand different ways but image captioning models, on the other hand, are traditionally trained to generate a single "best" (most like a reference) caption. Unfortunately, this process encourages captions that are informationally impoverished: Such captions often focus on only a subset of possible details, while ignoring other potentially useful information in the scene. We then introduce a simple, yet novel, method: "Image Captioning by Committee Consensus" (IC^3), designed to generate a single caption that captures details from multiple viewpoints by sampling from the learned semantic space of a base captioning model, and carefully leveraging a large language model to synthesize these samples into a single comprehensive caption. Our evaluations show that humans rate captions produced by

IC^3 more helpful than those produced by SOTA models more than two-thirds of the time, and IC^3 improves the performance of SOTA automated recall systems by up to 84%, outperforming single human-generated reference captions and indicating significant improvements over SOTA approaches for visual description.

Next, we turn our attention to automatic speech recognition (ASR). In chapter 7, we first look at how we can learn to perform ASR tasks leveraging context drawn from video data. While traditionally, research in automated speech recognition has focused on local-first encoding of audio representations to predict the spoken phonemes in an utterance, such approaches relying on such hyper-local information tend to be vulnerable to both local-level corruption (such as audio-frame drops, or loud noises) and global-level noise (such as environmental noise, or background noise) that has not been seen during training. In this chapter, we introduce a novel approach that leverages a self-supervised learning technique based on masked language modeling to compute a global, multi-modal encoding of the environment in which the utterance occurs. Then, using a new deep-fusion framework to integrate this global context into a traditional ASR method, we demonstrate that the resulting method can outperform baseline methods by up to 7% on Librispeech; gains on internal datasets range from 6% (on larger models) to 45% (on smaller models).

Beyond video, we can also learn from other forms of external context to efficiently build ASR models. In chapter 8, we investigate the potential of leveraging external knowledge through off-policy generated text-to-speech key-value stores, to allow for flexible post-training adaptation to new data distributions. In our approach, audio embeddings captured from text-to-speech are used, along with semantic text embeddings, to bias ASR via an approximate k-nearest-neighbor (KNN) based attentive fusion step. Our experiments on LibiriSpeech and Amazon Alexa voice assistant/search datasets show that the proposed approach can reduce domain adaptation time by up to 1K GPU-hours while providing up to 3% WER improvement compared to a fine-tuning baseline, suggesting a promising approach for adapting production ASR systems in challenging zero and few-shot scenarios.

Even though in chapter 8, we use a large catalog of text, it's also possible to just use a small set of local recent history to learn context hints. In chapter 9, we introduce CLC: Contrastive Learning for Conversations, a family of methods for contrastive fine-tuning of models in a self-supervised fashion, making use of easily detectable artifacts in unsuccessful conversations with assistants. We demonstrate that our CLC family of approaches can improve the performance of ASR models on OD3, a new public large-scale semi-synthetic meta-dataset of audio task-oriented dialogues, by up to 19.2%. These gains transfer to real-world systems as well, where we show that CLC can help to improve performance by up to 6.7% over baselines.

In summary, this section begins a conversation regarding how context, both external and intrinsic to the user, can significantly enhance the capabilities of computational models across a range of tasks, including automatic speech recognition and image/video description. Here, we aim to demonstrate the critical role context plays in improving model performance, making models more adaptable, accurate, and reflective of the complex nuances of real-world interactions and through these demonstrations, we aim to not only push the boundaries

of current state-of-the-art models but to deepen our understanding of how context can be effectively integrated into computational systems for natural language generation.

Previously Published Works Appearing In This Section:

1. Chan, David M., et al. "IC3: Image Captioning by Committee Consensus" Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

2. Chan, David M., et al. "Multi-modal pre-training for automated speech recognition." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

3. Chan, David M., et al. "Using External Off-Policy Speech-To-Text Mappings in Contextual End-To-End Automated Speech Recognition." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

4. Chan, David M., et al. "Task Oriented Dialogue as a Catalyst for Self-Supervised Automatic Speech Recognition." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

Chapter 6

IC3: Image Captioning by Committee Consensus

Generating a high-quality description of an image is not only an open research problem, but it is also a challenging task for humans (Lin et al., 2014; Sharma et al., 2018; Young et al., 2014). Image captioning datasets usually acknowledge this fact; rather than providing a single gold standard caption for each image, they instead rely on several human annotators, each with their own personal biases and contexts, to provide multiple captions for each image, hoping that the set of collected captions collectively captures all of the relevant semantic information.

While a set of image captions can be useful, many applications, such as alt-text generation, demand a single succinct sentence that summarizes the information present in the image. This "summarized" caption usually takes a different structural form compared to the "single-viewpoint" captions sourced from crowd workers that make up the datasets. While single-viewpoint captions may contain a subset of the relevant information in an image, it is unlikely that they contain everything (MacLeod et al., 2017; Stangl et al., 2020).

Unfortunately, while the development of large vision and language models (VLMs) has led to progress on a variety of tasks including image captioning, models are trained to produce samples from the reference distribution of a captioning dataset such as MS-COCO (Li et al., 2022; Wang et al., 2022b; Alayrac et al., 2022; Yu et al., 2022; Chen et al., 2022). While not inherently flawed, this approach reproduces the dataset's single annotator viewpoint captions, containing some, but not all, of the semantic information present in the image. Thus we seek to answer the question: "How can we combine many single-viewpoint captions into a collective summary of the image containing the relevant semantic information?"

One way to obtain a more comprehensive caption, given a set of single-viewpoint captions from annotators, would be to have another human expert consider the set of captions from the committee of individual annotators, and create a new caption that combines complementary information while filtering out any syntactic or semantic errors. Motivated by this idea, we propose the Image Captioning by Committee Consensus (IC^3) approach, which utilizes off-the-shelf VLMs in conjunction with large language models (LLMs) to generate higher quality

Figure 6.1: In the IC^3 (Image Captioning by Committee Consensus) method, we first leverage standard image captioning models to generate descriptions covering a range of content within the image, similar to how human raters describe the image from independent and unique points of view. We then summarize the group of captions using a vision-free summarization model into a single, high-quality description of the image, suitable for use in visual description applications.

captions than would be possible with VLMs alone. Our key contributions are summarized as follows:

1. We introduce IC^3, a simple, yet novel, approach leveraging pre-trained image captioning models and large language models to generate semantically complete image captions from a collection of "single-viewpoint" captions.

2. We perform an extensive human evaluation of our method which demonstrates that human raters rate image captions generated by IC^3 higher in both helpfulness and correctness.

3. We demonstrate through several automated measures that captions generated using IC^3 contain significantly more semantic information than baseline captions. Notably, CLIP-recall with IC^3 captions can be improved by 22-84%, with improved coverage of objects and actions in the image.

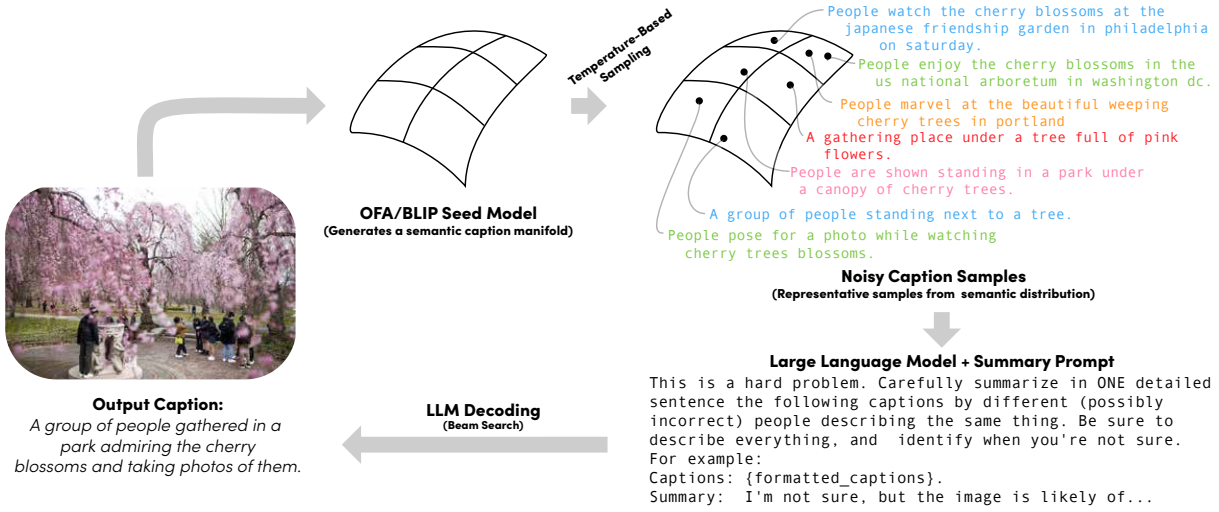


Figure 6.2: The IC^3 approach. Every captioning model defines a distribution across a caption semantic space. This distribution is unlikely to be unimodal, thus, while maximum likelihood decoding approaches such as beam search will capture a local maximum, this point is not likely to be representative of the full distribution of captions. Instead, IC^3 first generates a representative sample of captions from the semantic manifold using temperature-based sampling. This set naturally captures any means as well as the variance of semantic information in the image. Because this group of captions can be large, hard to parse, noisy, or incorrect, we use a large-scale language model, such as GPT-3, paired with prompt engineering, to summarize and filter the noisy group of captions. The resulting captions are more detailed and often more useful than captions generated by beam search alone.

6.1 Related Work

The idea that captioning models tend to produce single-viewpoint captions has been prevalent in the image captioning community for many years under several names (Wang and Chan, 2019). Notably, research has focused on quantifying and improving the “diversity" of output captions, including specific methods (Klein et al., 2022; Aneja et al., 2019; Dai et al., 2017; Mahajan et al., 2020; Mahajan and Roth, 2020; Wang et al., 2017, 2016) and metrics (Holtzman et al., 2020; Zhu et al., 2018; Wang et al., 2020; Shetty et al., 2017; Deshpande et al., 2019; Chan et al., 2022d; van Miltenburg et al., 2018; Wang and Chan, 2019). As an alternate approach to increasing and quantifying diversity, some methods (Gan et al., 2017; Yang et al., 2020b; Zha et al., 2019; Fang et al., 2022) have focused on explicitly modeling the variance in the caption space, and introduced human, or statistical controls to reduce the variance, turning the multi-modal problem into several uni-modal problems. While these methods are effective at describing the same image multiple times from multiple perspectives, they have not demonstrated an effective approach that generates a single caption covering all of the information in each of the diverse captions.

Dense captioning methods (Johnson et al., 2016; Yang et al., 2017; Li et al., 2019b; Yin et al., 2019; Kim et al., 2019) attempt to generate a full description of all of the objects in the image, however, dense image captions are long and unwieldy, and often contain redundant or repetitive information. Similar long-form captions have been explored in the form of paragraph captioning (Zha et al., 2019; Krause et al., 2017; Liang et al., 2017; Mao et al., 2018; Chatterjee and Schwing, 2018; Luo et al., 2019), however in all cases, no efforts have explored using additional post-processing or models to distill the relevant information for alt-text or for downstream applications. In this work, we explore beyond single-view captioning and move towards captions that are short, succinct summaries of the full visual context.

A natural way of summarizing and filtering a dense caption, paragraph caption, or set of captions, is with a pre-trained model for summarization. While end-to-end methods for abstractive and extractive text summarization exist (Allahyari et al., 2017), recently, large language models (LLMs) such as GPT-3 (Brown et al., 2020), LAMDA (Thoppilan et al., 2022) and PALM (Narang and Chowdhery, 2022) have demonstrated remarkable zero-shot performance when performing language-only summarization tasks (Brown et al., 2020; Liu et al., 2021b; Goyal et al., 2022b; Chintagunta et al., 2021; Kieuvongngam et al., 2020), so it is natural that they would be capable of summarizing multimodal information in a zero-shot way. Indeed, recently, large-scale vision and language models (VLMs) and large-scale language-only models (LLMs) have revolutionized a number of sub-fields in AI, including in the image captioning space. Models such as BLIP (Li et al., 2022), OFA (Wang et al., 2022b) and Flamingo (Alayrac et al., 2022) have all demonstrated strong performance in single-view image captioning tasks, and indeed, many of these approaches are rated as good or better than human users in some evaluations.

Surprisingly, vision-blind LLMs have also become particularly prevalent in multimodal image/language spaces, primarily using a language-only prefix generated by a set of pre-trained tools. Mokady et al. (2021) explores using a continuous embedding as a prompt for a GPT-style language model and demonstrate strong single-viewpoint image captioning performance, while Hu et al. (2022) and Tiong et al. (2022) leverage natural language prompts along with GPT to achieve SOTA performance on visual question answering.

Closest to our approach are (Zhu et al., 2023a) (developed concurrently with the proposed work) and Zeng et al. (2022). Zeng et al. (2022) leverages a CLIP-based model to extract key tags from the image, and then uses GPT-3 along with a specialized prompt to generate a stylized image caption, in an attempt to emulate the Socratic method. Zhu et al. (2023a) further employs the Socratic method by employing Chat-GPT and BLIP-2 (Li et al., 2023b) to ask and answer questions about the image, respectively. Finally, Chat-GPT summarizes the QA transcript into the image description. Our proposed approach primarily differs from Zhu et al. (2023a) and Zeng et al. (2022) in the method of visual data extraction. Instead of using the Socratic method, which requires repeated high-quality questioning and high-quality VQA models to elicit data, or imprecise image tagging models, our approach relies on existing image-captioning models augmented with temperature based sampling, which are able to generate a diverse set of (possibly noisy) information about the image from multiple sampled viewpoints. This avoids a repetitive (and computationally expensive) QA loop, which with

imperfect models can not only introduce significant noise, but also can fail to uncover detail outside the questioning distribution. Also related to our work is Xie et al. (2022), which uses similar tags to generate a paragraph-caption, but does not explore filtering the image, or using existing caption distributions.

6.2 IC3: Image Captioning by Committee Consensus

In this work, we introduce a simple framework for visual description, based on a committee generation then summarization process, which we call "Image Captioning by Committee Consensus" (IC^3). The approach consists of two stages. In the first stage, we leverage a standard pre-trained image captioning model to sample several (potentially noisy) captions using temperature-based sampling from the caption distribution. This generates a representative samples from the caption distribution, each possibly describing different parts of the image. In the second stage, we leverage a summarization model to summarize the information in each of the captions in a short and succinct way that can be presented to a user. An overview of our method is given in Figure 6.2.

The goal of IC^3 is to generate an output caption from a given image, by first sampling from a frozen image captioning model, and then summarizing these generated captions into a single "summary caption". More formally, given an image I, we aim to produce a sequence of m tokens $x_1 \dots x_m$ describing the image. Formally, an image captioning model can be described as a function $\mathcal{M}$ which takes I and a set of tokens $a_1 \dots a_{k-1}$ in some vocabulary V, and produces a probability distribution $P(a_k \in V | I, a_1 \dots a_{k-1})$, the probability of the next token in the sentence given all previous tokens and the image.

Traditionally, image captioning models generate a caption C, where:

$$C = \arg\max_{a_1 \dots a_{k \leq N}} \prod_{i=1}^{k} P(a_i | a_1 \dots a_{i-1}, I) \tag{6.1}$$

Finding the argmax is particularly challenging, as it is an optimization over all possible sequences. Usually, to avoid these challenges, a technique such as beam search (Li et al., 2016; Shen et al., 2017) is used to reduce the number of possible candidates. Recently, however, it has been shown by several papers, including Chan et al. (2022c) and Caglayan et al. (2020) that captions generated using beam search contain only the mutual information between the references, and that such captions are often bland, and uninteresting. To avoid this, we instead take a different approach. Instead of maximizing the likelihood, we generate a set of samples, $K = \{k_1 \dots k_i\}$, from the model using temperature-based sampling of the distribution:

$$k_i = a_1 \dots a_{m_i} \propto \exp\left(\frac{\log P(a_1 \dots a_{m_i} | I)}{T}\right) \tag{6.2}$$

where T is a temperature parameter. At temperature 1, the resulting samples $K = k_1 \dots k_i$ are an unbiased estimate of the distribution of reference captions. This means that, unlike

the maximum likelihood estimate caption, the sampled captions will contain variance in information commensurate with the variance of human descriptions.

Unfortunately, while the caption set K is a good description of all of the details that we might care about in the image, a set of captions can be hard to parse for a downstream user. Thus, we would like to summarize the set K as a single caption, by removing any redundant (or incorrect) information, and combining any details mentioned in one description, and not in one of the others. To do this, we leverage a summarization model $\mathcal{S}$, which maps the set of captions K to our final single output caption C. In IC^3, the summarization model is visually blind - that is, the image is not taken into account at all during the summarization phase. We discuss our choice of summarization model in subsection 6.2.2. In our work, C is generated using beam-search from the summarization model, giving us a maximum likelihood estimate of the best summary of the input captions.

6.2.1 Image Captioning Models

The first stage of the method is to sample a set K of candidate captions from an underlying pre-trained image captioning model, $\mathcal{M}$. In this work, we explore two underlying image captioning engines, the BLIP model (Li et al., 2022), and the OFA model (Wang et al., 2022b), which both represent high-quality image-captioning models pre-trained on large-scale image and language data, then fine-tuned for captioning on the MS-COCO dataset (Lin et al., 2014). More details on the specific image captioning models can be found in Appendix B.3.1.

Temperature Selection: We want to generate a sample of captions that lies as close to the reference distribution as possible. For most models, this will be at or close to temperature 1. To validate this, we use the TRM-CIDEr metric introduced in Chan et al. (2022d) to measure the distance between the reference distribution and the generated captions at a set of temperatures between 0 and 2. We found that for the BLIP model, the optimal temperature was 1.15, and for the OFA model, the optimal temperature was 0.95.

Selecting size of K: To select the number of captions that are generated, we used a small validation set and found that 10 captions represented a good trade-off between the length of the prompt, and the quality of the model. Sampling larger numbers of candidate captions can improve the total captured information but can decrease the performance of the summarization model, and be more costly to evaluate (See Appendix B.2.3 for an ablation).

6.2.2 Summarization Models

The choice of the summarization model $\mathcal{S}$ is a key decision when implementing IC^3, as the model should be capable of high-quality zero-shot abstractive summarization. We found that using a large language model (LLM) for zero-shot summarization is effective in generating high-quality output summaries of the input captions (See Appendix B.2.1).

For the results in section 6.3, we use GPT-3 (Brown et al., 2020), as it produced strong abstractive summaries of the candidate captions, however we discuss (and explore) in detail the choice of summarization model Appendix B.2.1.

6.2.3 Prompt Selection

To use a large-scale language model for summarization, we follow the techniques in Brown et al. (2020), and design an input text prompt passed to the language model, to encourage the generation of a correct output. The choice of prompt is defined by several motivations, including encouraging the model to summarize the information in the sampled captions, particularly the uncertainty of the captions, encouraging the model to convey this uncertainty in a natural manner (if the model is unsure about what is in the scene, it should identify when this is the case) and making sure that the generated caption is comprehensive, and contains all of the unique viewpoints presented in each of the sampled descriptions. Our final prompt used for experiments with GPT-3 was:

`This is a hard problem. Carefully summarize in ONE detailed sentence the following captions by different (possibly incorrect) people describing the same scene. Be sure to describe everything, and identify when you're not sure. For example: Captions: {formatted captions}. Summary: I'm not sure, but the image is likely of...`

Encouraging and surfacing uncertainty: In our prompt design, we aim to encourage the model to account for potential uncertainty/noise in the sampled captions. In many cases, high disagreement among the captions can indicate uncertainty in the language distribution, so we encourage the model to identify when the individual captions differ using language such as `possibly incorrect`, `is likely of` and `I'm not sure` in the prompt. The effect of encouraging and surfacing uncertainty is demonstrated in Table B.2 in the appendix. This shows that choosing this language significantly increases the likelihood that models generate uncertain language, and that such captions are rated as more correct on average by human raters.

This is a hard problem: Following Kojima et al. (2022), who showed that adding short interrogative/instructive sentences to the beginning of a prompt can improve zero-shot performance, we also add the short sentence "this is a hard problem". We found that this generally improved the quality of the model by a small amount, with diminishing returns as the quality of the candidate captions improved as seen in the ablation in Table B.3.

Use of capitalization: In our exploration of the prompt space, we found that in some cases, the models choose to generate long concatenations of the input captions instead of generating a single short and concise sentence. Thus, to help alleviate this issue, we found that

capitalizing the “ONE" when asking for sentences encouraged the GPT models to produce shortened captions, usually consisting of a single sentence (reducing the average caption length from 120.03 to 107.89 characters).

Style Transfer and Contextual Captions: In addition to the above goals, it is interesting future work to explore how the prompt can be used to shape the generated output caption, either for zero-shot transfer to other languages, or to guide the generation of the text to pay attention to specific details in the image. While we do a cursory exploration of such an approach in Appendix B.7 and Appendix B.8, future work in this area is essential.

6.2.4 Evaluation

N-gram matching scores such as CIDEr (Vedantam et al., 2015) do a poor job at comparing distributions of texts. An example of this is a single caption which is the concatenation of two non-overlapping references. Because for each reference, there exist n-grams in the candidate that do not overlap with that reference, the candidate will score poorly. However the candidate has the same (or more) information than either of the two original reference sentences alone. Thus, along with extensive human evaluation, we introduce two novel automated measures of caption quality, which directly address information retrieval.

CLIP Recall: One measure of the quality of a caption is its ability to distinguish between images in a dataset (Hessel et al., 2021a; Wang et al., 2022a). In this work, we leverage CLIP (Radford et al., 2021b) as a recall model, and use it to estimate an approximate specificity of each of the captions generated by our approach. Specifically, for each image i, we compute the CLIP embeddings $\mathcal{I}_i$ and the corresponding caption $\mathcal{C}_i$. We then compute the CLIP-Score (Hessel et al., 2021a) between $\mathcal{I}_i$, and every other generated caption $\mathcal{C}_j$, and from this, compute the mean reciprocal rank (MRR) of the caption $\mathcal{C}_i$, and the recall at 1, 5 and 10. High values of MRR suggest that the captions are more discriminative within the test set (and thus, are often more detailed in nature).

Content Coverage: In addition to the specificity of the individual caption, we also would like to measure how much of the total information provided by *all* of the references is included in the summarized caption. To do this, we first compute the caption $\mathcal{C}_i$ for each image and fetch the references $\mathcal{R}_i^j, 1 \leq j \leq N$ for each image. Let $\mathcal{N}(\mathcal{C}_i)$ be the set of nouns in a caption, $\mathcal{C}_i$, and $\mathcal{V}(\mathcal{C}_i)$ be the set of verbs. Let

$$I_{\mathcal{N},i}(n) = \begin{cases} 1, & \text{if } n \in \cup_{j=1}^{N} \mathcal{N}(R_i^j) \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

We compute exact noun overlap for $\mathcal{C}_i$ as:

$$\text{Noun Overlap} = \frac{1}{|\cup_{j=1}^{N} \mathcal{N}(R_i^j)|} \sum_{n \in \mathcal{N}(\mathcal{C}_\rangle)} I_{\mathcal{N},i}(n) \tag{6.4}$$

Verb overlap is defined analogously for $\mathcal{V}$. We compute fuzzy overlap similar to exact overlap, however instead of Equation 6.3, we use:

$$I_{\mathcal{N},i}(n) = \begin{cases} 1, & \text{if } ||E(n)-E(x)||_2^2 \leq \phi, x \in \cup_{j=1}^{N} \mathcal{N}(R_i^j) \\ 0, & \text{otherwise} \end{cases} \tag{6.5}$$

where E is a word-embedding function (we use embeddings from the Spacy package (Honnibal et al., 2020)), and $\phi = 0.1$ is a threshold.

Human Evaluation: To test the performance of our model in real-world conditions, we leverage human evaluations on the Amazon Mechanical Turk platform. We perform two styles of human evaluation. In "context-free" evaluation, raters are asked to rate two components: The "Helpfulness" of the caption to somebody who cannot see the image (on a scale of 0 to 4), and the factual "Correctness" of the caption (on a scale of 0 to 5). In "head-to-head" evaluation, raters are presented with two captions and asked which is more "helpful" for somebody who cannot see the image, as well as which is more factually "correct". Full details on the exact questions asked and the experimental design are given in Appendix B.5.

Reference Baseline: Because IC^3 can be used to augment any existing captioning method, we also explore augmenting the human reference captions with IC^3. To do this, we use the reference captions (REF) as candidates for the summary pipeline, which are then summarized by the LLM to generate the REF+IC^3 caption. Such an approach removes the additional variance introduced by the candidate captioning model and demonstrates the potential of IC^3 applied to a near-perfect captioning approach.

6.3 Results & Discussion

In this section, we compare captions generated using our baseline seed image captioning models, BLIP (Li et al., 2022), BLIP-2 (Li et al., 2023b), and OFA (Wang et al., 2022b), to captions generated using IC^3. We leverage two image captioning datasets for evaluation: MS-COCO (Lin et al., 2014) and the Flickr-30K dataset (Young et al., 2014) (see Appendix B.3.2).

Figure 6.3 and Appendix B.6 give some qualitative examples of our method compared to several baseline methods. We can see that descriptions using IC^3 are often longer, and contain more detail than their counterpart baseline models. Further, most display uncertainty about the content of the image in a natural way, which the baselines are not able to accomplish (see Appendix B.2.2).

6.3.1 Human Evaluation

Recent works (Chan et al., 2022d; Caglayan et al., 2020) have confirmed that human evaluation remains the gold standard for visual description evaluation, despite progress in

OFA + IC³: *A layered cake with chocolate and cream on top, sitting on a red plate, possibly with a knife sticking out of it, although it could also be an ice cream cake, a pancake, a crepe cake, or a stack of pancakes.*
OFA: *A cake sitting on top of a red plate.*
Socratic Models: *A variety of sweet treats on display at a deli.*
Human: *A plate that has a dessert on it.*

OFA + IC³: *A young man wearing an orange jacket and hat performing tricks on a skateboard at a skate park or enclosed pool at night.*
OFA: *A man riding a skateboard up the side of a ramp.*
Socratic Models: *A peaceful scene of a skate park in the city.*
Human: *A person doing a skateboard trick up a bowl*

OFA + IC³: *Two small figurines, possibly Star Wars related, sitting at table in getting ready to eat a pizza.*
OFA: *A couple of figurines sitting on top of a table next to a pizza.*
Socratic Models: *A group of people enjoying a delicious pizza at a local pizzeria.*
Human: *Two toys are sitting with a toy cup and real pizza.*

Figure 6.3: Some qualitative examples of IC^3 paired with the OFA model. We can see in all of these cases that OFA + IC^3 surfaces significantly more detail about the image, including both foreground and background details, as well as maintains the syntactic quality, relevant world-knowledge, and high-level details of the maximum-likelihood caption.

automated evaluation of image captioning. As discussed in subsection 6.2.4, we perform two experiments: head-to-head experiments and mean opinion score evaluation. The results of the head-to-head experiments on MS-COCO are shown in Table 6.1, where we can see that IC^3 augmented models significantly outperform the baselines on both helpfulness and correctness (Helpfulness: OFA + IC^3 vs. OFA, $p = 0.0008$; BLIP + IC^3 vs. BLIP, $p = 0.008$; BLIP2 + IC^3 vs. BLIP2, $p = 0.003$; REF + IC^3 vs. REF, $p = 1.73e^{-5}$. Correctness: OFA + IC^3 vs. OFA $p = 0.0428$; BLIP + IC^3 vs. BLIP, $p = 0.0280$; BLIP2 + IC^3 vs. BLIP2, $p = 0.898$; REF + IC^3 vs. REF, $p = 0.0019$; $n = 89$).

Table 6.2 shows the performance of IC^3 in terms of mean opinion score, and demonstrates that even in a calibration-free setup, where no extra evidence is presented, IC^3 methods significantly outperform their baseline counterparts when rated for helpfulness (Helpfulness: OFA + IC^3, $p = 0.0237$; BLIP + IC^3, $p = 0.0419$; REF + IC^3, $p = 0.0293$; $n = 121$). Numerically, IC^3 outperforms baselines on the correctness measure, however we found in all three cases that the difference was not statistically significant. We believe the the reduction in margin is caused by several effects: (1) without a point of reference for the potential quality of the captions, AMT workers cannot tell which captions are deserving of high scores and (2) both OFA and BLIP are strong captioning models, so a random sample of MS-COCO images may not contain difficult images that separate the two methods.

To investigate this hypothesis, we ran several additional human studies on a set of challenging examples, which we call the Hard MRR splits (see Appendix B.3.2), which contain the 200 most challenging images for CLIP to recall. We show the head-to-head experiments in Table 6.3, and see that once again, in head-to-head experiments, IC^3 significantly outperforms baseline methods (OFA + IC^3, $p = 0.0225, n = 28$, BLIP + IC^3, $p = 0.0074, n = 52$). In

Table 6.1: Head-To-Head human evaluation performance of models augmented with IC^3 on the MS-COCO dataset. Table shows % of instances preferred by users.

Model	Helpfulness ↑	Correctness ↑
Blip-2 + IC^3	**51.97%**	**44.10%**
Blip-2	37.49%	42.67%
Tie	9.78%	11.6%
Blip + IC^3	**52.05%**	**42.90%**
Blip	33.44%	36.28%
Tie	14.51%	20.82%
Ofa + IC^3	**52.91%**	**48.93%**
Ofa	32.72%	33.94%
Tie	14.37%	17.12%
Ref + IC^3	**55.79%**	**48.80%**
Ref	36.65%	36.27%
Tie	7.46%	13.97%

Table 6.2: Human rater mean opinion score for IC^3 on MS-COCO. Helpfulness (H, 0-4), Correctness (C, 0-5).

Candidate Generator	Baseline		+ IC^3	
	H ↑	C ↑	H ↑	C ↑
Ofa	2.876	3.891	**2.965**	**4.010**
Blip	2.901	**3.951**	**2.921**	3.881
References	2.932	3.966	**2.985**	**3.985**

MOS experiments (Table 6.4), IC^3 augmented OFA and Reference captions both significantly outperform their baselines ($p < 0.05, n = 41$) on both BLIP and OFA Hard MRR sets, but the experiments with BLIP on the BLIP Hard MRR set are inconclusive ($p = 0.682, n = 41$), suggesting that in some cases, IC^3 is unable to overcome all of the challenges with the seed captioning model. The fact that the head-to-head performance on the BLIP Hard MRR split in Table 6.3 is stronger for IC^3, coupled with the fact that reference captions augmented with IC^3 perform better on this set suggests that IC^3 can manage some of the underlying noise,

Table 6.3: Head-To-Head human evaluation performance of IC^3 on the hard MRR MS-COCO splits.

Model	Helpfulness ↑	Correctness ↑
Blip + IC^3	**48.06%**	**41.78%**
Blip	28.50%	30.43%
Tie	21.01%	25.60%
Ofa + IC^3	**51.10%**	**48.90%**
Ofa	32.04%	29.52%
Tie	15.06%	19.78%

Table 6.4: Human rater mean opinion score for IC^3 on Hard-MRR subsets. Helpfulness (H, 0-4), Correctness (C, 0-5).

Candidate Generator	Baseline		+ IC^3	
	H ↑	C ↑	H ↑	C ↑
OFA Hard Subset				
Ofa	2.452	3.651	**2.713**	**3.713**
References	2.649	3.675	**2.728**	**3.902**
BLIP Hard Subset				
BLIP	**2.708**	**3.827**	2.648	3.704
References	2.887	3.887	**2.934**	**3.918**

does not fully compensate for a lack of calibration.

6.3.2 Automated Evaluation

As discussed in subsection 6.2.4, we also perform automated evaluations of the method on both the MS-COCO and Flickr-30K datasets. The performance of IC^3 in CLIP recall is first demonstrated in Table 6.5, where for MS-COCO, CLIP recall MRR is improved by 27.6% under OFA, and by 46.5% under BLIP, suggesting that IC^3 augmented captions significantly outperform SOTA captions in indexing scenarios. Similar improvements exist in Table 6.6, where IC^3 improves CLIP MRR by 22.49% for OFA and up to 84.46% for BLIP. These results suggest that IC^3 surfaces significant additional detail compared to individual baseline and

Table 6.5: CLIP Recall for IC^3 augmented captions in the MS-COCO Dataset (Karpathy Test Split). MRR: Mean Reciprocal Recall, R@K: Recall @ K.

Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑
Ref + IC^3	**0.776**	**0.691**	**0.883**	**0.930**
Ref	0.593	0.480	0.724	0.808
Ofa + IC^3	**0.748**	**0.656**	0.857	0.914
Blip + IC^3	0.734	0.639	0.848	0.908
Blip2 + IC^3	0.746	0.652	**0.863**	**0.921**
Ofa	0.586	0.472	0.717	0.798
Blip	0.501	0.382	0.634	0.736
Blip2	0.589	0.473	0.725	0.811

Table 6.6: CLIP Recall for IC^3 captions in the Flickr-30K test set. MRR: Mean Reciprocal Recall, R@K: Recall @ K.

Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑
Ref + IC^3	**0.856**	**0.836**	**0.836**	**0.938**
Ref	0.708	0.679	0.679	0.798
Ofa + IC^3	**0.806**	**0.782**	**0.782**	**0.889**
Blip + IC^3	0.736	0.707	0.707	0.829
Ofa	0.658	0.629	0.629	0.745
Blip	0.499	0.463	0.463	0.581

reference sentences, leading to strong recall performance, and suggesting that IC^3 augmented captions can lead to benefits when applied to indexing and search. **On all datasets, IC^3 outperforms single human reference captions**, suggesting that summarizing multiple viewpoints is essential for strong automated recall performance.

Table 6.7 and Table 6.8 both demonstrate the summarization ability of IC^3 augmented methods, as IC^3 outperforms all baseline methods in recalling content from the dataset, often by relatively large margins. The verb recall is often lower (though still improved) across all approaches, suggesting that IC^3 focuses recalling content over action in an image. We further quantify IC^3's summarization capability in Appendix B.2.5 where we find that increasing the diversity of the input candidates can improve noun/verb recall, however has little impact on

Table 6.7: Content coverage performance on IC^3 augmented captions in the MS-COCO Dataset (Karpathy Test Split). N: Noun Recall, V: Verb Recall

	Exact		Fuzzy	
Model	N ↑	V ↑	N ↑	V ↑
Ref + IC^3	**0.552**	**0.354**	**0.767**	**0.616**
Ref	0.255	0.137	0.567	0.398
Blip2 + IC^3	**0.364**	**0.229**	**0.667**	0.529
Blip + IC^3	0.353	0.223	0.663	**0.534**
ofa + IC^3	0.351	0.211	0.656	0.498
Blip2	0.277	0.185	0.582	0.442
Blip	0.266	0.196	0.573	0.486
ofa	0.275	0.171	0.583	0.412

Table 6.8: Content coverage performance on IC^3 augmented captions in the Flickr-30K Test Dataset.

	Exact		Fuzzy	
Model	Noun ↑	Verb ↑	Noun ↑	Verb ↑
Ref + IC^3	**0.548**	**0.350**	**0.763**	**0.684**
Ref	0.246	0.147	0.543	0.490
Blip + IC^3	0.283	**0.200**	0.604	**0.585**
ofa + IC^3	**0.296**	0.195	**0.607**	0.571
Blip	0.205	0.134	0.505	0.507
ofa	0.230	0.147	0.533	0.495

MRR. These results suggest that IC^3 summarizes any salient information as required.

While Appendix B.3.3 discusses the performance of our methods on N-Gram measures, such measures are relatively misleading, as we generate captions that *differ significantly* from reference captions, thus, the N-Gram metrics are naturally lower compared to maximum-likelihood baselines.

6.4 Limitations

While IC^3 significantly outperforms baseline captioning approaches, as well can outperform single human image captioning references, it also suffers from several distinct limitations.

Hallucination: While IC^3 often produces high-quality summaries of the associated captions, it has several distinct failure modes, mostly coming down to hallucinations induced by the underlying captioning model. In some cases, objects that are hallucinated by the model can propagate to the summary, even if they are internally inconsistent with other captions in the candidate set K. Another distinct failure mode of the captions is when uncertainty in the samples is interpreted as two distinct parts of the image. For example, if 50% of the captions refer to a dog, and 50% of the captions refer to a cat, the model may infer that both a dog and a cat are present in the image, even though only a single, unknown, animal is there. Examples of these failure cases are shown in Appendix B.8. We believe that such failure cases can largely be solved by introducing a visually aware summarization model, however, as of writing, no sufficiently large-scale general-purpose multi-modal model exists which is capable of serving this purpose.

Controllability: One of the key applications of image captioning systems is alt-text generation. As discussed in recent work (Archibald, 2023), alt-text generation is largely contextual, which means that for each image, the alt-text of the image should depend on the context that such an image is included in. While IC^3 introduces a natural pathway for including context through the summarization model, we have found (see Appendix B.8.2), that IC^3 is somewhat resistant to prompts that encourage surfacing background information. Exploring how to make IC^3 surface arbitrary information in the image instead of focusing primarily on foreground information is key direction for future work.

The Cost of using LLMs: The use of many closed source large language models can represent a significant financial, human, and environmental cost (Bender et al., 2021). We recognize that for some researchers and students, the financial cost of using a large zero-shot model such as GPT-3 can be prohibitive, making IC^3 difficult to compare against, especially for large-scale experiments such as the Karpathy test set for MS-COCO and the Flicker-30K datasets (which consists of 5K images each). Using GPT-3, IC3 costs about \$0.0109/Image, and with GPT-3.5, that cost falls to \$0.001/Image. Notably, this is significantly less than Chat Captioner (Zhu et al., 2023a), which can cost as much as \$0.27/Image, which made it infeasible to run large-scale experiments. The experiments/ablations/all GPT-3 tuning in this paper was performed for \$250 (USD). Our approach, while not necessarily cheap, is several orders of magnitude less expensive than training/evaluating fine-tuned vision and language models such as Flamingo (1536 TPUs/15 days, roughly \$1,780,531 using on-demand TPU pricing) or BLIP-2 (16 A100 GPUs, 6 days, \$11,796 using AWS on-demand pricing). Furthermore, we hope that this cost will not be prohibitive long-term. GPT-3.5 is an order

of magnitude cheaper, and has similar performance to GPT-3, and open-weight models such as Koala and Vicuna, seem promising for the future of affordable LLMs (see Appendix B.2.1), making IC3 even more accessible to students and researchers.

6.5 Conclusion

In this work, we introduced IC^3, a method for image captioning that first samples captions from multiple viewpoints and then summarizes and filters them to create high-fidelity descriptions. As far as we are aware, IC^3 is the first work to demonstrate a pipeline for generating a single caption by integrating distributionally-faithful candidate captions, and does so without changing model architecture or retraining by leveraging summarization to produce a single omnibus caption capturing the full distribution of information. Further, IC^3 is the first work for paragraph captioning or image captioning that uses summarization of distributionally-faithful caption samples and the first to demonstrate in human experiments that long-form captions encoding this distribution are preferable to single reference captions. Human users rate IC^3 captions at least as helpful as baseline captions up to 80% of the time, and such IC^3 captions are capable of inducing up to 84% relative improvements in recall approaches over baseline captioning methods. While our implementation of IC^3 is relatively simple, it demonstrates significant gains over traditional paradigms, suggesting that this is only the beginning for caption sampling and summary methods.

Chapter 7

Multi-modal pre-training for automated speech recognition

As discussed in the introduction to this section, despite considerable research, automated speech recognition (ASR) remains an extremely challenging task, especially in noisy environments. Correctly understanding spoken phonemes requires an understanding of speech patterns, as well as an understanding of myriad varieties of background noise, much of which may never have been encountered by a model during the training process. Many traditional ASR methods focus on a local understanding of phonemes, predicted from small 10-30ms segments on audio. Unfortunately, such local representations may leave ASR models vulnerable to extreme noise such as frame drops or sudden loud noises at the local level. Not only are such models vulnerable to local disruptions, but these models can be affected by global-level noise that has not been seen during the training process.

In this chapter, we target the problem of such global-level noise in utterances. Many ASR datasets such as Librispeech (Panayotov et al., 2015) are collected in lab-specific environments, and even a number of large corporate datasets are collected from a canonical set of situations which leaves a long tailed distribution of noisy environments uncovered. While local level disruptions can in part be solved by introducing semantic-level language modeling (Gulati et al., 2020), the global out-of-distribution noise problem has no such simple solution. Recently, the vision and NLP communities have introduced several methods based on self-supervised representation learning, which make use of large amounts of unlabeled data to build representations which can augment low-data downstream tasks. Such representations can provide exposure to the long-tailed data distribution, and have been shown to reduce the amount of data required to learn robust representations for downstream vision and language tasks (Sun et al., 2019).

We hypothesize that by leveraging self-supervised learning to learn representations of the global environment, we can improve the performance of ASR models. If we allow the model to additionally condition the phoneme output on a robust representation of the environment, models should be able to respond correctly in a wider variety of global noise environments (as the model has some environmental experience, even though it may be outside of the ASR

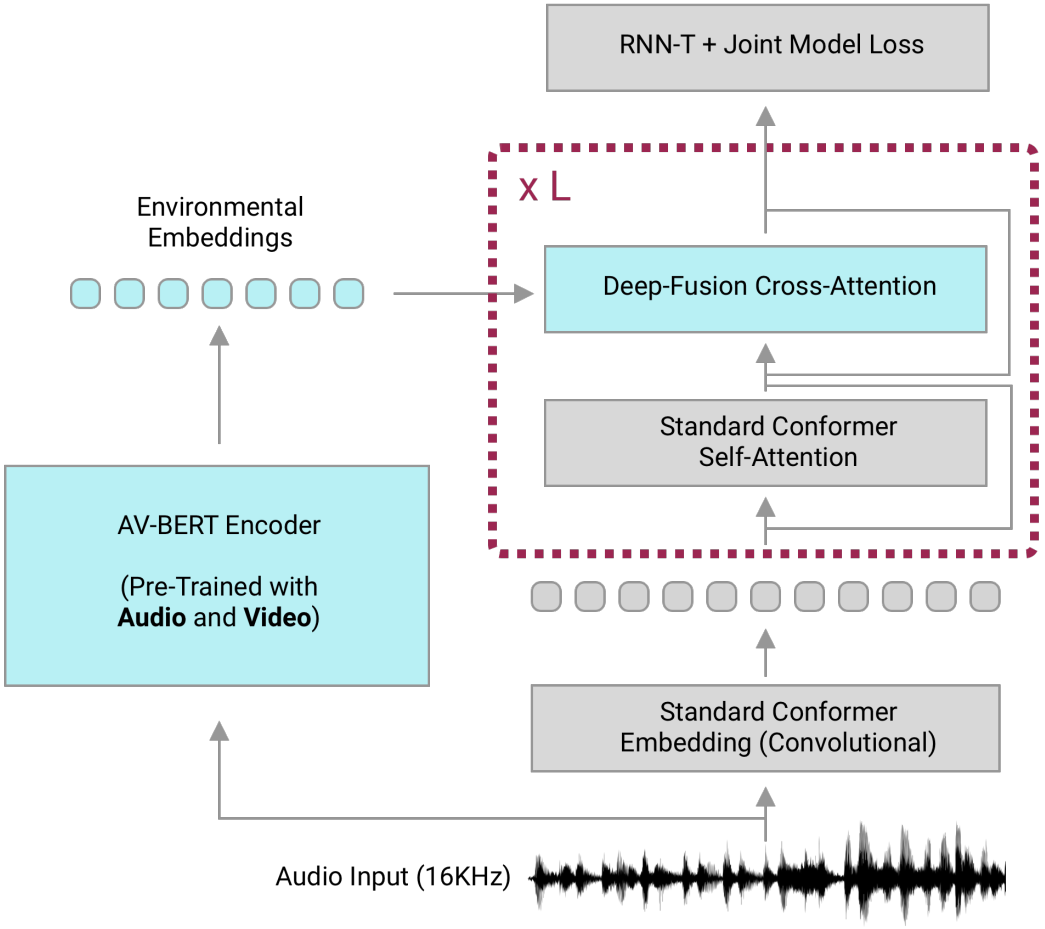


Figure 7.1: An overview of our proposed approach to the ASR training process using deep-fusion with environmental embeddings. Our audio is fed to the pre-trained environmental representation model, trained on large-scale multimodal data. We then use a stack of L deep-fusion cross-attention layers in the base conformer architecture to deeply fuse the environmental representations with a standard conformer model. The RNN-T and joint model loss remain unchanged from (Gulati et al., 2020).

domain). In this work we provide evidence to support this hypothesis, by demonstrating the following:

1. We develop a pre-training scheme, AV-BERT, based on masked language modeling, for learning robust environmental representations (subsection 7.2.1).

2. We introduce a novel deep-fusion scheme for training on joint local/global representations in the ASR domain (subsection 7.2.2).

3. We demonstrate the benefits of generating robust environmental representations with additional modalities such as visual information (even when such visual information is not present at test time) (section 7.3).

4. We discuss directions for further improvement of ASR techniques leveraging global multimodal representations (section 7.4).

7.1 Related Work

Traditional ASR methods such as the RNN-T (Graves et al., 2013) have two main components, an audio encoder and a joint transducer model. In this framework, the audio encoder is largely responsible for generating the local probabilities of any particular phoneme, while the joint transducer model introduces language/semantic level information. Recent evolution in the audio encoder space focus primarily on improving the performance of the local level representations. The Transformer-T (Zhang et al., 2020c) and Conformer (Gulati et al., 2020) are both methods which introduce self-attention into the computation of the local embeddings. Because of the transducer-level loss, methods are still primarily local-first, with each token representation focusing on predicting the phoneme present in the given 10-30ms audio frame (Hsu et al., 2021). The local-level dependence can introduce a vulnerability to global out of domain data, as demonstrated by Chiu et al. (Chiu et al., 2021).

Instead of focusing on building global audio representations directly, many methods leverage self-supervised learning to build additional exposure to the long-tail of the ASR distribution. Self-supervised learning has been used to great effect in the ASR community, primarily in the context of student teacher models such as those in Watanabe et al., Zhang et al., Manohar et al. and Movsner et al. (Watanabe et al., 2017; Zhang et al., 2020f; Manohar et al., 2018; Mosner et al., 2019). These methods have the remarkable property of both distilling the representations learned by a large/slow teacher model trained on seed data to a smaller representation, as well as improving the overall performance of the model. We speculate (and believe that it is important future work to confirm) that the additional performance gained by student teachers is a direct response to exposure to a large tail of environmental effects, regardless of the ASR content.

While student-teacher models make up the majority of self-supervised learning in ASR, recent techniques such as HuBERT (Hsu et al., 2021) and COLA (Saeed et al., 2021) have

demonstrated that learning general representations of audio can be useful for ASR pre-training. Our proposed method could be considered a natural extension of the ideas in HuBERT, as we extend the ideas of building a speech-only general representation to building a multi-modal environment-level representation of the audio. Since our proposed method expands the scope of the environmental representation with not only environmental audio, but also visual data, we believe that our models capture even more robust distributional information than is possible using speech-data alone.

Outside of automated speech recognition, building an environment-level multi-modal representations have consistently been shown to be effective methods for instilling global domain knowledge. Akbari et al. (Akbari et al., 2021) recently demonstrated that building representations in a contrastive learning framework by ingesting video, audio and text information can lead to state of the art performance on the Kinetics (Carreira et al., 2018) classification dataset, and Alayrac et al. (Alayrac et al., 2020) demonstrated that self-supervision over the video, audio and language streams simultaneously can often lead to representations that perform better on global classification tasks in both the vision and audio domains. Perhaps most importantly, Wang et al. (Wang et al., 2021a) recently confirmed that even in the absence of the video representation at test time, audio-video self-supervised model representations outperform audio-only self-supervised models on *audio only downstream tasks*.

The finding by Wang et al. (Wang et al., 2021a) is somewhat unintuitive - as it suggests that video information during pre-training can help to organize the audio representations in such a way that the models perform better at test time on audio-only problems. In designing our proposed method, we exploit the same effect, which we hypothesize (and will have to confirm), is due to the videos acting as pseudo-semantic labels over the unsupervised audio data. Unlike the model proposed by Wang et al.(Wang et al., 2021a), our proposed method makes direct use of this hypothesis by using masked language modeling as a joint training objective, as opposed to contrastive representation learning. In contrastive representation learning, samples from the audio and video domains are pushed into a joint latent space, traditionally at a global level. This mode of training inherently suggests that the models should lie in the same latent space, which we believe to be suboptimal for automated speech recognition, where we want to focus on globally aware local-first representations. To help avoid this, our proposed method focuses on masked language modeling objectives, in a framework first explored in BERT (Devlin et al., 2019), and further extended to multiple modalities in methods such as VideoBERT (Sun et al., 2019), UniT (Hu and Singh, 2021), and Multi (Tsai et al., 2019).

In the context of automated speech recognition, the exploration of global multimodal representations has been largely unexplored. Our proposed method seeks to close a gap in the research: exploring if multimodal representations can be used to improve automated speech recognition tools even when *the video information is not present at test time*. We frame our multimodal representations in the context of larger goal of building environmental representations, as demonstrated in section 7.3, which shows that even building audio-only global representations from out of domain data can lead to improved ASR performance, and that augmenting with multiple modalities only improves this performance benchmark.

7.2 AV-BERT: Multimodal Pre-Training for ASR

Our method consists of a two-stage approach (an overview is given in Figure 7.1) inspired by ideas from VideoBERT (Sun et al., 2019), HuBERT (Hsu et al., 2021) and Alayrac et al. (Alayrac et al., 2020). In the first stage we build a video-augmented audio representation using a pre-training task based on masked language modeling. In the second stage, we use these video-augmented audio representations to provide additional context to a conformer-based ASR model.

7.2.1 Multimodal Pre-Training

While many methods for learning multimodal representations focus on self-supervised learning with a contrastive objective, our proposed method, AV-BERT, differs in that it uses a masked language modeling objective. Our pre-training encoder model, shown in Figure 7.2, takes inspiration from the UniT (Hu and Singh, 2021) model for unified transformer architectures, however instead of using multiple encoders, we use a single unified encoder layer. We take further inspiration from ViViT (Arnab et al., 2021) and use a video-patch based encoding to the transformer, while taking inspiration from HuBERT's (Manohar et al., 2018) iterated quantization training method for masked-language modeling from raw signals. In this section, we dive deeper into each of the components, and discuss our modeling choices from the perspective of an ASR-first multimodal pre-training model.

To build a multimodal representation learning method based on masked language modeling principles, we first consider a token-based representation of modalities. We draw the representation from a discrete quantization of both the video and audio domains.

For the video modality, we extract non-overlapping patches of the input data, which are further quantized. Audio data is batched and quantized in a similar way. While we could use the quantized tokens directly as input to the masked language model as was done in VideoBERT (Sun et al., 2019), similar to HuBERT (Hsu et al., 2021) and ViViT (Arnab et al., 2021) we use the pixel/waveform patches directly on the input layer (while still classifying based on the quantization). This allows the model to see the full input audio/video, while gaining the benefits of a masked language model. In addition, this allows the model to respond to subtle changes in the input which cannot be captured by our audiovisual language (which only consists of a total of 12288 tokens). It is possible to use a larger audiovisual language, however doing so leads to an increase in computational complexity that can be detrimental to training speed (and thus, training performance).

Thus, to form the input to our masked language model, we first use a set of modality-specific convolutions to embed both the video and audio in the model dimension. We then apply a modality-specific learned embedding, as well as learned position embedding (Devlin et al., 2019). For the audio, we use the frame index as the position. For the video, we apply a pair of position embeddings – one across the flattened spatial dimension, and one across the temporal dimension (the same as in the Timesformer (Bertasius et al., 2021)). We then

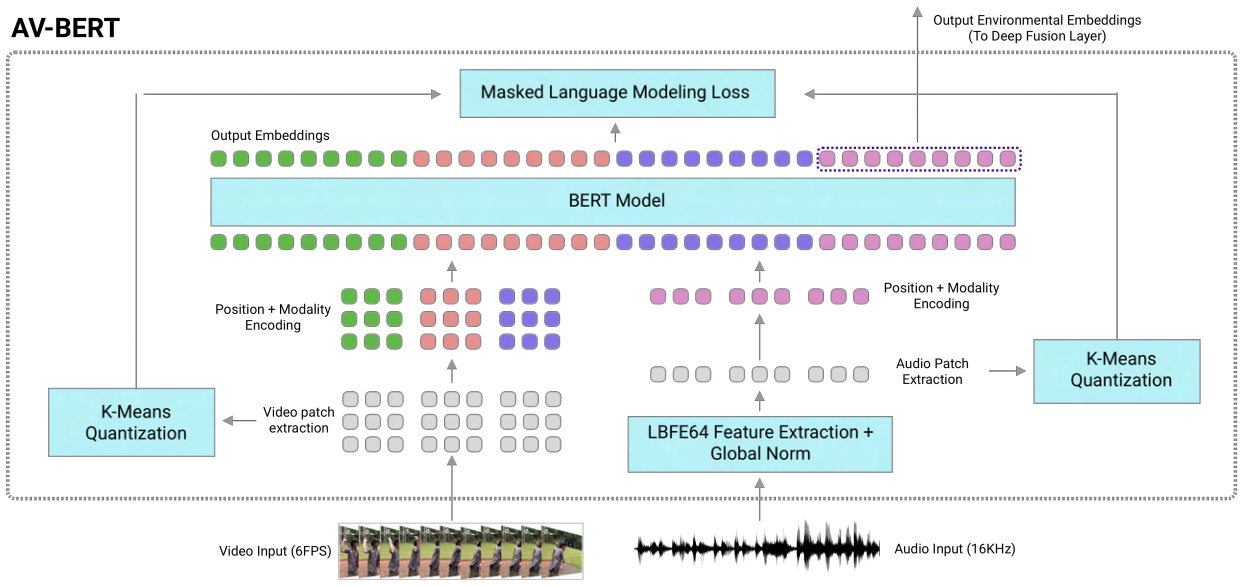


Figure 7.2: An overview of our pre-training model. First, a set of patches are extracted from the multimodal inputs. Next, these patches are quantized using k-means and embedded directly using convolutional layers, modality encodings, and positional encodings. The embedded patches form the input sequence, which is passed to a standard BERT masked-language model. The quantized token labels are used along with the output of the masked BERT model to perform masked-language prediction.

flatten the spatial dimensions, and concatenate the video and audio sequences, to form the input to the model.

To perform masked language modeling, we use an architecture similar to BERT (Devlin et al., 2019), which allows for full cross-attention between all points (both in the audio and video modalities, as well as spatiotemporally). This can lead to very long sequence lengths, so we compensate by reducing the per-node batch size, and distributing the model across several GPUs. Because of the distribution across many GPUs, we do not use batch normalization, and instead use instance normalization to help handle the inputs. This replacement could potentially lead to a degradation in accuracy over the original model presented by (Devlin et al., 2019) (which we will verify empirically), however it significantly increases training speed.

The training of the AV-BERT model is heavily dependent on the choice of masking technique. If we mask tokens uniformly with some rate, it is unlikely that the model will learn cross-model representations, as both audio and video are highly local representations (information in one location tends to share a large amount of mutual information with other neighbors). A natural solution to this approach would be to mask entire modalities at a time. This approach. however, can be too broad, as often the modalities are not heavily correlated

enough to reconstruct a quantized representation of the other.

To combat this breakdown in representation, we apply a progressive masking technique, where we begin the training by masking local information to encourage local-level representations, and progressively increase the size of the masks during training. This encourages the model to first learn local representations, and then eventually learn more global representations as the training process continues.

We perform our pre-training using the publicly available splits of the Kinetics-600 dataset (Carreira et al., 2018). The Kinetics-600 dataset consists of 366K 10 second videos, each with a corresponding audio track and an associated action label (from 600 classes). For video, we reduce the frame-rate to 6FPS and resize the short side of the video to 256 pixels, and take a 256x256 center crop of the resulting resized video. For the audio, we resample the raw input audio to 16KHz, and stack 3 adjacent LBFE64 features from the resulting audio frames. The features are whitened using a global-norm, then clipped to lie between $(-1.2, 1.2)$ for numerical stability.

In our pre-training experiments, we use a model dimension of 128, with six BERT encoder blocks. The model is implemented in Tensorflow (Abadi et al., 2016), and is trained using 32 Nvidia-V100 GPUs, each with a batch size of 8, for 100 epochs (or until the validation perplexity converges, whichever comes first). To perform the optimization, we use an Adam (Kingma and Ba, 2015) optimizer with a fixed learning rate of $3e^{-4}$, and $\beta_1 = 0.9, \beta_2 = 0.99$.

7.2.2 Automated Speech Recognition Downstream Task

For the downstream automated speech recognition task, we have two goals: (a) maintain the performance of current state of the art machine learning techniques, and (b) augment the current models with additional global-level environment context to improve phoneme recognition. In order to accomplish these goals, we modify the conformer architecture (as shown in Figure 7.1) to include additional cross-attention layers, which attend across the vector-level representations generated by our pre-trained AV-BERT model. This method allows the model to selectively pay attention to global context information learned by our model, while preserving the local-first approach favored by ASR techniques. This helps resolve one of the major challenges faced by HuBERT (Hsu et al., 2021) in the ASR domain: when you focus on learning global representations, you can fail to encode the information necessary for local-first tasks such as phoneme detection. During the training of the downstream model, we freeze the representations learned by AV-BERT, to both reduce the computational complexity and to maintain a more global representation level even after significant training.

We evaluate the proposed model on the LibriSpeech (Panayotov et al., 2015) dataset, which consists of 970 hours of labeled speech. Because our audio embedding method is frozen during the training process, to ensure that there is no domain shift, we follow the same audio preprocessing technique as in subsection 7.2.1 with additional SpecAugment (Park et al., 2019). In addition to Librispeech, we present results on several internal datasets: “Base" representing general speech, “Query", representing standard speech queries, “Rare" representing the long-tailed distribution of rare words, and “Messages" representing longer

Table 7.1: Results summary of word error rate for the Librispeech dataset with no additional language model. The baseline model replaces the cross-attentions with self-attention (to closely preserve parameters) using the same training profile (See section 7.2). “A" is the audio-only model, and “A/V" is the full Audio/Video BERT.

Method	Params (M)	test-clean	test-other
LAS			
Transformer (Synnaeve et al., 2019)	370	2.89	6.98
Transformer (Karita et al., 2019)	-	2.2	5.6
LSTM (Gulati et al., 2020)	360	2.6	6.0
Transducer			
Transformer (Zhang et al., 2020c)	139	2.4	5.6
ContextNet (M) (Han et al., 2020)	31.4	2.4	5.4
ContextNet (L) (Han et al., 2020)	112.7	2.1	4.6
Conformer			
Conformer (M) (Gulati et al., 2020)	30.7	2.3	5.0
Conformer (L) (Gulati et al., 2020)	118.8	2.1	4.3
Ours			
Conf. (M, base)	79	2.21	4.85
Conf. (L, base)	122	2.11	4.29
A + Conf. (M)	79	2.15 **(+2.7%)**	4.82 **(+0.6%)**
A/V + Conf. (M)	79	2.10 **(+4.8%)**	4.72 **(+2.7%)**
A/V + Conf. (L)	122	1.98 **(+7.0%)**	4.10 **(+4.4%)**

message-based utterances. All customer-specific data has been de-identified from these internal datasets. For these results, we use the same model architecture, however train on a corpus consisting of 120K hours of labeled speech, and 180K hours of unsupervised speech using in-house teacher distillation.

For our ASR model we use a model dimension of 512 with a feed-forward dimension of 1024. Our model consists of 24 self-attention/cross-attention blocks, and has a convolutional downsampling on the input with a kernel size of 3 and a stride of 2. We use a standard joint model similar to RNN-T (Graves et al., 2013) based joint model with a graph-based ASR decoding framework. The optimization process is shared with AV-BERT, and described in subsection 7.2.1, with the exception that we use a per-node batch size of 28.

7.3 Results & Discussion

Our main results are presented in Table 7.1. We report results on two models, a model using AV-BERT trained with both the video and audio components of the Kinetics dataset, as well as a model trained with only the audio from Kinetics. The Baseline model is identical to the proposed model except all multi-modal cross-attention layers are replaced

Reference:	should i buy from the princess starfrost set royale high
Base (M):	should i buy from the princess stare froset in we're all rawhide
Ours (M):	should i buy from the princess star frost set royale high
Reference:	read all of lisa left eye lopes songs including the thirteen more
Base (M):	read all of lisa **** *** loeb songs including the thirteen horn
Ours (M):	read all of lisa left eye lopez songs including the thirteen more
Reference:	... signalman he lead tenor for telephone wires so soldiers ...
Base (M):	... signal map he'd late tenoff telephone wise ** soldiers ...
Ours (M):	... signalman he lead teno for telephone wires so soldiers...

Figure 7.3: Qualitative examples showing improvements on utterances in our model, vs the baseline model. Blue indicates a deletion, pink indicates a substitution and yellow indicates an insertion. The first example demonstrates an additional robustness to unfamiliar terms, which our proposed model has additional exposure to through out of domain pre-training. In the second example, a local noise event causes the baseline model to suffer, however our model is able to compensate with it's global-first representations. In the third example, global noise is present in the sample, however our model with audio/video pre-training can compensate for this global noise distribution due to its exposure to out of domain data.

with self-attention layers to preserve the parameter count, and the representational capacity.

The results show that both models — one trained on audio only and another with audio + video embeddings — outperform the baseline model. As has been previously shown, larger models are able to outperform smaller models, with the large model achieving larger gains over the smaller medium model. Training with audio-generated embeddings alone induces a performance jump over the baseline. By allowing the model to generate context-level embeddings on out of domain data, the audio-only models are able to gain pseudo-exposure to a longer tail of possible noise scenarios. Surprisingly, the model achieves better relative performance gains on the test-clean dataset, which should have much cleaner audio than the test-other dataset. We hypothesize that this is caused by the relatively high variance in the test-clean dataset. Improving the performance on a small number of samples in test-clean can lead to higher percentage gains. Because the test-other dataset is larger, the effect is less pronounced.

We further validate the method with experiments on internal Alexa AI datasets in Table 7.2. We can see that while the method produces very large performance gains over baselines with the smaller models, the effect in larger models is less pronounced. This demonstrates the power of context-level embeddings to help with models that have less representation power. By providing access to a self-supervised embedding, the model is able to compensate for a large amount of the lost representational power of the network. In the larger models, using contextual embeddings is less powerful, since the models both have access to a larger

Table 7.2: Relative improvement over baseline WER for Alexa-AI datasets methods without a language model. We can see that adding contextual embeddings can significantly improve performance.

Method	Base	Rare	Query	Messages
Conformer (M)	0	0	0	0
+ Audio (M)	+30.1%	+17.9%	+26.7%	+20.1%
+ Audio/Video (M)	+45.6%	+31.2%	+38.7%	+17.2%
Conformer (L)	0	0	0	0
+ Audio/Video (L)	+5.1%	+5.4%	+4.2%	+5.9%

representation space and more data to train on. We still, however, see the benefit of adding out of domain data, as we can see that the test set sees much better performance. In our experiments, we found that the training accuracies remained unchanged in all of these models, which suggests somewhat that the models are limited by the input data and model capacity, rather than model architecture. Our audio/visual embeddings compensate for both of these issues by training on large-scale out of domain data, and leveraging additional capacity from the pre-training models.

In Table 7.2, we can see that the performance of the large model is most dramatic in both the messages and rare words datasets. These datasets are dominated by complex terms, and longer utterances which are unlikely to appear in the in-domain training data, the same place where our model is expected to demonstrate the best performance. The gains in the smaller model are less pronounced in both of these datasets. We believe that in the smaller datasets, the models do not have sufficient internal complexity to leverage the embeddings to their fullest extent, and while the embeddings do improve the performance, they are not able to capture the subtle connections between the global context and the local speech phonemes that models with more parameters are able to capture.

Table 7.3 demonstrates the relative performance of our large A/V model vs the baseline model when evaluated on the Alexa AI base dataset. We can see here that the model primarily reduces the word error rate by fixing insertions in deletions at the cost of substitutions. This is intuitive: the model is able to account for additional noise in the dataset more effectively, and thus can reduce deletions and insertions caused by noise, however the model is less effective at handling issues caused by incorrect phoneme recognition, leading to substitutions. In addition, we can see that the model performs better on areas where there are very few samples, validating the hypothesis that the model improves performance on the long-tail of the distribution of audio categories.

Category	NUM	ΔWER (%)	ΔSUB (%)	ΔINS (%)	ΔDEL (%)
System	4972	0	-12.12	-1.83	22.86
Global	3887	-0.25	-0.51	-20	10.14
Music	2809	6.58	5.65	9.26	7.32
Home Automation	2453	4.96	4.69	20.29	-12.07
Notifications	1244	1.33	5.22	-4.92	0
Knowledge	779	-0.93	-5.41	-8.33	11.07
Weather	408	0	8.51	8.51	-225
Calling And Messaging	182	-5.54	-6.35	0	-7.3
Video	104	20.48	20.74	39.79	0
Household Lists	86	11.25	0	32.08	100
App	86	-28.62	-41.4	0	11.11
Shopping	46	-10.53	22.25	0	-44.5
Original Content	41	22.3	28.4	0	0
Daily Briefing	34	0	-49.62	0	50.38
Books	28	16.84	0	0	100
Sports	25	33.4	37.35	0	0
Recipes	23	0	0	0	0
System Settings	19	16.67	25	0	0
User Profile	14	100	100	100	0
None	14	8.59	25.1	0	50.21
Translation	12	0	0	0	0
Gallery	10	-24.95	-33.58	0	0
Local Search	9	80.05	66.73	100	0
Game	8	100	100	100	0

Table 7.3: Table demonstrating the performance split across different domains in the Alexa AI base dataset. Negative numbers represent worse performance by the A/V Conformer L model over the baseline, while positive numbers represent performance gains. We notice that in many cases, the largest gains come in the long tail of the categories, where fewer utterances are available. In the more global categories, the performance remains similar, or can improve if the model is more likely to encounter rare words, such as in the notifications class, or the model encounters noise (such as is often the case in the music class).

7.4 Conclusion

While this model represents an initial step towards using multi-modal embeddings in automated speech recognition, there is still a long way to go. In many ASR applications, multi-modal inputs are available, which means that we have the capacity and ability to leverage both modalities at test time. This would strengthen the connection between the audio and visual data, and allow for more information to be leveraged during test-time for disambiguation, further improving performance. While in some cases the visual data may be present, it is not always the case that visual and auditory data are correlated. While in this paper we rely on the law of large data to compensate for uncorrelated inputs, it is perfectly reasonable to expect that even in the infinite data scenario, the portion of uncorrelated audio

and video present in collected videos from YouTube is non-zero. Dealing with uncorrelated semantic inputs remains important, and interesting, future work.

In addition to exploring uncorrelated inputs, it is also interesting to improve the contextual embedding model. In this work, we use a simple vision + audio model inspired by videoBERT and HuBERT, however recently models such as VATT (Akbari et al., 2021) expand the context to include not only audio and video, but language as well. Further augmenting the context is likely to exaggerate the performance gains, as the model is able to even more efficiently represent the context of a scenario, and compensate for even longer tail elements of the ASR distribution.

In conclusion, with this paper, we have introduced an initial approach for exploring global-level contextual embeddings for the automated speech recognition pipeline. We build a novel self-supervised vision + audio encoder, and demonstrate the performance of this method by using deep-fusion to directly connect the contextual embeddings with the local-first conformer model. Our model demonstrates strong performance on Librispeech, and presents a new direction for exploration into multi-modal ASR.

Chapter 8

Domain Adaptation with External Off-Policy Acoustic Catalogs for Scalable Contextual End-to-End Automated Speech Recognition

One of the most challenging problems in automated speech recognition (ASR) is specializing large-scale models, particularly speech encoders, for downstream applications that often (a) have fewer labeled training examples, and (b) rapidly evolving distributions of speech data. The traditional approach to this problem is to frequently collect fresh data, which can be used to re-train and specialize models, leveraging tools such as domain-prompts (Dingliwa et al., 2022), incremental-learning (Baby et al., 2022), knowledge distillation (Zhao et al., 2022), or hand-written grammars (Gandhe et al., 2018) to reduce the impact of re-training the model for the downstream application. Unfortunately, for data that changes on a rapid basis, such as product listings or applications requiring per-customer specialization, such methods, while effective, are either inherently slow or remain computationally infeasible.

In this chapter, we propose a method that leverages context from external text data catalogs – large lists that can contain as much as 10 million specialized words or phrases – to improve the performance of models during both the fine-tuning process, and when specializing an already fine-tuned model to a new dataset. Here are the key highlights of our approach: first, we generate a key-value external knowledge store that maps an audio representation of each text element of the catalog (usually consisting of 1M-10M examples) to a semantic representation of the text. Next, we train a model that leverages this external store by attending over retrieved key/value pairs, which we retrieve through approximate k-nearest neighbors. The external, constant, and off-policy key-value store can be updated during specialization, requiring only an updated list of phrases for each new model instead of additional fine-tuning.

Leveraging external text data to improve the performance of audio encoders in ASR models has been studied for a long time. Perhaps the closest work to our proposed model

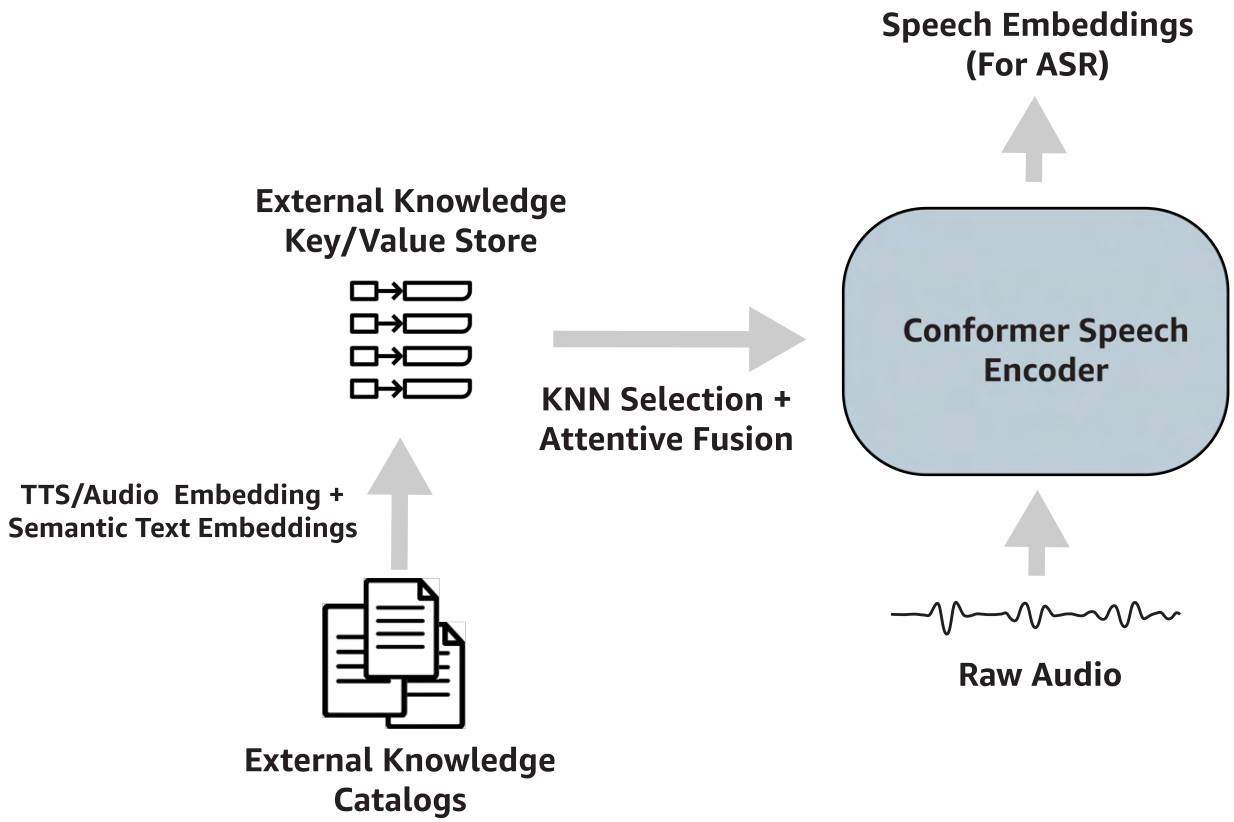


Figure 8.1: An overview of our method leveraging text-to-speech mappings for contextual ASR. Using data from a text catalog, we generate audio and text representations to generate mappings from audio key to text value. To leverage these mappings for ASR, we implement a K-Nearest Neighbors attention in the speech encoder during the fine-tuning (or training) phase.

was presented by Chen et al. (2019), who used attention over a *local* set of LSTM-based grapheme/phoneme embeddings to augment the audio encoder. They found that biasing the encoder with only 40 contextual text entities per utterance leads to improvements of up to 75% on specialized test datasets. Similarly, Sathyendra et al. (2022) and Chang et al. (2021) demonstrate WER reductions when small (<100) contexts are fused with an RNN-T in a multi-head attention-based process. Our method differs in that it is designed primarily for *domain specialization*, whereas existing biasing methods are focused on *personalization*. This is shown foremost in the scale of the catalogs – while in prior work, each utterance may have at most 100 utterances in their context, we leverage catalogs with up to 10M samples. Thus, our approach is designed to compensate for general domain shift, rather than supplementing ASR performance through limited personalization. Further, while current biasing approaches focus on late-stage fusion, we use deep fusion in the model network, which we demonstrate

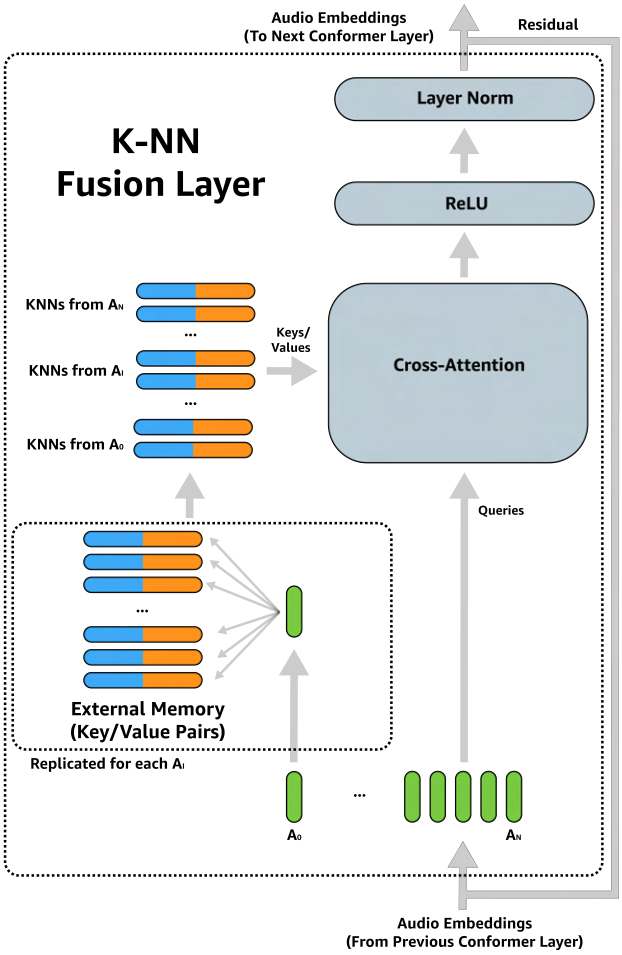


Figure 8.2: Overview of K-NN fusion layer. For each audio frame embedding, we extract approx KNNs using audio keys from our catalog, which form a context key/value store for a standard cross-attention layer (Vaswani et al., 2017), where queries are incoming audio frame embeddings.

(Table 8.2) is more effective. While biasing the *speech encoder* has been under-explored, many works (Novotney et al., 2022; Shenoy et al., 2021; Zhao et al., 2019; Liu and Lane, 2017; Jaech and Ostendorf, 2018; Kim and Metze, 2018; Lin et al., 2015; Williams et al., 2018; Munkhdalai et al., 2022) have shown the importance of biasing the language model in the ASR stack.

Outside of ASR, it has been shown that models augmented with external memory generated from large-scale text data have the potential to outperform similarly sized models without external knowledge. Borgeaud et al. (2022) recently demonstrated that leveraging external-knowledge lookup for natural language models can lead to efficiency improvements of up to 25x, and Wu et al. (2022) showed that expanding the context of standard text transformers through external cached key-value pairs can lead to significant perplexity improvements on the standard language modeling task. General memory augmentations have also been shown to be useful in translation (Khandelwal et al., 2021), RL (Goyal et al., 2022a), and image generation (Tang et al., 2022). This domain adaptation approach can combined with multi-modal learning (Chan et al., 2022a) and context-content factorization (Chan and Ghosh, 2022a) for better audio representation learning; it can also be used for metric learning (Mahadevan et al., 2018), continuous learning (Zhang et al., 2020a), or incremental learning (Li et al., 2019a; Zhang et al., 2020b).

Inspired by Borgeaud et al. (2022) and Wu et al. (2022), we apply a context embedding approach with a focus on ASR, leveraging TTS-generated audio data and semantic text embeddings to bias the speech encoder of a conformer model. To the best of our knowledge, using TTS to encode textual context has not been explored in prior work. Our key contributions are three-fold. (1) We outline the first method (to our knowledge) to leverage large-scale text data for contextual biasing of the speech encoder. (2) We show that our approach combined with an approximate K-NN lookup yields improved WER on ASR models, particularly when encoded catalogs match the target domain. (3) We show that our approach provides accurate solutions under the constraint of quick reactions to distribution changes (e.g., fast catalog updates for sporting events, changes in personal catalogs), without model retraining.

8.1 Learning from External Catalog Contexts

An overview of our method is given in Figure 8.1. Our approach consists of two key components: (1) A method for generating key-value mappings between the audible speech and a text representation of the catalog, which we call an “external memory" and (2) An attention-based module for fusing the “external memory" with the existing speech encoder. The external memory must be capable of offline and off-policy updates to enable memory alteration without incurring re-training costs.

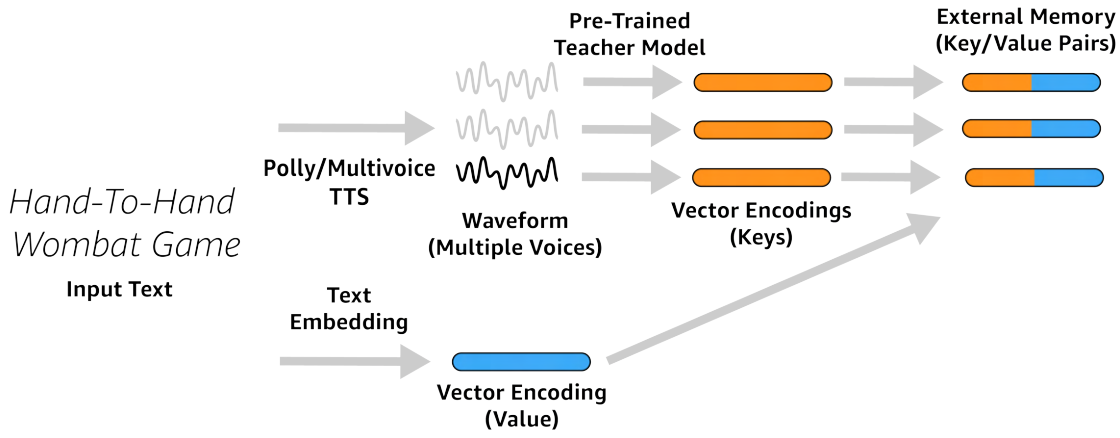


Figure 8.3: Overview of our text-catalog encoding process. For each catalog entry, we generate TTS-based audio encoding that forms the "key" vector in the key-value pair. The value is a semantic text-embedding of the entry. Key/value pairs are assembled into the external memory, referenced in Figure 8.2.

8.1.1 Generating the External Memory

An overview of the external-memory generation process is shown in Figure 8.3. Our approach generates the external memory consisting of audio-embedding key/text-embedding value pairs from a text-only catalog. To generate the audio-embedding key, we use text-to-speech (TTS) to generate waveform representations of the audio data, and then embed these waveform representations using the pre-trained speech encoder model. To generate the text-embedding values, we leverage off-the-shelf semantic text embedding methods, including 1-hot, GLoVE (Pennington et al., 2014) and BERT-style embedding (Devlin et al., 2019) approaches.

TTS: We explore two TTS modules to generate audio for the audio-embeddings: the Amazon Polly TTS service[1], and an Alexa-AI Internal text to speech (TTS) library optimized for synthetic ASR data, Multivoice-TTS (Vallés-Pérez et al., 2021). For both TTS methods we use ten voices drawn from en-US and en-GB locales. Silence (0.1s) is inserted before and after each utterance.

Audio Embedding: While audio embeddings for the external catalog could be constructed in several ways, similar to Wu et al. (2022), we aim to make audio-embeddings as close to on-policy self-attention embeddings as possible. Thus, we use the mean of the self-attention representations of the baseline model (no fine-tuning) at an intermediate layer, as audio embeddings. In this work, we always assume the existence of a suitable seed model from which the catalog audio embeddings, a_i, can be generated, and they are generated offline prior to

[1]https://aws.amazon.com/polly/

training the model described in subsection 8.1.2. While training the model and catalog jointly would eliminate off-policy speech embeddings, such a method presents technical challenges.

Text Embedding: We explore several methods of generating the text embeddings (the "value" in the catalog K-V pairs). Initially, for small catalogs, we explored learned one-hot embeddings, however, while one-hot embeddings can lead to better performance (Table 8.1), they are not scalable – as they cannot be computed offline (and thus, cannot be inserted during test time). To generate scalable text embeddings, we explore two semantic text-embedding approaches: GLoVE embeddings (Pennington et al., 2014), which are built using word co-occurrence probabilities, and BERT-style embeddings (Devlin et al., 2019), which are learned from large statistical models. GLoVE embeddings are 300 dimensional, and computed using the publicly available vectors, and our BERT-style embeddings are computed using the `all-MiniLM-L6-v2` model in the `sentence-transformers` package (Thakur et al., 2021).

8.1.2 External Memory Fusion

An overview of the external memory fusion process is given in Figure 8.2. The speech encoder in our proposed work is based on the Conformer encoder (Gulati et al., 2020), augmented with additional K-Nearest-Neighbor (KNN) fusion layers. In each KNN fusion layer, for each audio frame embedding a_i of the utterance A, we query the external memory $E = (k_i, v_i), 1 \leq i \leq |E|$ for a set of m nearest neighbors:

$$\mathcal{N}_{a_i} = \underset{N \subset E, |N|=m}{\arg\min} \sum_{(k_j, v_j) \in N} ||k_j - a_i||_2^2 \tag{8.1}$$

We then construct the context for the layer as $\mathcal{C} = \cup_{a_i \in A} \mathcal{N}_{a_i}$. From $\mathcal{C}$ we can construct two matrices, $K_c \in \mathbb{R}^{m|A|, d_{\text{key}}}$ and $V_c \in \mathbb{R}^{m|A|, d_{\text{value}}}$, consisting of the keys and values respectively. The output of our K-NN fusion layer is then:

$$F(A, E) = A + \text{LN}\left(\text{ReLU}\left(\text{softmax}\left(\frac{(AW_q)K_c'}{\sqrt{d}}\right)(V_c W_v)\right)\right) \tag{8.2}$$

where LN is LayerNorm. Unfortunately, because we are working with large catalogs, the computation of Equation 8.1 can be very expensive. Thus, instead of computing the exact nearest neighbors, we rely on approximate nearest neighbors, which can be computed much more efficiently. To efficiently extract approximate nearest neighbors from our large-scale catalogs, we leverage the FAISS (Johnson et al., 2019) library to generate Optimized Product-Quantization-transformed keys (64 dimensions) (Ge et al., 2013), which are searched using a Hierarchial Naviagable Small Worlds (HNSW) index with 2048 centroids encoded with product-quantized fast-scan (Malkov and Yashunin, 2018). Such an approach leads to only a 15% increase in forward-pass latency, even when running with catalogs with over 7M key/value pairs.

8.1.3 Experimental Design

Catalog Data Sources: In our work we explore several different catalog data sources. For Librispeech, we build a simulated catalog using the 2500 rarest tokens present in either the training or test datasets. Our internal Alexa catalog focuses on assistant queries in a media domain, and consists of 15K movie titles. In both cases, we build a unique catalog for training and testing, allowing us to explore how well the model performs under distribution shift of the catalog at test time.

ASR Base Model: Although in practice our method could be applied to many different speech encoders, we use the Conformer encoder (Gulati et al., 2020). For the decoder, we use a 1-layer LSTM decoder with 320 hidden dimensions. While we explore several encoder sizes, we primarily follow Gulati et al. (2020) for Librispeech and use a 16-layer encoder with a hidden dimension of 144 (10.3M Params). For Alexa data, we use a conformer model with 208.37M parameters.

8.2 Results & Discussion

Librispeech: Our key results are shown for Librispeech in Table 8.1 for several choices of TTS, Text Embeddings, and NNs/Frame (K). We can see that overall, augmenting models with additional data leads to stronger performance than models without external data. For Librispeech, when training with the train catalog and testing with the test catalog, we get strong transfer performance, exceeding that of when we use the training catalog for both training and testing, suggesting additional zero-shot specialization. While 1-hot vectors outperform BERT vectors, we must train these vectors for each catalog, leading to an inability to do test-time specialization. BERT outperforms GLoVE in all cases (with GLoVE causing regressions on test-time specialization). Figure 8.4 demonstrates that our method can capture and apply domain data from the catalogs. In this experiment, the model is trained with a catalog containing 300K training-set unique bigrams, and we show the performance of this model using ten test catalogs, each consisting of 30K bigrams, taken either from the test set or dev set. As the fraction of bigrams in the test data that are available in the test catalog increases, the performance of the model improves – showing our approach can use the information in test catalogs effectively in a zero-shot learning setup.

We also run several ablations with Multivoice-TTS, BERT Embeddings and 8 NNs/Frame. Table 8.2 explores the performance of our approach when placing the external knowledge augmentation at different layers of the network. While using external knowledge in all layers is the most effective approach, we find that such an approach is latency-prohibitive, as it increases the latency of a forward pass of the model by ≈85%. Using a single layer increases latency by only ≈15%, while two layers increase latency by ≈23%. Table 8.3 explores the performance of the method on Librispeech for differing model sizes. As the number of parameters increases, the gains provided by external memory decrease.

Table 8.1: Word Error Rate on Librispeech data with a small (10.3M param) model. MV-TTS refers to Multivoice-TTS.

Catalog	**TTS**	**Text**	**K**	**test-clean**	**test-other**
Baseline				5.77	13.34
Train	Polly	1-Hot	4	5.75 (0.34%)	13.30 (0.29%)
	Polly	1-Hot	8	5.72 (0.86%)	13.19 (1.10%)
	Polly	1-Hot	16	5.71 (1.03%)	13.15 (1.42%)
	Polly	BERT	8	5.74 (0.52%)	13.26 (0.60%)
	MV-TTS	1-Hot	8	5.52 (4.33%)	12.96 (2.84%)
	MV-TTS	BERT	8	5.68 (1.63%)	13.05 (2.18%)
Test	Polly	GLoVE	8	6.33 (-8.84%)	14.56 (-9.15%)
	Polly	BERT	8	5.71 (1.03%)	13.24 (0.75%)
	MV-TTS	GLoVE	8	6.15 (-6.17%)	14.32 (-6.84%)
	MV-TTS	BERT	8	**5.34** (8.05%)	**12.84** (3.86%)

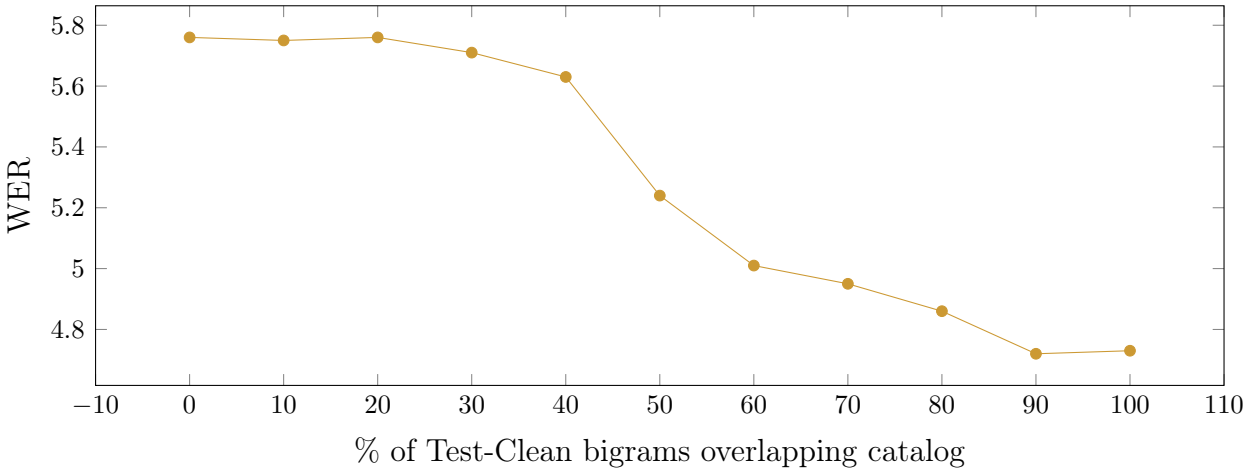
Table 8.2: Librispeech test-set Relative WER *improvement* for models augmented with catalog data in different layers.

Dataset	**1**	**3**	**12**	**16**	**3,12**	**all**
clean	1.02%	3.65%	6.65%	2.63%	7.79%	8.05%
other	0.71%	2.88%	2.97%	1.08%	3.41%	3.86%

Table 8.3: Librispeech test-set Relative WER *improvement* over baseline fine-tuning using differing model sizes (M: Millions of params).

Dataset	**5M**	**10M**	**50M**	**100M**	**300M**
clean	28.9%	8.05%	4.28%	1.66%	0.08%
other	19.3%	3.86%	2.65%	-0.07%	0.01%

Figure 8.4: Librispeech test-clean WER vs. test catalog/data overlap.



Alexa: To further validate our method, we additionally explore a real-world simulation of our model's ability to generalize to test data. We started with a seed model B, and trained two derived models: B_{FT}, fine-tuned on both the TTS Catalog for Alexa ($\mathcal{C}$, section 8.1) and an additional 120K hours of de-identified Alexa data, $\mathcal{D}$, and B_{cat}, which trains the proposed method on $\mathcal{D}$, with catalog $\mathcal{C}$. In both cases, the full model (the speech encoder, and if applicable, the fusion model) are fine-tuned. The results (Table 8.4, rows 1/2) demonstrate that even with significantly fewer GPU hours, our approach achieves similar WER. We then transfer both models to the test dataset (consisting of real speech) without additional tuning. We see in Table 8.4 (rows 3/4) that our trained model achieves better performance, suggesting that the model has learned to generalize better than the model trained with fine-tuning alone.

Finally, we update our fine-tuned and catalog models to include the test data, T. Test data is incorporated into the fine-tuned model through GPU-based training, while the test data is incorporated into the catalog model through catalog generation/concatenation. Table 8.4 (rows 5/6) demonstrates that even with *no additional GPU training* our approach ($B_{\text{cat}+T}$) achieves similar performance to fine-tuning ($B_{\text{FT}+T}$).

8.3 Conclusion

This chapter introduces the first approach for large-scale contextualization of speech-encoder representations using text-only catalog data. We strongly believe that our method represents a promising step forward for ensuring the recognition of rare words and efficient transfer novel test-time distributions. While this chapter is a first step towards contextualized speech encoders, many directions for future work remain including investigating embeddings for the catalogs (such as grapheme/phoneme embeddings), exploring other languages and word pronunciations, and understanding the performance in larger-scale rapidly changing

Table 8.4: Performance on Alexa data. T-C: Time for Catalog Generation. T-FT: Time for fine-tuning. WER-I: Relative word-error rate *improvement*. Multivoice-TTS, BERT, and 8 NNs/Frame.

Model/Test Data	**T-C** (min)	**T-FT** (GPU-Hours)	rel. **WER-I**
B_{FT}/Test-TTS	0	2048	7.1%
B_{cat}/Test-TTS	33	1600	6.8%
B_{FT}/Alexa Test	-	-	0.52%
B_{cat}/Alexa Test	-	-	4.12%
$B_{\text{FT}+T}$/Alexa Test	0	1024	19.66%
$B_{\text{cat}+T}$/Alexa Test	28	0	21.27%

real-world distributions.

Chapter 9

Task Oriented Dialogue as a Catalyst for Self-Supervised Automatic Speech Recognition

When users interact with assistant systems in task oriented ways, they build rich conversational contexts, which contain information that may be relevant to future requests along with feedback on the performance of the system. When users are dissatisfied, they express that intent in many ways, from direct corrections of the system response, to repeating and rephrasing the original question (Kwan et al., 2023). This discourse provides a source of contextual user interaction signals that are relatively untapped in Automatic Speech Recognition (ASR).

Indeed, traditional systems for ASR have primarily focused on single-utterances (Radford et al., 2023; Baevski et al., 2020; Hsu et al., 2021; Chan et al., 2022b, 2023a; Mitra et al., 2023), which, although flexible, overlook the wealth of contextual cues available in task directed dialogues. While work has been done in natural language understanding (NLU) to exploit these cues (Min et al., 2021), their potential in ASR has remained largely unexplored, primarily due to the limited availability of task-driven dialogue datasets in the audio domain (Chang et al., 2023; Si et al., 2023). Current efforts to integrate context from non-dialogue sources into ASR often involve training models explicitly with external per-turn contextual inputs, often leveraging context attention mechanisms (Chang et al., 2023; Kim and Metze, 2018; Chan et al., 2023a; Chang et al., 2021; Chen et al., 2019; Sathyendra et al., 2022; Wei et al., 2021; Yang et al., 2023; Chan et al., 2023b; Mahadevan et al., 2019). While per-turn context is important for the ASR task, these methods do not draw from dialogue structures, nor do they account for interactive feedback present in labeled dialogues.

Instead of directly training on per-sample or per-turn context (e.g., contact names (Chen et al., 2019; Sathyendra et al., 2022), or external dictionaries (Chan et al., 2023a)), we explore the potential of learning implicit contextual signals of user interactions, which remains relatively untapped in ASR-based dialog systems. Following work demonstrating the benefits of contrastive learning in ASR (Chan and Ghosh, 2022b), closest to our work may be Chang

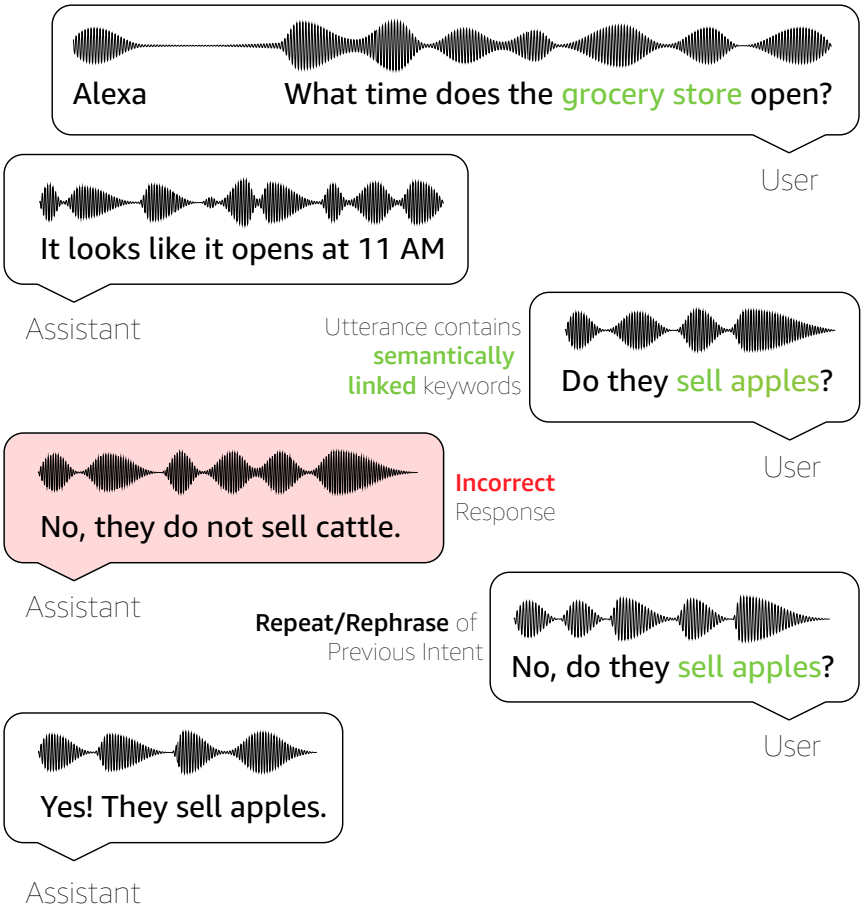


Figure 9.1: Task oriented dialogues can contain a multitude of relevant information for performing automated speech recognition. In this work, we explore how we can learn from both semantically linked keywords within dialogues, and failed dialogue turns.

et al. (Chang et al., 2023) who propose reducing ASR errors with contrastive learning between noisy and clean audio transcripts from task-oriented dialogues – however, their work focuses only on single turns of dialogues, not contextual dialogue cues. Our primary contributions are:

- We propose a new family of self-supervised fine-tuning losses, CLC, which incorporate self-supervised information from task oriented dialogues (TODs), and show that learning from TODs, even those with errors, provides benefits over fine-tuning.
- We introduce a new semi-synthetic benchmark meta-dataset, the Open Directed Dialogue Dataset (OD3), designed to enable further research in conversational interactions for ASR.

9.1 Contrastive Learning for Conversations

In this chapter, we introduce two novel auxiliary losses, termed "Contrastive Learning for Conversations" (CLC), designed to enable learning from both successful and unsuccessful task-directed conversations with assistants (section 9.1), as well as a new synthetic dataset for the evaluation of contextual automated speech recognition models in task directed domains (subsection 9.1.1).

Learning from Past and Future Dialogues: As shown in Figure 9.1, utterances in a dialogue can contain important contextual hints useful for recognizing low-frequency words in the sentence. While we may not have access to past or future utterances at inference, we can often learn from these hints during training. The first auxiliary loss we introduce follows this key motivation; auditory information within a dialogue should share more semantic and representational overlap than auditory information from a second, unrelated dialogue.

This insight induces a natural contrastive loss: the speech encoder representations of audio within a session should be closer in the latent space (on average) than the representations between sessions. To implement a "Past-Future" contrastive loss, we consider the utterances $u_1, \ldots, u_N$ in a dialogue. Let the speech encoder be defined as $e_i = \epsilon(u_i) \in \mathbb{R}^{T \times k}$, where k is the dimension of the speech encoder embedding, and T is the number of frames of audio in the dialogue. We further introduce three "head" encoders, $\xi_{past}(e_i) \in \mathbb{R}^d, \xi_{current}(e_i) \in \mathbb{R}^d, \xi_{future}(e_i) \in \mathbb{R}^d$, which embed the sequential embeddings from the encoder ϵ of the current, past, and future frames into single vectors (of dimension d) representing the current, past, and future contexts. These head encoders take the form of global pooling followed by two layers of a shallow MLP with ReLU activations, LayerNorm, and Dropout. We can then compute the following contrastive loss terms (similar to (Khosla et al., 2020)) for a batch of $1 \leq i, j \leq N$ samples (where embeddings are L2-normalized):

$$L_{future}^{i,j} = -\log \left[\frac{\exp(\xi_{current}(e_i) \cdot \xi_{future}(e_j)/\tau)}{\sum_{k=1}^{N} \exp(\xi_{current}(e_i) \cdot \xi_{future}(e_k)/\tau)} \right]$$

$$L_{past}^{i,j} = -\log \left[\frac{\exp(\xi_{current}(e_i) \cdot \xi_{past}(e_j)/\tau)}{\sum_{k=1}^{N} \exp(\xi_{current}(e_i) \cdot \xi_{past}(e_k)/\tau)} \right]$$

The "Past-Future" auxiliary loss is then a weighted sum:

$$L_{pf} = \frac{1}{N} \left[\alpha \sum_{i=1}^{N} L_{future}^{i,i} + \beta \sum_{i=1}^{N} L_{past}^{i,i} \right] \tag{9.1}$$

Here, we choose cosine distance (the dot-product) as the similarity function. The result of the loss function is that we aim to maximize the mutual information between the encoding of the current, future and past frames within a dialogue, while minimizing the mutual information between the current frames and frames from other dialogues. Note here that

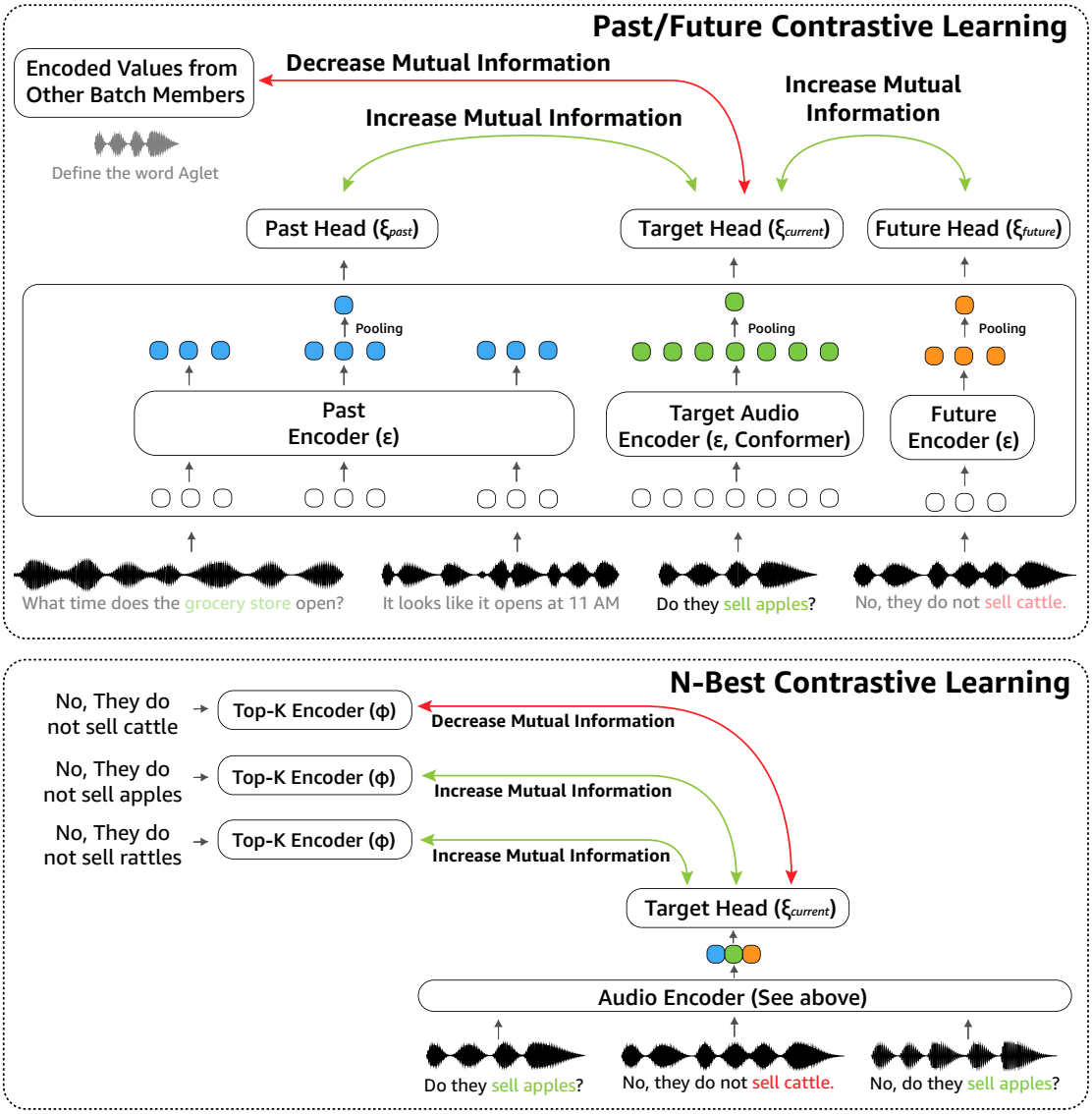


Figure 9.2: Overview of CLC approaches. The Past-Future loss maximizes agreement between current, past, and future embeddings. The N-best loss minimizes agreement between current embeddings and top predictions of rephrases, while maximizing agreement otherwise.

it's important that $e_i \neq e_j$, that is, the embedding of the past should not be identical to the embedding of the future (as they have different ASR content). Instead, we encourage high mutual information between the different segments, by leveraging contrastive learning on projection heads stemming from the shared representation. α and β are hyper-parameters which control the strength of the binding in the loss function, and τ is a temperature parameter. In our experiments, we found through grid hyper-parameter search of $\alpha, \beta \in [0.0001, 100]$ (logarithmic sweep) and $\tau \in [0.1, 1]$ (linear sweep) that $\alpha = 1.0, \beta = 0.7, \tau = 0.1$ is the most effective.

Learning from Failures: We can extract valuable semantic information from conversations, even those that don't proceed smoothly. It is often possible to detect dialogues where unsuccessful ASR has triggered repeats and rephrases of previous content by understanding when subsequent user turns have high semantic overlap, or tracking NLU failures in downstream systems. In these cases, we can further leverage contrastive learning to improve the performance of the model. Ideally, when there is a repeat or rephrase in a dialogue, we want to reduce the mutual information between the conformer encoder embedding of the initial turn triggering the repeat or rephrase, and the answer produced by the model in that dialogue. While we could use reinforcement learning to optimize for this signal (and it is interesting future work to do so), we often train models offline, and as the model trains, its decisions deviate from the original policy, leading to a breakdown in the learning process. Instead, as we know the "bad" solution, we can use supervised contrastive learning (Khosla et al., 2020) to improve the model. When there is no rephrase, we want to increase the mutual information between the semantics of the top-1 prediction of the model and the current frames. When there is a rephrase, we want to decrease the mutual information between the semantics of the top-1 prediction, and instead encourage the model to produce a different output from the top-k. While it is possible that worse hypotheses with high similarity exist in the hypothesis set (leading to incorrect labels), we observe empirically that our models have high oracle WER, allowing this method to achieve a weak approximation to oracle re-ranking of the candidate set, which improves overall performance when smoothed over a large training set.

An overview of our n-best approach is given in Figure 9.2. For each sample u_i, let $\phi_1(u_i) \dots \phi_K(u_i)$ be the semantic embeddings of the top-k predictions of the i'th utterance (using beam-search decoding) and $\xi_{current}(e_i)$ be an embedding of e_i for u_i. Using a similar set of heads to the network above, we compute positive and negative losses:

$$L^i_{pos} = -log \left[\frac{\exp(\xi_{current}(e_i) \cdot \phi_1(u_i)/\tau)}{\sum_{k=1}^{K} \exp(\xi_{current}(e_i) \cdot \phi_k(u_j)/\tau)} \right]$$

$$L^i_{neg} = -log \left[\frac{\max_{j \neq i} [\exp(\xi_{current}(e_i) \cdot \phi_j(u_i)/\tau)]}{\sum_{k=1}^{K} \exp(\xi_{current}(e_i) \cdot \phi_k(u_i)/\tau)} \right]$$

Let $\mathcal{R}$ be the set of utterances which trigger a repeat/rephrase, and $\mathcal{S}$ be the set of utterances which are considered successful. We can then combine the positive and negative losses as

follows:

$$L_{nbest} = \frac{\gamma}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} L^i_{neg} + \frac{\kappa}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} L^i_{pos} z \tag{9.2}$$

where γ and κ are hyper-parameters controlling the trade-off between negative and positive reinforcement. Discovering the sets $\mathcal{S}$ and $\mathcal{R}$ can be challenging, however, we can detect repeats and rephrases with relatively high accuracy using semantic vector matching (such as matching BERT embeddings). Using grid search with $\gamma, \kappa \in [0.0001, 100]$ (logarithmic sweep), we found $\gamma = 0.1, \kappa = 1.0$ was most effective.

9.1.1 Data

While a predominant portion of interactions with assistant systems revolves around task-directed dialogues, the availability of datasets (Table 9.4) encompassing task-directed audio interactions remains quite limited. Moreover, even within datasets that do incorporate such interactions, a conscious effort has been exerted to remove flawed turns (turns in which a dialogue assistant responds incorrectly, and must be corrected by the user). To evaluate our CLC methods for self-supervised fine-tuning, we use two datasets: a private collection of de-identified real-world conversations with a conversational assistant, and a new semi-synthetic meta-dataset, OD3, replicating flawed conversations often seen in real-world assistant interactions. The OD3 dataset is released as part of this work under the CC-BY-NC-SA (4.0) license.

9.1.1.1 Real-World (Alexa) Data

To demonstrate the performance of our method, we train and evaluate our models on 130K hours of de-identified agent-centric task-directed dialogues constructed from independent interactions with Amazon Alexa. These dialogues have a maximum of five utterances each (with an average of 1.2 turns per goal). Dialogues are constructed around a seed utterance by collecting interactions within $\rho = 90$ seconds on each size of the utterance. This process is repeated recursively until there are no more interactions. In the case that there are more than five utterances, we halve ρ, and repeat the process. This continues until either we have less than 5 utterances in the final set or we hit a minimum time gap of 15 seconds. During testing, only the past and current context is available to the model (the future remains hidden).

9.1.1.2 OD3: A new dataset for conversational learning

In addition to the results on real-world interactions in this chapter, we further introduce a new semi-synthetic meta-dataset, OD3 (Open Directed Dialogue Dataset), which is designed to allow the community to explore further research into leveraging flawed conversational interactions to improve model performance. OD3 is a collection of 63K conversations (600K turns, 1,172 hours of audio) drawn from existing natural language task-oriented dialog datasets, and augmented with synthetic audio. OD3 is further augmented with turns containing *repeats*

and *rephrases* of previous failed utterances. We compare our dataset with some others in the field in Table 9.1.

Constructing OD3: To construct OD3, we start with several seed datasets of natural language task oriented dialog data: KVRET (Eric et al., 2017), Multi-Woz (Budzianowski et al., 2018), DSTC11 (Track 5) (Zhao et al., 2023), NOESIS-II (Gunasekara et al., 2020) and SIMMC-2.1 (Kottur et al., 2021). Here we focus on multi-turn dialogue, (as opposed to single-turn datasets such as those for question-answering like NMSQA (Lin et al., 2022)), as they contain the most relevant contextual information. This gives us a pool of ≈ 63K unique dialogues (≈ 597K turns) containing no explicitly labeled errors or flaws. Because these datasets are not augmented with audio for each of the conversational turns, we leverage the NeMo Text Normalizer (Kuchaiev et al., 2019) and the YourTTS method (Casanova et al., 2022) (voice cloning) to generate audio for each of the conversations. In all of the conversations, we hold the voice for the agent constant, and each voice used in voice cloning is randomly selected from the English subset of Common Voice (Ardila et al., 2020) (which is CC0 licensed). We found that in some cases, the TTS induces errors in the generated speech, which we found correlated with a high number of deletions in the resulting ASR models. To clean the dataset, we filter out ≈4K utterances inducing a significant number of deletions in both our tested and third party ASR models. While we run our experiments in this chapter on the clean data, we additionally release the noisy versions of the data as they could be useful for investigation into alternate directions of research.

We synthetically introduce errors and noisy conversations into the data. For that, we first compute ASR for each dialog turn using OpenAI's Whisper Large (v2) model (Radford et al., 2023). We consider conversational turns with WER higher than 15% candidates for the injection of either a *repeat*, or a *rephrase* of the intent. We then insert repeats and rephrases into 20% of the possible candidate conversations. To insert a *repeat*, we introduce two conversational turns: a response for the agent which is a non-specific error response (such as "I'm sorry, I don't understand"), and a repeat of the phrase which triggered the ASR errors (re-sampled from the original TTS model). Inserting a *rephrase*, on the other hand, is much more complicated. Similar to the case of repeats, we first introduce a non-specific agent error message. We then generate a rephrase of the original triggering utterance using in-context learning with the MPT-30B language model (Team, 2023), combined with the prompt: `Our automated speech recognition model found "<input string>" hard to parse, so we rephrased it to use easier to understand words as "...`

We found that this prompt generated reasonable rephrases of the candidate sentences. For example, *"Are there noisy neighbors?"* was rephrased as *"Is the place quiet enough?"*. This gives us a total of ≈ 625K turns of dialogue in ≈ 62K sessions, and 1,172 hours of audio.

9.1.2 Models

For the speech encoder ϵ, we use a conformer architecture (Gulati et al., 2020), with 17 layers, latent dimension of 1024, and two stride-two convolutional sub-sampling layers (≈

Table 9.1: Statistics for OD3. OD3 is much larger than existing TOD datasets, while including both audio and noisy conversations.

Dataset	Dialogues	Turns	Audio	Errors
DSTC-2 (Henderson et al., 2014)	1,612	23,354	✓	
KVRET (Eric et al., 2017)	2,425	12,732		
MultiWOZ (Budzianowski et al., 2018)	8,438	115,424		
DSTC-10 (Kim et al., 2021)	107	2,292		
SpokenWOZ (Si et al., 2023)	5,700	203,074	✓	
OD3 (Ours)	**62,974**	**623,145**	✓	✓

200M parameters). We use a 1-layer LSTM decoder with latent dimension of 320, and a 4K token vocabulary. The encoder/decoder are initialized with a model pre-trained on 120K hours of de-identified Alexa seed data. During training, we apply both kernel regularization and bias regularization with weight $1e^{-6}$, and dropout with weight 1.0. We optimize the overall loss:

$$L_{overall} = L_{asr} + \lambda L_{pf} + \delta L_{nbest} \tag{9.3}$$

The models are trained for at most 120 epochs with the Adam optimizer, following a linear increase, hold, exponential decay learning rate schedule starting at $1e^{-8}$, increasing linearly over 50K steps to hold at $4e^{-5}$ for 250K steps, and then decay back to $1e^{-6}$ over a further 300K steps. We use gradient clipping with limit 0.3, and a dynamic batch size (depending on input feature length) ranging between 128 and 1024. As contrastive learning cannot naively be scaled across GPUs, we leverage techniques similar to BASIC (Pham et al., 2023) and perform memory efficient contrastive mini-batching.

9.2 Results & Discussion

We first demonstrate the performance of our method on the Alexa session data. From the results in Table 9.2, we can see that all three settings of CLC improve the overall WER/SER of the model, particularly over zero-shot models. We notice that setting $\lambda = 1$ is the most effective at reducing overall WER, as in most situations, contextual information from previous (and future) turns can provide more powerful hints to the content of an utterance. While δ is helpful as well, it is less important to overall WER.

Table 9.3 shows the performance of CLC across different values of α and β for L_{pf}. We can see that taking into account both past and future information is important. Unsurprisingly, past information is a more powerful indicator of the current ASR context; however it's important to note that pre-training with the information from the future allows the model to improve the predictive ability of the audio representations, leading to improvements (particularly in SER). Table 9.3 also shows the performance for values of γ and κ in the L_{nbest} loss. We can see here that placing too much weight on the γ term leads to a destabilization of

Table 9.2: Results on Alexa data, both overall and only on turns inducing repeats or rephrases. WERR (↑): Percent relative WER Improvement. SERR (↑): Percent relative SER improvement.

Model	**Overall**		**Repeats/Rephrase**	
	WERR	**SERR**	**WERR**	**SERR**
Zero-Shot (No Fine Tuning)	-23.02%	-17.46%	-4.65%	-5.75%
Baseline (Fine Tuned)	-	-	-	-
CLC ($\lambda = 1, \delta = 0$)	2.75%	2.88%	3.0%	3.39%
CLC ($\lambda = 0, \delta = 1$)	2.60%	2.39%	3.75%	3.87%
CLC ($\lambda = 1, \delta = 1$)	4.31%	3.88%	5.07%	5.31%

Table 9.3: Results on Alexa data for different values of α and β ($\tau = 0.1$) in L_{pf}, as well as γ and κ in L_{nbest} for small scale (batch size 128) experiments. WERR (↑): Relative WER Improvement. SERR (↑): Relative SER improvement.

Model	**WERR**	**SERR**
Baseline (CLC, $\lambda = 0, \delta = 0$)	-	-
CLC ($\alpha = 1, \beta = 0, \gamma = 0, \kappa = 0$)	3.28%	2.26%
CLC ($\alpha = 0, \beta = 1, \gamma = 0, \kappa = 0$)	2.74%	3.68%
CLC ($\alpha = 1, \beta = 1, \gamma = 0, \kappa = 0$)	4.50%	5.34%
CLC ($\alpha = 1, \beta = 0.7, \gamma = 0, \kappa = 0$)	5.17%	4.67%
CLC ($\alpha = 0, \beta = 0, \gamma = 1, \kappa = 0$)	-11.81%	-10.43%
CLC ($\alpha = 0, \beta = 0, \gamma = 1, \kappa = 1$)	-1.88%	-2.21%
CLC $\alpha = 0, \beta = 0, \gamma = 0, \kappa = 1$)	6.23%	5.59%
CLC ($\alpha = 0, \beta = 0, \gamma = 0.1, \kappa = 1.0$)	6.77%	6.25%

Table 9.4: Results on the OD3 dataset (overall and repeat/rephrase inducing). WER (↓): Word Error Rate, BERT-S (↑): Bert Score.

Model	Overall		Repeat/Rephrases	
	WER	**BERT-S**	**WER**	**BERT-S**
Baseline (206M)	11.13	0.9762	16.17	0.9690
CLC ($\lambda = 1, \delta = 0$)	9.57	0.9801	14.12	0.9702
CLC ($\lambda = 0, \delta = 1$)	9.38	0.9803	13.94	0.9721
CLC ($\lambda = 1, \delta = 1$)	8.99	0.9812	13.81	0.9737

Table 9.5: Zero-shot results on OD3 for several open-source models. Models in this table are not directly comparable (trained on differing data, setups, hyperparameters, optimizers etc.), but serve as a benchmark for performance on OD3 under several varying setups. WER (↓): Word Error Rate, BERT-S (↑): Bert Score.

Model	Overall		Repeat/Rephrases	
	WER	**BERT-S**	**WER**	**BERT-S**
CLC best model	8.99	0.9812	13.81	0.9737
Whisper S (200M) (Radford et al., 2023)	11.24	0.9775	14.17	0.9727
Whisper L (1.3B) (Radford et al., 2023)	8.51	0.9852	12.37	0.9792
Conformer (100M, Librispeech) (Gulati et al., 2020)	19.26	0.9612	22.19	0.9571
Wav2Vec 2 (433M, Librispeech) (Baevski et al., 2020)	19.41	0.9582	22.03	0.9544
Streaming Conformer (45M) (Tsunoo et al., 2021)	14.38	0.9701	16.70	0.9665

the loss, however small magnitude γ values can help with overall performance. We believe that this destabilization is caused by the high variance of the $\max_{j \neq i} [\exp(\xi_{current}(e_i) \cdot \phi_j(u_i)/\tau)]$ term, and it is future work to explore how functional implementations such as a soft-max could reduce the gradient variance stemming from this loss term.

Table 9.2 also shows the performance of our method when restricted to only defective utterances: utterances triggering repeats and rephrases in the dataset. We can see that setting $\delta = 1$ is helpful, since the additional losses nudge the model away from high-confidence decisions in detected repeats/rephrases and makes an impact on the model's ability to correctly recognize challenging samples. Note that WERR/SERR gains are statistically significant over the large-scale test set ($\approx 1K$ hours of test audio).

On OD3, our approach produces even more defined results, demonstrated in Table 9.4, where our model produces a 19.22% improvement over baselines, clearly showing how learning from additional contextual clues can benefit ASR models. Interestingly, despite a high word error rate, the semantic similarity, as indicated by the BERTScore (Zhang et al., 2020e) remains high — this suggests that ASR errors, while numerous, do not significantly impact the semantic meaning. Several major questions remain unanswered for future work, for example, it remains an open question how the approaches scale with model parameters, as well as

understanding to what extent different mixes of pre-training data alter the performance of the model.

Even for models with strong language models, large vocabularies, and training data focused on open-domain conversational language, Table 9.5 shows that OD3 is challenging. Models demonstrated increased insertions and substitutions, as there are a large number of challenging low-frequency words that must be recognized accurately. It's interesting to see that the streaming conformer (Tsunoo et al., 2021) (trained on Gigaspeech) outperforms some of the larger models. This is likely due to the training data mix: training smaller models on more robust datasets is more effective than training larger models on sparse or biased data.

9.3 Conclusion

This chapter introduces CLC, a self-supervised fine-tuning approach for enhancing contextual automated speech recognition (ASR) in task-oriented dialog systems. We also introduced OD3, the largest-ever dataset for task-oriented automated speech recognition. By leveraging both successful and unsuccessful conversational interactions, our method enhances the underlying ASR model's ability to handle challenging and contextually rich utterances. In real-world data, we demonstrate as much as 6.77% improvement over baselines. Further, for OD3 we show up to a 19.22% improvement over baselines. We hope that our approaches and datasets will help address ASR challenges within intricate and error-prone dialog settings, elevating user experiences and enabling more effective interactions between humans and AI agents.

Chapter 10

Discussion

In this section, we looked at how context (both external and intrinsic) can be used to enhance the capabilities of CNLG models. chapter 6 demonstrated that understanding the users themselves, and how to combine users can lead to strong models for visual description. chapter 7 showed that we can use video context during training to improve ASR performance (even if that video isn't present at test time). chapter 8 showed that not only can we use video, but we can use text as well. chapter 9 showed that there is an implicit signal that we can extract from conversations.

In general, while this section demonstrated that context, both internal and external can be useful in improving model performance, there is still significant work to be done in the areas of integrating contextual signals with CNLG models. Several key areas stand out to me:

1. **Grounded Context Models:** Currently, the models that we have introduced take the context as input, but they are not able to directly ground the output to the sources of context. I.e. models are incapable of directly citing their sources beyond a prompt-based request to do so. Developing methods such as Grad-cam (Selvaraju et al., 2017), which can directly visualize how tokens in the output are related to tokens in the context may help build more explainable models, and models that are more trustworthy to users.

2. **Increasing the size, and scope, of contexts:** In this work, we explored contexts of several sizes, ranging from only a single video source to hundreds of thousands of catalogue entries. I believe that one of the most interesting directions for future work is how we build retrieval-augmented models to build the contexts that they operate on. Some questions include *How do we build models that are query aware, and efficient, while retrieving the relevant context with high recall?* and *How do we encode context in such a way that is token efficient?* (to avoid constraints with model context length). Further, recent work has shown that not all context tokens are created equal: Liu et al. (2023), Li et al. (2023a) and Sun et al. (2021) have all shown that not every token is considered equally by an LLM, and it remains an open question to understand what causes these attentional blindnesses and to correct for them in model performance.

3. **Multi-source context models:** In the work so far, we have focused on including only a single source of context at any time. In practice, however, it would make sense to combine all of these sources of context to build models that are capable of contextual learning from multiple different sources, and multiple different inputs. It remains to be seen how such a model would work – are the contexts multiplicative? Do they include redundant information? Such questions require significant additional research to fully understand.

4. **How much does context learning help in the absence of context?** Context is not always available to a model. Perhaps it is because the context was never collected by the UI, or perhaps it is because the model doesn't have all sources of information available at every time. It remains a very interesting question to understand how our context models function in the absence of context. In chapter 7, we showed that by pre-training with context, even in the absence of context during test time, we saw increased performance. Such performance also held in the experiments in chapter 9, suggesting that learning with context seems to help models, even when that context isn't present. Why is this the case? And what can we do to leverage this effect efficiently? Some initial exploration is provided in our work on hallucination (Jhamb et al., 2022), but this question remains open, and of great interest.

This section only provides a glimpse of what is possible for contextually aware models for conditional natural language generation. In chapter 14, we discuss even more visions for contextually aware models, as I strongly believe the potential is endless, and we have only begun to scratch the surface of the problem.

Part III

Evaluating

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.
>
> *William Thomson (Lord Kelvin)*

Evaluating models for conditional natural language generation (CNLG) is not only crucial for measuring their performance and utility to downstream users, but also for understanding their limitations, biases, and ethical implications of their deployment in real-world applications. Usually, when we evaluate our CNLG models, we make use of (1) A large dataset of human-annotated context/response samples and (2) a natural language distance that determines the distance between the model-generated output and the reference responses for that context. But as we have discussed in earlier sections, in many cases CNLG is not a classification problem: i.e. there are many possible valid outputs, with the particular outputs in the reference set largely depending on the human who generated that reference, and the context in which the references were collected. Unfortunately, traditional automated metrics for evaluating conditional natural language generation rely on pairwise comparisons between a single generated text and the best-matching gold-standard reference. This method is effective when ground truth data diversity can be attributed to noise, however, it falls short when diversity in references holds valuable contextual information, as in visual description or summarization, as it does not evaluate the ability of a model to generate text matching the diversity of the ground truth samples. Thus, we have to ask the question: how can we evaluate conditional NLG models when our datasets have only a small sample of the possible outputs (sometimes even just one human reference)? In this section of the dissertation, we investigate two novel methods for evaluation in such scenarios.

First, in chapter 11, we challenge the adequacy of existing metrics in semantically diverse contexts and introduce a novel approach for evaluating conditional language generation models, leveraging a family of meta-metrics that build on existing pairwise distance functions. These meta-metrics assess not just single samples, but distributions of reference and model-generated captions using small sample sets. We demonstrate our approach through a case study of visual description which reveals not only how current models prioritize single-description quality over diversity, but further sheds light on the impact of sampling methods and temperature settings on description quality and diversity.

Next, in chapter 12, we explore how we can leverage the latest advances in large language models to judge the distance between captions while accounting for the variance that is generated by user contexts. In our evaluations, CLAIR, our novel method that leverages the zero-shot language modeling capabilities of large language models (LLMs), demonstrates a stronger correlation with human judgments of caption quality compared to existing measures. Notably, on Flickr8K-Expert, CLAIR achieves relative correlation improvements over SPICE of 39.6% and over image-augmented methods such as RefCLIP-S of 18.3%. Moreover, CLAIR provides noisily interpretable results by allowing the language model to identify the underlying reasoning behind its assigned score.

Underlying both of these approaches is a key idea: *conditional natural language generation is not a classification problem.* In this section, we show that instead of treating it as such, by treating CNLG as a problem with several correct answers we can more closely align model evaluation with the multifaceted nature of human communication, thereby enhancing the utility, fairness, and transparency of CNLG models in serving diverse human needs.

Previously Published Works Appearing In This Section:

1. Chan, David M., et al. "Distribution Aware Measures for Conditional Natural Language Generation." Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. 2024.

2. Chan, David, et al. "CLAIR: Evaluating Image Captions with Large Language Models." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Chapter 11

Triangle-Rank Metrics for Distribution Aware Conditional Natural Language Generation

Recent models for conditional language generation, particularly in the field of visual description, have shown dramatic improvements in both fluency and the ability to ground generated language in context (Liu et al., 2021a; Zhou et al., 2020; Mokady et al., 2021; Chen et al., 2018). Standard metrics for these tasks such as BLEU, ROUGE, METEOR, and CIDEr, compare a generated text with a reference set of texts and compute some measure of quality for the generated text. By construction of these metrics, a model will achieve the best performance by generating a single high-scoring text. In contrast, it has been widely observed that large language models such as GPT-3 (Brown et al., 2020) or LAMDA (Thoppilan et al., 2022) generate the most realistic texts at temperatures close to one, where the set of potential texts generated is often very diverse. More significantly, if we look at an example of an image from MS-COCO and its set of reference captions (Figure 11.1), we notice that each (human-generated) reference contains a unique subset of the overall information in the image:

> "A woman in a red robe is sitting at a dining table."
> "A woman in a red flowered shawl sits at a table while a man wearing jeans is in the kitchen looking at her."
> "A person sits at a table and another person stands in the kitchen."
> "A woman is sitting at a table wearing a robe while a man is cooking."
> "Man and woman in a kitchen looking in the same direction."

Important features like the red robe, the man, the gaze of the two people etc, are mentioned only in one or a few captions. Metrics that encourage generating information from *only one* of these captions will generally fail to capture much of the important detail in the image. This holds for more than just image description. For many conditional language generation tasks such as video captioning, abstractive summarization, translation, and open-ended question-answering, it is often beneficial to be able to sample from a diverse distribution of generated

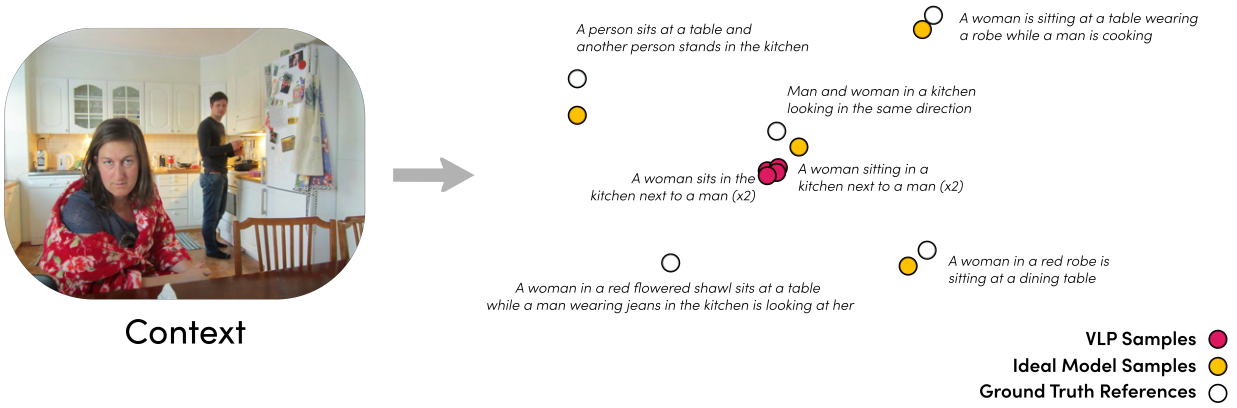


Figure 11.1: Samples from these two models achieve similar BLEU scores, however, the samples from a SOTA model (VLP) lie near a center of the distribution, and fail to capture the dispersion of natural language in the ground truths, while the samples from an ideal model better match the ground truth distribution. In this work, we introduce metrics which better measure deviations between samples from candidate and reference distributions, compared to single-sample pairwise metrics.

outputs. If we compute a maximum-likelihood generated caption from a state-of-the-art model (Zhou et al., 2020) we get:

"A woman sitting in a kitchen next to a man."

In this description, we see that only information common to most or all of the reference captions is preserved. This is intuitive, since including more information runs the risk that no reference caption contains that information, leading to a low score. It seems the designers of metrics such as BLEU are already aware that direct use of shortest distance to a reference caption favors generated captions which are even shorter and more impoverished, and thus, the BLEU score, and many others, also include a term encouraging longer texts. However, the (log-) text length heuristic in standard metrics is intuitively a poor proxy for actual diversity. Thus, since models optimize for standard measures, drawing multiple maximum-likelihood samples using beam search from SOTA models only produce repetitions, or slight variations of the above caption.

Thus, we encounter an issue in the evaluation of conditional text generation models with multiple available references. With multiple references, typically the metric score is based on the maximum score over a set of ground truths (e.g. max pairwise score for a particular n-gram as in BLEU), leading measures to erroneously incentivize the production of text minimizing the expected pairwise distance to the reference set, i.e. near a strong mode in the training text distribution, causing the issues discussed above. Changing the metric aggregation method (e.g. sum as in ROUGE) does not substantially alter this situation, as the model still strives to produce a high-scoring output that is close to nearby references which will be maximized at a smoothed mode in the training text distribution (Caglayan

et al., 2020; Yeh et al., 2021).

An over-reliance on simple aggregations for multiple candidates and references has, over time, compounded into several issues: The first, discussed further in section 11.2, is that, as observed in visual description by Chan et al. (2022c) and dialog generation by Caglayan et al. (2020), human-generated captions tend to receive lower scores than model-generated captions using automated measures, even though they actually receive higher scores under human evaluation. The second, discussed in section 11.1, is that diversity of candidate texts is largely relegated to reference-unaware measures, encouraging models to diverge from ground truth distributions to hit diversity targets.

In this chapter, we aim to solve these problems by introducing several novel automated ways of measuring the performance of conditional text generation models. Our measures encourages models to not only to generate samples at the locus of a distribution but also with sufficient variance, since they are designed computing the divergence between candidate and reference *distributions*. While some recent methods have been designed to closely measure the divergence between full distributions of text data in the unconditional case (Pillutla et al., 2021), no such methods exist for conditional generation, which often operates on the level of 10s of reference samples and candidates. Our contributions are summarized as follows:

1. We demonstrate that existing automatic metrics that use simple aggregations of candidate and reference distributions are insufficient, and we introduce a new paradigm that instead involves sampling from these distributions, and comparing the samples.
2. We introduce two new families of metrics which *extend* existing semantic distances: triangle-rank metrics, and kernel-based metrics, designed to measure the divergence between small text samples from candidate and reference distributions.
3. We explore how our new metrics behave in the context of visual description (both image and video description) and show that by measuring distributional effects, we can capture nuances in the data that existing metrics cannot explore.

11.1 Related Work

This work is not the first to notice the shortcomings of traditional metrics for the automated evaluation of conditional language generation models. In visual dialog, Caglayan et al. (2020) find that a number of the automated metrics proposed for visual dialog do not match well with human judgment, while in visual description, Chan et al. (2022c) find that current automated metrics do not assign high scores to human-generated descriptions. This work not only quantifies such issues but proposes a method for addressing these cases without developing novel metrics for measuring text semantic distance. In this section, we review related works, roughly divided into three groups; methods for evaluating text quality, text diversity and distribution aware metrics.

Measuring the Quality of Generated Text The evaluation of machine-generated text has long been an active area of research, which has continuously evolved to keep pace with

accelerating advances in text generation. As a consequence of the tools available and the state of early text generation approaches, classical measures have primarily focused on evaluating the quality of generated text with respect to ground truth references using surface-level text statistics. Most notably, these include n-gram matching based metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). More recently, the rapid progress enabled by large-scale language models has motivated new evaluation techniques which go beyond superficial n-gram statistics and toward measures that aim to capture the underlying semantics of language (Shimanaka et al., 2018; Clark et al., 2019; Zhang et al., 2020d; Sellam et al., 2020). These approaches leverage high dimensional representations of generated and reference text provided by a state-of-the-art language model, such as BERT (Devlin et al., 2019) in the case of BERTScore (Zhang et al., 2020d) and BLEURT (Sellam et al., 2020). While such methods are focused on measuring the semantic distance between two pairs of natural language texts, the evaluation of the diversity of the generated captions has largely been done independently of quality.

Measuring the Diversity of Generated Text Until recently, measures of diversity for generated text have been largely secondary to measures of quality, since the pursuit of human-like generated text has been the primary focus of the field. In fact, many diversity measures quantify surface-level statistics of the generated text (van Miltenburg et al., 2018), such as metrics based on the number of unique tokens, unique sentences, or unigram frequency statistics, such as Zipf coefficients (Holtzman et al., 2020). Similarly, n-gram-based diversity measures such as self-BLEU (Zhu et al., 2018), compute scores between samples from a model. Unfortunately, these approaches do not consider the diversity of a model's outputs with respect to the diversity of human references, and are primarily focused on the diversity of the vocabulary, rather than the aggregate semantic diversity, factors that our proposed work aims to address.

Distribution Aware Measures of Generated Text MAUVE, proposed by Pillutla et al. (2021), measures the divergence between multi-candidate samples and multiple ground truths using density estimates in a text embedding space. This approach measures both text dispersion and quality simultaneously, however, MAUVE is designed for unconditional text generation with many thousands of candidate and ground truth samples available. While MAUVE works well in these scenarios, it does not work well when only a few references are available (due to the K-means approximation) (see appendix C.2.4). Such a low-reference scenario is common in conditional NLG, making MAUVE unsuitable for many potential applications, and motivating the need for more sensitive measures.

11.2 Distribution Aware Measures for Conditional NLG

In this section, we introduce our two primary contributions. First, we introduce and demonstrate the need for a paradigm for multiple candidate evaluation for conditional language generation, and second, we introduce several simple augmentations to existing pairwise metrics, designed to alleviate the sensitivity issues induced by evaluating conditional language generation models with only a single candidate text. Our family of augmented metrics, which we call Triangle-Rank Metrics (TRMs), represents the first step towards optimizing metrics that force models not only to generate samples at the locus of a distribution but also with sufficient variance, hopefully alleviating the field-wide issues that optimizing standard pairwise-metrics can induce.

11.2.1 Multi Candidate/Reference Evaluation

Traditionally, most methods for conditional language generation have been designed to sample a single candidate example using beam search, designed to be a maximum likelihood sample of the data. This single candidate is compared against the reference data. Unfortunately, as discussed earlier, models can easily exploit such aggregations. For example, when the best score amongst the ground truths is chosen (the "min-distance" aggregate), models generate texts optimizing the *expected minimum distance to the reference distribution*. Such a text is, by definition, the mode of the distribution. This mode likely represents some amount of central tendency, as we observe such captions to be bland and uninformative (See C.2.5, (Chan et al., 2022c; Yang et al., 2019)).

Thus, a single candidate may not be sufficient to understand if the model has learned to approximate the reference distribution. Consequently, we aim to develop methods that can sample several suitable candidate texts, each with high accuracy, while matching the diversity of the ground truth distribution. In this work, to extend methods to multiple candidate generation, we leverage temperature-based sampling or nucleus sampling (as indicated) to produce multiple candidates from each model's distribution. While beam search can generate multiple candidates, Vijayakumar et al. (2016) showed diversity among beams is relatively poor, leading to samples that diverge from the model distribution. This gives us a model which *generates multiple candidate samples*, and requires an evaluation metric which *compares multiple candidate samples to multiple reference samples*.

Extending Existing Metrics for Multi-Candidate Evaluation Currently, no standard pairwise metrics (Papineni et al., 2002; Agarwal and Lavie, 2008; Lin, 2004; Vedantam et al., 2015; Zhang et al., 2020d) support a comparison between multiple candidates and multiple references, and the most efficient extension of existing metrics to multi-candidate, multi-reference situations is a non-trivial task. In this work, we naively extend the existing pairwise metrics to multiple candidates through the use of mean aggregation. Thus, for a standard pairwise score S, set of candidates $(c_1, \ldots, c_n) = C$ and a set of references

Figure 11.2: Intuition for TRMs. For samples from different distributions (left), in-distribution edges will often be short, but for identical distributions (right), edge rank-distributions will be more uniform.

$(r_1, \ldots, r_m) = R$, we assign the output score S_{agg} as:

$$S_{\text{agg}} = \frac{1}{N} \sum_{i=1}^{N} S(c_i, R) \tag{11.1}$$

11.2.2 Triangle-Rank Metrics (TRMs)

While existing metrics for semantic similarity are powerful for determining the pairwise semantic distances between two utterances (Papineni et al., 2002; Agarwal and Lavie, 2008; Lin, 2004; Vedantam et al., 2015; Anderson et al., 2016), these measures cannot accurately measure the distance between distributions. How, then, can we leverage already strong pairwise tools in a multiple candidate scenario? Unfortunately, many statistical techniques for measuring the distances between samples require points to lie in a metric space (Basseville, 2013) - however, most text distances neither respect symmetry nor triangle inequality.

We propose a novel answer based on an application of the triangle-rank statistic for statistical testing proposed by Liu and Modarres (2011). The triangle-rank statistic has several promising properties: it neither requires symmetry nor the triangle inequality in the metric space (it only requires $d(x, x) = 0$), and it is computed using only pairwise distances, meaning that we can easily reuse existing text semantic distance functions when computing the statistic.

For the purpose of explanation, it can be helpful to think of texts as points on an arbitrary manifold (based on the selected text distance function). To compute the triangle-rank statistic for a given distance S, a set of candidates $(c_1, \ldots, c_n) = C$ and a set of references $(r_1, \ldots, r_m) = R$, we first extract all directed triangles $(t_1, \ldots) = T$, such that one point lies in C and two points lie in R. We refer to the edge between points from the same distribution as $e_{t_i}^{\text{IN}}$ and the other two edges as $e_{t_i}^{E_0}$ and $e_{t_i}^{E_1}$. We then compute the score for each of the edges. For $(a, b) = e_{t_i}^{\cdots}$, let

$$d(e_{t_i}^{\cdots}) = S(a, b) \tag{11.2}$$

We then compute indicators I_0, I_1, I_2 for each triangle t_i as follows:

$$\begin{aligned}
I_0(t_i) =& 1 \text{ if } d(e_{t_i}^{\text{IN}}) \leq d(e_{t_i}^{E_0}), d(e_{t_i}^{E_1}) \text{ else } 0 \\
I_1(t_i) =& 1 \text{ if } d(e_{t_i}^{E_0}) \leq d(e_{t_i}^{\text{IN}}) \leq d(e_{t_i}^{E_1}) \text{ or } d(e_{t_i}^{E_1}) \leq d(e_{t_i}^{\text{IN}}) \leq d(e_{t_i}^{E_0}) \text{ else } 0 \\
I_2(t_i) =& 1 \text{ if } d(e_{t_i}^{E_0}), d(e_{t_i}^{E_1}) \leq d(e_{t_i}^{\text{IN}}) \text{ else } 0
\end{aligned} \tag{11.3}$$

These indicators represent the rank of the same-sample edge (if it is the smallest, largest, or middle-sized edge). The directed statistic for the sample (C, R), $Q(C, R)$ is then computed as:

$$Q(C,R) = \left|\frac{\sum_{t_i \in T} I_0(t_i)}{|T|} - \frac{1}{3}\right| + \left|\frac{\sum_{t_i \in T} I_1(t_i)}{|T|} - \frac{1}{3}\right| + \left|\frac{\sum_{t_i \in T} I_2(t_i)}{|T|} - \frac{1}{3}\right| \tag{11.4}$$

For the experiments in this paper, we use an extension of the directed statistic, the undirected statistic, $TRM(C, R) = Q(C, R) + Q(R, C)$, which increases the sensitivity of the metric by taking into account rank statistics of both within-candidate and within-reference edges.

An intuition for how this statistic measures divergence between distributions is given in Figure 11.2. If the in-distribution edges are always short compared to the cross-distribution edges, this suggests that either the distance between the candidate and reference distributions is high (different locus), or the spread of the candidates in the semantic space is significantly less than that of the references (different spread). If the in-distribution edge is always the longest edge, it suggests that the spread or dispersion of the candidate samples is higher than the dispersion of the reference samples. Because this statistic takes into account the full distribution through triplets of samples, it does not suffer from the issues with aggregation discussed earlier. Not only does it solve these issues, but TRMs build on existing pairwise metrics, allowing us to increase sensitivity while retaining existing semantic distance measure and intuitions.

Notably, $Q(C, R)$ does not distinguish between situations where $I_0 = 1$ and $I_2 = 1$. Intuitively, a model that can generate a candidate that is closer to two references than the references are to each other ($I_0 = 1$) seems be better than another model where the candidate is far apart from one (or both) of the references ($I_2 = 1$), however this is not always a desirable situation (in fact, it is often a situation we wish to avoid). Consider the situation where the “mean" of all reference captions is generated by the candidate set. This caption is closer to any individual caption than any reference caption may be to other reference captions, however as seen in Figure 11.1, and discussed in prior work (Caglayan et al., 2020; Yeh et al., 2021; Chan et al., 2022c), such captions capture only mutual information in the references, and fail to match the full distribution.

11.2.3 Kernel-Based Metrics

While TRMs represent one method of augmenting existing pairwise metrics, a second possible approach relies on representing utterances as points in the embedding space of a model, particularly a large pre-trained model such as BERT (Devlin et al., 2019) or GPT (Brown et al., 2020). Evaluating the distance between two distributions based on representative samples on a Euclidean manifold is relatively well studied in GAN literature. One option, MAUVE, introduced by Pillutla et al. (2021), uses a K-Means density estimator to estimate the distribution of the points on this manifold and then computes a fixed divergence (such as Kullbeck-Libeller) between the two density estimates. Unfortuantely, MAUVE cannot

correctly estimate the density when there are few samples, such as in the case of conditional language generation, as the K-means density estimator requires at least K (usually at least 50) samples. In this work, we introduce several possible extensions to MAUVE as an alternative family of distribution-aware metrics, which we dub “Kernel-Based Metrics" (KBMs):

- **FID-BERT (C.1.6):** The Frechet Inception Distance (Salimans et al., 2016) represents the squared Wasserstein distance between multidimensional Gaussian distributions fitted to the components of the input. In the FID-BERT metric, we replace Inception embeddings with those from a pre-trained BERT model (Devlin et al., 2019).
- **MMD-BERT (C.1.7)**: A related metric is the maximum mean discrepancy distance function (Li et al., 2017), which leverages a density estimate of the data, and computes the maximum mean discrepancy between the density estimates for each sample. In our case, we leverage a Gaussian kernel estimate over the embeddings generated by a pre-trained BERT model (Devlin et al., 2019).

While we primarily explore BERT-based embeddings for KBMs, we explore additional text embedding methods in Appendix C.2.1.

11.3 Case Study: Visual Description

Visual description is a challenging task where a model must generate natural language descriptions of visual scenes. Datasets for visual description often set themselves apart from other datasets for conditional natural language generation (such as those for translation and summarization), as they contain more than one ground truth sample, making it possible to evaluate multi-reference measures. In this set of experiments, we look at two datasets for visual description: MSCOCO (image description) (Lin et al., 2014) and MSR-VTT (Xu et al., 2016) (video-description) (full dataset details in appendix C.1.2). We demonstrate first that current metrics are not sensitive enough to evaluate the performance of existing approaches, and then show quantitatively how a multi-candidate evaluation paradigm can close this gap, and how a distributionally sensitive metric, such as TRMs, can provide new insights.

Single caption evaluation is insufficient A natural first question to ask when evaluating the performance of a metric is, “given the data, is the metric sensitive enough to distinguish between captions from a model and caption from a reference distribution?" To answer this question, we evaluated the p-values using a permutation-test for each measure under the null hypothesis that the candidate and reference samples come from the same caption distribution. The p-values represent the probability of obtaining the observed result under the null hypothesis: a higher p-value means that it is immanently possible the results obtained are due to chance rather than any signal in the underlying experiment. It is important to highlight that in this paper, when we compare p-values, we are evaluating the *sensitivity* of the measures on a *single experiment* and *not comparing p-values between experiments*. It

Table 11.1: The p-value (lower is better) produced by measuring standard metrics under the null hypothesis that the candidate distribution is the same as the reference distribution (using single-image/video tests aggregated with HMP (Wilson, 2019)). With a single candidate text, the metrics are unable to make a statistically significant distinction ($p < 0.05$) between ground truth and candidate samples, motivating the need for multi-candidate evaluation. BERT refers to the BERT-Score (Zhang et al., 2020d). Additional experimental detail in C.1.5.

Model	**BERT**	**CIDEr**	**BLEU**	**METEOR**	**ROUGE**
(Video) MSR-VTT Test Set p-values					
TVT	0.658	0.409	0.781	0.457	0.477
O2NA	0.645	0.457	0.795	0.564	0.593
Human	0.515	0.531	0.829	0.530	0.566
(Images) MS-COCO Karpathy Test Set p-values					
CLIPCap	0.558	0.822	0.878	0.748	0.798
VLP	0.592	0.742	0.859	0.664	0.770
Human	0.640	0.668	0.874	0.635	0.684

is generally not the case that lower p-values correspond to better captions, rather, lower p-values when comparing two differing distributions indicate a more sensitive measure.

The results, shown in Table 11.1 demonstrate that under all existing measures, using a single description for the candidate dataset does not have sufficient sensitivity ($p < 0.05$) to tell different distributions apart, motivating a transition to a paradigm with significantly more sensitivity. This result confirms observations made in Yeh et al. (2021) and Liu et al. (2016): most metrics are unable to produce statistically significant results. Thus, even for standard metrics, it makes sense to sample more than one ideal candidate description and aggregate the metric score across these candidate descriptions. Such a sampling approach for evaluation does not preclude efforts toward generating single "omnibus" captions capturing details from several diverse captions. However, such captions will be much longer than typical human captions, and will score poorly under the standard metrics, as they would differ greatly from individual reference captions.

TRM and KBM metrics are more sensitive than naive aggregation In section 11.2, we proposed several new metrics which can be leveraged by switching to multi-candidate evaluation. Figure 11.3 shows the sensitivity of both the newly introduced metrics and existing metrics using the naive aggregation schemes discussed in section 11.2, as we increase the

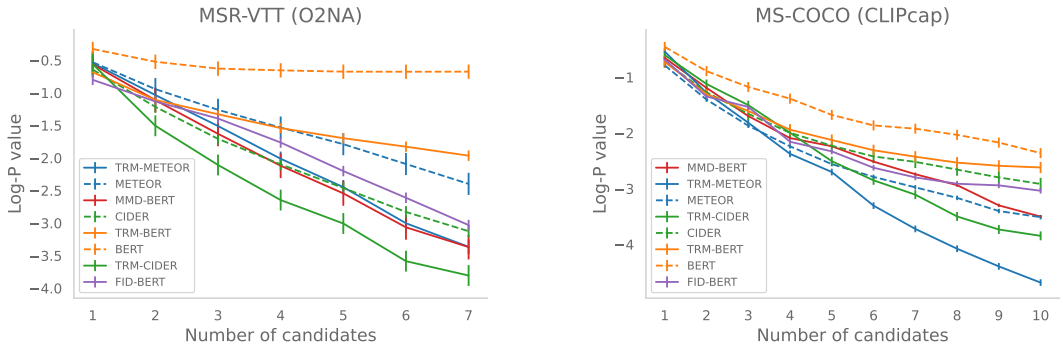


Figure 11.3: Plots showing the log p-values for the existing and proposed metrics as we increase the number of sampled candidate descriptions from the models. $\text{TRM}_{\text{METEOR}}$ achieves a 162% increase in sensitivity over METEOR, while $\text{TRM}_{\text{CIDEr}}$ represents a 49.3% increase over CIDEr-D for O2NA evaluated on the MSR-VTT dataset. Additional experimental details are given in C.1.5.

number of candidate samples from the model. While the sensitivity increases for all models to significance, our proposed metrics are much more sensitive with fewer candidate and reference descriptions. As an additional check, when tested on human captions, our metrics do not consider the two distributions significantly different ($p > 0.05$, see C.2.3). Our proposed metrics do not alter the manifold: so, for example, $\text{TRM}_{\text{METEOR}}$ and METEOR measure the same underlying intuitive divergences (n-gram recall with some additional synonym matching), however, our TRM method increases the sensitivity of the test, allowing us to measure the full distribution divergence, instead of using naive aggregates. For a practitioner, computing the full p-value of the data is unnecessary; we need only sample enough candidates to be sure of the statistical significance.

Multi-candidate evaluation illustrates a diversity vs. likelihood trade-off A metric's sensitivity to the full distribution can give us novel insights into the visual description task. Consider the two models, VLP (Zhou et al., 2020), a standard transformer-based model pre-trained on large-scale vision and language data, and CLIPCap (Mokady et al., 2021), a transformer-based model which is initialized with a large language model, and uses prefix-tuning with CLIP (Radford et al., 2021c) embeddings (Additional details in C.1.3). Figure 11.5 illustrates that $\text{TRM}_{\text{METEOR}}$ captures a subtlety in the model comparisons that METEOR does not capture alone: while VLP produces better descriptions at low temperatures, it becomes less fluent (likelihood) on average as we introduce diversity, leading to worse captions when sampling at high diversities. CLIPcap retains better fluency at high sampling temperatures, leading to improved performance in diverse captioning tasks. While $\text{TRM}_{\text{METEOR}}$ demonstrates this, METEOR monotonically decreases, giving little insight into this problem. The sensitivity of the TRM measure is also visible in qualitative samples, given

Candidate Set 1
METEOR (↑): 1.0
TRM-METEOR (↓): 0.574

The cows are grazing in a field.
The cows are grazing in a field.
The cows are grazing in a field.
The cows are grazing in a field.
...

Candidate Set 2
METEOR (↑): 0.393
TRM-METEOR (↓): 0.069

Animals grazing on grass in an enclosed area.
Several cows grazing in a field with trees in the background.
Cows grazing in a large green pasture in a distant scene.
A grassy field overlooking cows in a pasture.
...

References

Cows grazing in a pasture ringed with trees.
Polaroid-looking photograph of cows in a green pasture.
A herd of animals grazing on a lush green field.
The cows are grazing in a field.

Figure 11.4: A qualitative sample from CLIPcap. Candidate set one uses beam search (8 beams), while candidate set two uses nucleus sampling (with temperature one, top-k of 20 and top-p of 0.9). As the diversity increases, the $\text{TRM}_{\text{METEOR}}$ divergence decreases, but METEOR fails to correctly capture the diversity/correctness trade-off, leading to decreased scores for more complete caption sets that are still relatively high quality. Additional qualitative examples are provided in C.2.6.

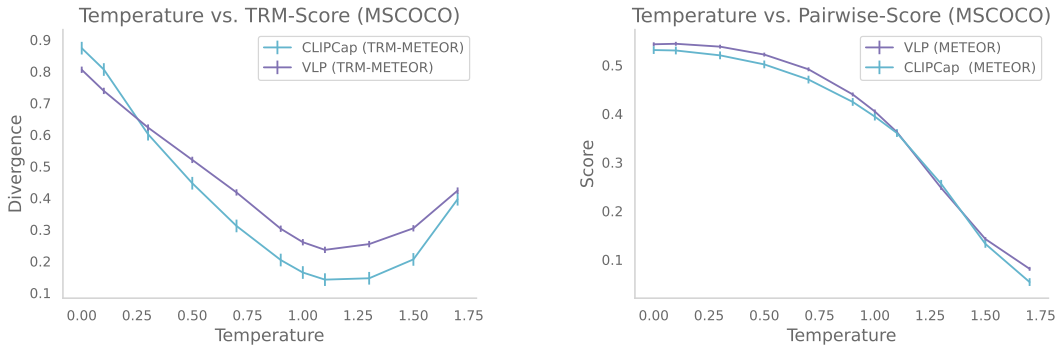


Figure 11.5: Plots indicating the impact of temperature on the metric scores. Left: $\text{TRM}_{\text{METEOR}}$ (↓) for CLIPcap and VLP. Right: Standard METEOR Score (↑) for CLIPcap and VLP.

in Figure 11.4, where we see TRM metrics are sensitive to both diversity and likelihood. These results confirm observations made by Zhang et al. (2021a) for open-ended language generation tasks such as storytelling and dialogue: a fair comparison of approaches must not only compare at the same level of entropy but at a range of entropy levels.

Sampling algorithms matter Not only does the temperature of the generation process matter when correctly trading off between diversity and description correctness (as seen in the previous discussion), but the sampling process itself matters. Figure 11.6 shows the performance at different temperatures of the Nucleus sampling method (Holtzman et al., 2020)

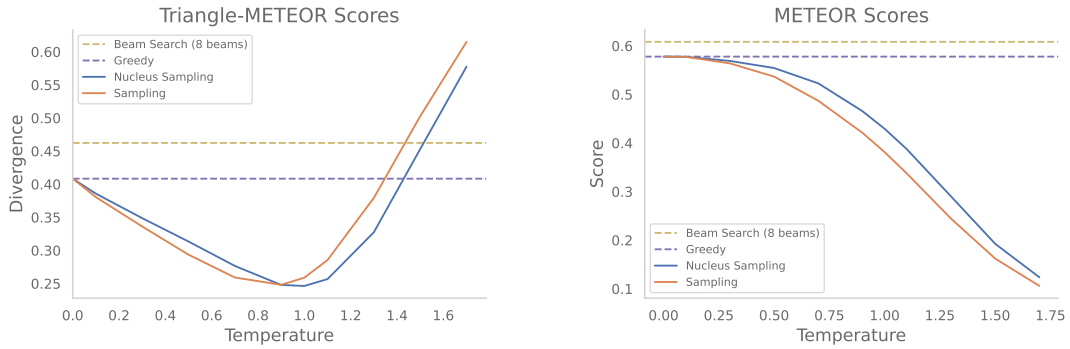


Figure 11.6: Plots indicating the impact of search technique on divergences. Left: $\text{TRM}_{\text{METEOR}}$ (↓) for TVT on MSR-VTT. Right: METEOR Score (↑). See C.1.8 for experimental details.

Table 11.2: Method evaluation efficiency on the MS-COCO dataset with 5 references and 10 candidates.

	METEOR	$\text{TRM}_{\text{METEOR}}$	CIDEr	$\text{TRM}_{\text{CIDEr}}$	MMD-BERT	FID-BERT	MAUVE
Samples/Sec	298.4 ± 18.3	161.18 ± 21.2	131.23 ± 12.6	97.54 ± 9.1	53.76 ± 38.7	17.45 ± 4.6	2.29 ± 0.78
Wall Time (Min)	2.26	4.18	5.14	6.92	12.55	38.68	294.78

vs. standard sampling, beam search, and greedy, approaches. While maximum-likelihood methods achieve the best METEOR scores, they have relatively high divergence, as they sample only a single description. Further, Figure 11.6 shows that $\text{TRM}_{\text{METEOR}}$ illustrates how Nucleus sampling allows models to achieve higher temperatures than standard sampling without diverging significantly from the distribution. METEOR alone does not indicate such an effect and only monotonically decreases.

TRM Measures correlate with human judgements It has long been known that humans are relatively poor at measuring the semantic distance between two sets of objects, particularly in the presence of distractors (Durga, 1980). While this is the case, we still find that proposed measures correlate with human judgement significantly more than existing measures, which we show in Table 11.3. To demonstrate the correlation of distributional measures with human judgement of distributional distance, humans were presented with two candidate caption sets (two image captioning models, OFA (Wang et al., 2022b) and BLIP (Li et al., 2022) using different temperatures), and asked which candidate caption set correlated better with a reference caption set on two measures: how much they overlapped factually (correctness), and how much information they provided about the references (coverage). Additional experimental details are available in C.1.9.

Clearly, distributional measures correlate more, and with significantly less information than existing measures aggregated using the max function. Notably, despite evidence that

Table 11.3: Pearson Correlation with human judgement, $N = 794$.

Method	**Coverage**	**Correctness**
Human	0.2247 ($p < 0.001$)	0.2247 ($p < 0.001$)
TRM-Meteor	0.1278 ($p < 0.001$)	0.1082 ($p < 0.001$)
TRM-BLEU	0.1271 ($p < 0.001$)	0.1510 ($p < 0.001$)
MMD-BERT	0.1288 ($p < 0.001$)	0.1243 ($p < 0.001$)
FID-BERT	0.0807 ($p = 0.011$)	0.0978 ($p < 0.001$)
METEOR	0.0162 ($p = 0.3978$)	0.0057 ($p = 0.7650$)
BLEU-4	0.0044 ($p = 0.8157$)	0.0026 ($p = 0.8884$)
ROUGE	0.0110 ($p = 0.5631$)	0.0381 ($p = 0.1845$)
CIDEr	0.0037 ($p = 0.8445$)	0.0261 ($p = 0.1725$)

existing decoding methods optimize for fooling humans over correctness (Ippolito et al., 2020), our method *is the only approach* which correlates *at all* with human judgement, suggesting that we have accomplished our goals of being distribution aware, improving the sensitivity of the base measures to human preferences.

11.4 Discussion and Limitations

Kernel-Based Metrics (KBMs) vs. Triangle-Rank Metrics (TRMs) A natural question to ask is: "which metric should practitioners choose when evaluating conditional language models?" KBMs have one major, distinct, advantage over the TRMs in that they are naturally differentiable, yet KBMs also have downsides. The first is that, unlike the TRMs, they require both a pre-trained BERT model and a kernel-density estimator which both have complex behavior affecting the performance of the model. The TRMs, however, can be specified on top of existing natural language distance functions, improving the ability of the user to intuit the model performance. Additionally, TRMs are bounded and have p-values that can be computed analytically. Finally, because the TRMs do not need a density estimate, they can be more sensitive with small sample sizes (see Figure 11.3), which is essential for conditional language generation where we have only a few gold-standard samples. Table 11.2 demonstrates another key benefit of TRMs: efficiency. The time per sample to compute TRMs, while higher than single metric standards, is lower than KBMs on average.

Perplexity We acknowledge that perplexity (likelihood of the test distribution) is another alternative metric to proposed methods. While methods **should** report the perplexity of their models, it is not standard practice, and it has been shown by Theis et al. (2016) that perplexity suffers from several major issues when evaluating generative models. For example,

a lookup table storing sufficiently many training examples will produce convincing results but have poor perplexity on the test data. On the other hand, van den Oord and Dambre (2015) demonstrate that even when perplexity is low, models may not generate high-quality test samples.

Reference-Free Metrics Some metrics, such as CLIP-score (Hessel et al., 2021b) for visual description, are immune to ground truth aggregation effects as they are computed in a reference-free way, and focus on pre-trained models' ability to ground vision and language information. Unfortunately, such large, black-box, models represent a liability as a metric as their capabilities are largely unknown, and untested (Floridi and Chiriatti, 2020; Caglayan et al., 2020). Further, the metric is only as good as the model, and CLIP has been known to suffer from numerous issues including counting, attribute-association, and spatial reasoning (Blattmann et al., 2022; Ramesh et al., 2022).

11.5 Conclusion

In this chapter, we introduce a robust framework for multi-candidate evaluation of conditional language generation models, show that existing metrics for semantic similarity can be seamlessly extended to this framework, and demonstrate that multi-candidate evaluation paired with more sensitive distribution-aware metrics can provide novel insights into existing models and methods. This work is only the beginning. It is necessary for future work to explore how a wider range of existing generation techniques and models perform under this new paradigm, and to understand the implications of distribution-aware evaluation in fields beyond visual description.

Chapter 12

CLAIR: Evaluating Image Captions with Large Language Models

As we have already seen so far in this thesis, automatically evaluating the quality of image captions is challenging. There are many dimensions to consider, such as grammatical quality, semantic relevance, correctness, and specificity, among others. To ensure fair evaluations, most image captioning works employ a suite of measures, each capturing different aspects. For instance, n-gram-based measures like BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2015) broadly measure content overlap, SPICE (Anderson et al., 2016) compares scene graph structures, and CLIPScore, TIFA, SeeTrue and VPEval (Hessel et al., 2021b; Hu et al., 2023; Yarom et al., 2023; Cho et al., 2023) directly incorporate visual information. Unfortunately, while significant strides have been made in automated evaluation, human preference studies remain the most reliable (yet costly) source of caption evaluation.

Fortunately, recent advances in large language models (LLMs) have opened new avenues for automatic evaluation. Models trained with reinforcement learning from human feedback (RLHF, Christiano et al. (2017)) or similar methods are particularly useful for open-ended evaluation tasks, including image captioning, due to their explicit training to align with human preferences.

In our work, paralleling several recent works which find that LLMs can act as effective "judges" for selecting the better answer from two candidates (Bubeck et al., 2023; Dettmers et al., 2023; Chiang et al., 2023), we explore the ability of LLMs to evaluate caption quality in the multimodal setting. We introduce CLAIR (Criterion using LAnguage models for Image caption Rating), a measure which scores a candidate caption based on the likelihood that it describes the same image as a set of references by directly asking an LLM to produce a numeric rating. We further query the LLM to provide a *reason* for its score, providing a level of interpretability to the scalar rating. As far as we are aware, this is the first work to explore replacing measures of *semantic text quality* with directly obtained LLM judgments, however concurrently, Zheng et al. (2023) have shown that directly providing an answer rating can align highly with human preferences on a range of standard language-based tasks, such as conversational instruction following.

```
You are trying to tell if a candidate set of captions is describing
the same image as a reference set of captions.
Candidate set:
{candidate captions}
Reference set:
{reference captions}
On a precise scale from 0 to 100, how likely is it that the candidate
set is describing the same image as the reference set? (JSON format,
with a key "score", value between 0 and 100, and a key "reason"
with a string value.)
```

Figure 12.1: CLAIR: a (surprisingly simple) large language model-based measure for image caption evaluation. We find that CLAIR not only correlates strongly with human judgments of caption quality but can also generate interpretable reasons for the generated scores.

Through several experiments on captioning datasets such as MS-COCO (Xu et al., 2016), Flickr8k (Mao et al., 2014), and PASCAL-50S (Vedantam et al., 2015), we find that CLAIR correlates surprisingly well with human preferences, outperforming prior captioning measures. We additionally propose CLAIR$_E$, where we Ensemble the outputs of several LLMs by taking the average score, leading to further improvements.

Despite a simple pipeline using an LLM prompt with minimal output parsing, CLAIR's strong correlation with human preferences suggests that it captures multiple dimensions of caption similarity at once – a feature that prior measures struggle to achieve alone. More generally, CLAIR demonstrates how language-only models can evaluate vision-language tasks. We show LLMs can provide not only reliable scalar ratings but also corresponding reasoning for a given rating, offering a valuable combination of accuracy and interpretability.

12.1 CLAIR: LLMs for Caption Evaluation

In CLAIR, we adapt the zero-shot in-context learning approach described in Brown et al. (2020) to score candidate captions with large language models (LLMs). This involves converting the caption evaluation problem into a human-readable text completion task which is solved by the LLM. Using the prompt in Figure 12.1, CLAIR first generates completions from the LLM and then parses those completions into both candidate scores and an explainable reason for the score. We use a greedy sampling method ($t = 0$) to encourage reproducibility in the results, while acknowledging the inherent nondeterminism in LLMs (see section 12.3). CLAIR's experimental implementation is surprisingly simple: it uses no in-context examples (is entirely zero-shot), and default inference parameters for the APIs. See Appendix D.1 for further implementation details.

The choice of language model directly affects the quality of the CLAIR measure – more

accurate models should produce evaluations that align better with human judgment. We explore three language models: GPT-3.5 (ChatGPT) (OpenAI, 2022a), Claude (Instant) (Bai et al., 2022), and PaLM (Chowdhery et al., 2022). Unfortunately, we found for several open-weight language models including Koala (Geng et al., 2023) and Vicuna (Chiang et al., 2023) that CLAIR aligned poorly with human judgment.

As CLAIR is language model-agnostic, we can leverage the different distributions learned by each model and combine their decisions in an ensemble approach we term CLAIR$_E$. We calculate individual CLAIR scores for each model and compute an unweighted average to obtain the ensemble score.

Benchmark measures: We benchmark against several existing measure of caption similarity. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Agarwal and Lavie, 2008) and CIDEr (Vedantam et al., 2015) all primarily measure n-gram overlap (however have different weighting schemes between n-grams, and across precision/recall). We also compare against SPICE (Anderson et al., 2016), which compares caption parse trees and focuses on matching perceived action and object relationships in addition to n-grams. While the aforementioned measures are commonly reported in image captioning works, we also compare against several modern measures, including CLIP-Score (Hessel et al., 2021b) which uses the recent CLIP (Radford et al., 2021c) model for reference-free evaluation.

12.2 Evaluation & Discussion

To evaluate the quality of the measure, we run several evaluations that compare scores generated by CLAIR to both human judgments of caption quality and other image captioning evaluation measures. We additionally provide several qualitative examples in Figure 12.2. A unique benefit of CLAIR is that it provides not only numeric scores but is also introspectable, as it can identify which details in the candidate caption set match the reference set.

Sample-level human correlation: We first ask the question, how well does CLAIR correlate with human judgments of caption quality at a sample level? We do so by exploring the performance on three datasets, COMPOSITE, Flickr8K-Expert, and MS-COCO (See Appendix D.1 for details).

The results of our sample-level correlation experiments are shown in Table 12.1. We can see that CLAIR outperforms language-only measures (e.g., 0.604 to 0.403 for SPICE), and in most cases, outperforms vision-augmented measures. CLAIR$_E$ achieves strong sample-level correlation on all datasets; for instance, CLAIR$_E$ closes the gap to inter-human agreement by 0.097 over vision-based measures and 0.132 over language-based measures. The improvements of CLAIR$_E$ over CLAIR suggest that each language model may have some bias (similar to each human), yet the ensemble of models correlates more strongly with human judgments. A reasonable concern might be that the models underlying existing approaches are significantly smaller than those in CLAIR, and trained on less data. To address this, we introduce

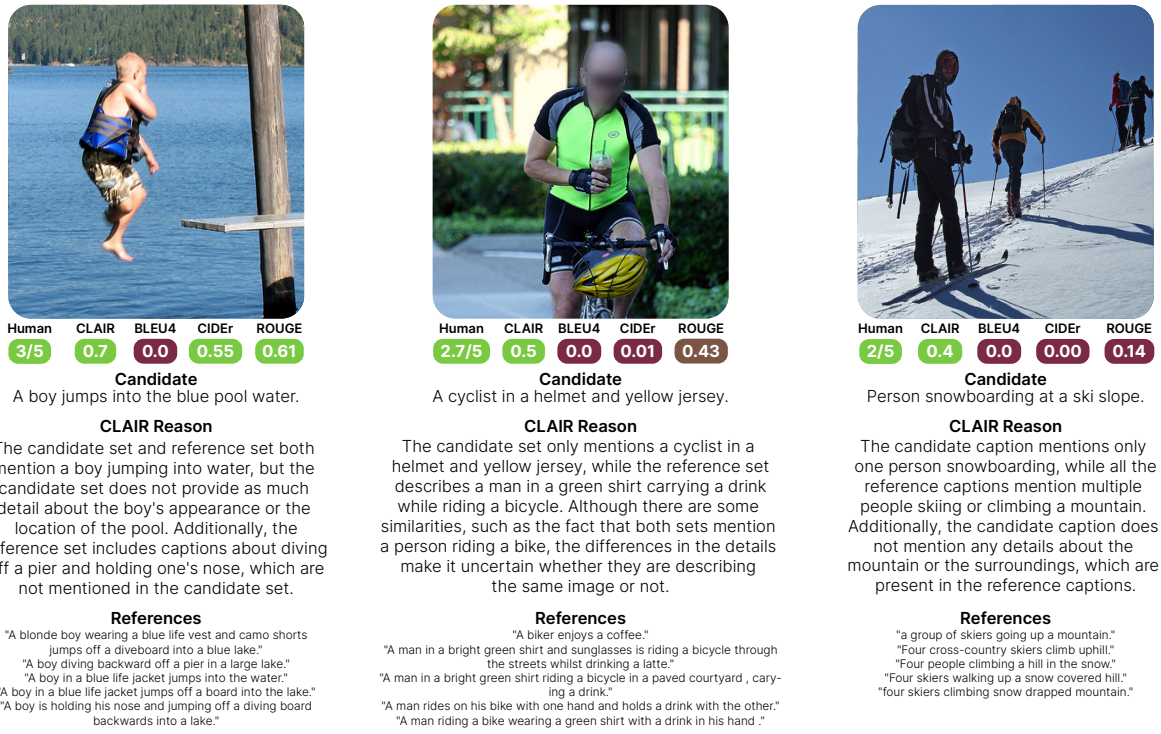


Figure 12.2: Several qualitative examples of CLAIR from the Flickr8K-Expert dataset. CLAIR not only correlates better with human judgments of caption quality but also provides detailed explanations for its score. CLAIR scores normalized by 100.

and compare against RefCLIP-X, which replaces the CLIP model in RefCLIP with a CLIP ViT-bigG/14 model trained on LAION 2B (Ilharco et al., 2021). Even in this case, CLAIR demonstrates significantly improved performance. Note that in Table 12.1, we use τ_b, instead of the form τ_c, due to the ties generated by the CLAIR model.

System-level human correlation: In addition to computing the sample-level correlation on the MS-COCO dataset, we use the annotations from the five models considered by Rohrbach et al. (2018) to compute the system-level correlation. For each of the methods, we compute the mean human score on the test samples, and mean metric score on the test samples, followed by the Kendall's rank correlation coefficient (Kendall's tau, strength of ordinal association) between these values (the set of five mean human scores, and the set of five metric scores). Our results, given in Table 12.2, demonstrate that CLAIR ranks the five methods in a novel way that is more accordant with human rankings of the methods. These results further suggest that CLAIR has the potential to redefine which methods are preferable to humans compared to existing n-gram approaches.

Table 12.1: Sample-level correlation (Kendall's τ_b) with human judgments. All p-values < 0.001. *: Model has access to additional visual context.

	Dataset		
Measure	COMPOSITE	Flickr8K	MS-COCO
BLEU@1	0.313	0.323	0.265
BLEU@4	0.306	0.308	0.215
ROUGE-L	0.324	0.323	0.221
BERT-S	0.301	0.392	0.163
METEOR	0.389	0.418	0.239
CIDEr	0.377	0.439	0.262
SPICE	0.403	0.449	0.257
CLIP-S*	0.498	0.511	-
RefCLIP-S*	0.512	0.526	-
RefCLIP-X*	0.523	0.549	0.274
CLAIR			
+ GPT3.5	**0.604**	0.616	0.296
+ Claude	0.542	0.563	0.320
+ PaLM	0.580	0.546	0.355
CLAIR_E	0.592	**0.627**	**0.374**
Inter-Human	-	0.736	-

Decision Making: In addition to evaluating the correlation with human judgments, we also evaluate the capability of the measure to perform discriminative analysis. The PASCAL-50S dataset (Vedantam et al., 2015) contains a set of 4000 human-annotated caption pairs. For each pair of captions, humans label which caption in the pair is closest to the reference set for the image. The caption pairs fall into four groups: "HC:" two human-written captions matching the image, "HI:" one human caption, and one machine-generated caption, with only one matching the image, "HM:" a matching human caption and a matching machine-generated caption and "MM:" two matching machine-generated captions. See Appendix D.1 for more dataset information.

The performance on PASCAL-50S is given in Table 12.3. We can see that CLAIR_E outperforms all existing text-only measures (e.g., by 5.18% overall score over CIDEr), and in many cases, even outperforms measures that have access to the image at test time. Note that it is relatively weaker than image-augmented models in the HC setting; however, since both captions are correct, the model often cannot judge which is better purely the text. Models such as RefCLIP-S that have access to the image are naturally better discriminators in this

Table 12.2: System-level correlation between the average CLAIR score and human model evaluation for 5 models trained and evaluated on MS-COCO. All p-values < 0.05.

Measure	Kendall's τ_b	Spearman's ρ	Pearson r
BLEU@1	0.399	0.600	0.706
BLEU@4	0.799	0.899	0.910
ROUGE-L	0.600	0.700	0.792
METEOR	0.600	0.700	0.666
CIDEr	0.399	0.600	0.856
SPICE	0.399	0.600	0.690
CLAIR			
+ GPT3.5	0.799	0.899	0.869
+ Claude	**1.000**	**1.000**	0.868
+ PaLM	**1.000**	**1.000**	**0.954**
CLAIR_E	**1.000**	**1.000**	0.903

case. We suspect that CLAIR's discriminative performance could be further improved by giving the LLM a choice between the two captions; however, we leave this optimization to future work.

Groups of Captions: While CLAIR is capable of comparing a single candidate caption to a set of reference captions, it is also capable of comparing *sets* of candidate captions to sets of reference captions. This task is necessary when evaluating the ability of a model to generate captions that are diverse and that fully describe the conditional text distribution. We evaluate on the COCO-Sets dataset (Chan et al., 2022d), 794 caption sets rated by AMT workers on two scales: how closely a candidate set matches the reference set in terms of both correctness and content coverage (See Appendix D.1 for details). The results of this experiment are given in Table 12.4. We can see that CLAIR outperforms well when measuring the quality of a group of captions, and approaches the inter-human correlation on the (very) challenging task. CLAIR also outperforms TRM-METEOR and TRM-BLEU (Chan et al., 2022d), suggesting that LLMs can judge both the content and diversity of the caption sets.

12.3 Limitations

While CLAIR correlates well with human judgments of caption quality, it has several limitations:

Table 12.3: Accuracy of measures when matching human decisions for PASCAL-50S (5 reference captions). *: Model has access to additional visual context.

Measure	HC	HI	HM	MM	All
BLEU@1	51.20	95.70	91.20	58.20	74.08
BLEU@4	53.00	92.40	86.70	59.40	72.88
ROUGE-L	51.50	94.50	92.50	57.70	74.05
METEOR	56.70	97.60	94.20	63.40	77.98
CIDEr	53.00	98.00	91.50	64.50	76.75
SPICE	52.60	93.90	83.60	48.10	69.55
TIGEr*	56.00	99.80	92.80	74.20	80.70
CLIP-S*	56.50	99.30	96.40	70.40	80.70
RefCLIP-S*	64.50	99.60	95.40	72.80	83.10
CLAIR					
+ GPT3.5	52.40	99.50	89.80	73.00	78.67
+ Claude	**57.90**	98.50	91.30	62.90	77.65
+ PaLM	54.70	98.30	87.30	64.00	76.08
CLAIR_E	57.70	**99.80**	**94.60**	**75.60**	**81.93**

Non-Determinism and Parsing Errors: Because CLAIR depends on the output of a language model, the measure can be non-deterministic and noisy. For instance, it may fail to elicit a judgment (e.g., "As an AI language model, I cannot see, and thus, cannot determine if the image captions match the references"), or rarely, generate malformed JSON output. To address these issues, we perform multiple queries to the LLM, sometimes at higher temperatures if necessary. As a consequence, the measure may differ between runs, although we found the variance to be relatively insignificant (< 0.01 in many of the experiments). Additionally, since the language models used are not open-source, the models are subject to arbitrary change, replacement, or removal, which limits the efficacy of the measure as a long-term comparable measurement. We hope that increasing open access to language models with efforts such as Koala (Geng et al., 2023) and Vicuna (Chiang et al., 2023), will help to alleviate these challenges in the future.

Increased Cost: CLAIR relies on language models which contain many billions of parameters. These language models have not only monetary cost but also human and environmental costs (Bender et al., 2021) which can reduce its utility as a target during training, such as for self-critical sequence training (Rennie et al., 2017). While API-based LLMs may be considered costly, even open-source LLMs have a cost (which can often be hard to quantify). CLAIR on the MS-COCO dataset uses an average of 226.148 tokens per sample (on OpenAI's

Table 12.4: Pearson correlation with human judgments when evaluating sets of captions on MS-COCO ($N = 794$).

Measure	$\text{Coverage}_{\text{p-value}}$	$\text{Correctness}_{\text{p-value}}$
BLEU@4	$0.004_{0.816}$	$0.003_{0.888}$
ROUGE-L	$0.011_{0.563}$	$0.038_{0.184}$
METEOR	$0.016_{0.398}$	$0.006_{0.765}$
CIDEr	$0.004_{0.844}$	$0.026_{0.173}$
TRM-METEOR	$0.128_{<0.001}$	$0.108_{<0.001}$
TRM-BLEU	$0.127_{<0.001}$	$0.151_{<0.001}$
MMD-BERT	$0.129_{<0.001}$	$0.124_{<0.001}$
FID-BERT	$0.081_{0.011}$	$0.098_{<0.001}$
CLAIR		
+ GPT3.5	$\mathbf{0.195}_{0.011}$	$\mathbf{0.187}_{0.014}$
+ Claude	$0.110_{0.099}$	$0.124_{0.145}$
+ PaLM	$0.129_{0.081}$	$0.085_{0.172}$
CLAIR_E	$0.183_{0.027}$	$0.156_{0.018}$
Inter-Human	$0.225_{<0.001}$	$0.274_{<0.001}$

API), representing a cost of $0.0067 per sample (GPT-4), or $0.00033 per sample (GPT 3.5). For PALM, this drops to $0.000113 per sample. We hope that over time, advances in LLM inference (such as quantization and distillation), coupled with improvements in architecture will continue to yield lower-cost alternatives with strong performance on the caption evaluation task.

Hallucination: While CLAIR does suffer from potential hallucination, we strongly believe that this weakness does not diminish the fact that CLAIR still correlates strongly with human judgment. In CLAIR, hallucinations in the score manifest as "incorrect" judgements of similarity, while hallucinations in the explanations manifest as poorly grounded explanations of the score/quality. Hallucinations in the score should be considered false negatives (blind spots instead of hallucinations). In the case of hallucinations in the explanations, such hallucinations may lead to misinterpretation, but arguably less misinterpretation than a black box method, and may even indicate misunderstandings in the model. Hallucination is a well-known challenge of current LLMs and is the subject of a great amount of research on instruction-tuning, RLHF, RLAIF, and other methods. As hallucination and instruction-following performance of the base models improves, CLAIR inherit similar improvements.

Explainability: While CLAIR generates explanations for each rating, CLAIR has no strict scoring rubric. Much like human judgments, there is no direct way of attributing changes in score to changes in caption quality. For similar reasons, it is difficult to evaluate the quality of the generated explanations. Qualitatively, the explanations are generally reasonable and consider multiple axes of judgment.

12.4 Conclusion

This work introduces CLAIR, an LLM-based evaluation measure for image captioning. CLAIR's superior performance compared to highly-engineered measures indicates a remarkable fact: LLMs are well aligned with human judgments of caption quality, even more so than some measures designed specifically for semantic similarity. CLAIR is only a glimpse into how LLMs can be used for evaluation tasks, and image captioning is only the beginning. We hope that our work will inspire further exploration of similar measures in other vision and language domains, such as visual storytelling (Huang et al., 2016b), where human evaluation of generated text remains a challenging task.

Chapter 13

Discussion

As we build more contextual models, we need to continue to develop evaluation measures that are capable of closely matching human judgments and measurements of quality. In this section, we explored two aspects of this evaluation framework. We first introduced triangle rank measures, focusing on a new paradigm of evaluating not only the single best-generated sample but the diversity of multiple generated candidates simultaneously. Such an approach allows us to understand not only when models are correct, but when they closely match a human distribution of generated text. Going beyond using explicit measures to perform such matching, we further introduced CLAIR, which replaces a fixed distance metric between texts with an LLM-based distance measure that implicitly captures semantic distance. Doing so, we found that LLMs correlate strongly with human judgments of performance, enabling a wide range of possible directions for automated evaluation in underspecified tasks.

The measures introduced in this section are only the beginning of a new era of natural language evaluation tools based on statistical methods and distribution alignment, rather than direct n-gram comparison. There are many possible directions for further research in this space:

1. **Development of More Granular Meta-metrics:** Building upon the foundation laid by *TRMs*, future research could focus on developing more granular meta-metrics that can assess specific aspects of diversity and quality in the generated text. Currently, TRMs are based on n-gram or global text similarity measures, but this need not be the case, as the distance measures can be anything (within reason). Extending the distance measure could help disentangle different dimensions of performance, such as novelty, relevance, and coherence, providing a more nuanced understanding of model capabilities.

2. **Leveraging Multimodal Large Language Models:** With the advancement of multimodal large language models such as GPT4-V and LLaVA (Liu et al., 2024b), there's a significant opportunity to explore how these models can be utilized in evaluating CNLG, especially in contexts where understanding requires integrating information across text, image, and possibly audio or video modalities. Measures such as CLAIR

may be more effective when paired with a large multimodal model, rather than a language model alone (though doing so raises some questions about correlation, and what using a large multimodal model for evaluation means for out-of-distribution inputs).

3. **Bias and Fairness in CNLG Evaluation:** Investigating the potential biases in current evaluation methods remains a challenging problem. It is important to explore how both TRMs and CLAIR introduce novel evaluation biases and to correct and mitigate these biases. Such directions are not only limited to the analysis of existing measures, or the creation of new measures but could include tasks such as creating more diverse datasets for evaluation or creating evaluation protocols that involve diverse groups of human judges.

4. **Interpretability and Explainability in Evaluation:** CLAIR is one of the few "explainable" measures for NLG evaluation, in that it naturally responds as to why it generated a particular score. That being said, the reasons generated by CLAIR are unverified, and it remains to be seen if (and if so, in what ways) such reasoning or explanations correlate with human judgment. This is a challenging task, however uncovering and understanding explanations, and determining if such explanations are grounded is a question that I would desperately like answered.

5. **Cross-Lingual and Cross-Cultural Evaluation:** Expanding evaluation methods to be more inclusive of different languages and cultures, addressing the challenge of evaluating CNLG models in a globally diverse context. This involves developing evaluation metrics that can operate across languages and cultural contexts, possibly leveraging multilingual large language models and culturally diverse datasets.

Exploring these directions for future work promises to significantly enhance the evaluation, fairness, and effectiveness of conditional natural language generation models, ensuring they are more closely aligned with the diverse and nuanced needs of users worldwide. By tackling these challenges, we can pave the way for advancements that not only push the boundaries of AI capabilities but also ensure these technologies serve society in more ethical, inclusive, and meaningful ways.

Part IV

Discussion & Conclusion

Chapter 14

Discussion and Future Research Opportunities

In this dissertation, we comprehensively investigate ways in which we can understand the data distributions of our multi-modal problems, build models accounting for context, both intrinsic and extrinsic, and evaluate the resulting contextually aware models in a human-centric way. Returning to the key question of the dissertation, *How can we understand, build, and evaluate models for contextual natural language generation?* We can see that indeed, it is clear that the language that we generate, how we generate that language, and how we evaluate that generated language is all a product of the context within which that language was created. While we have, in this dissertation, delved deeply into all three questions, there remains considerable work required to understand how context fits into the language generation picture, particularly in our application domains, and how we can build and evaluate models that are aware of this context. In this section, I want to discuss several key directions for future research overall, beyond those that we have already discussed in chapter 5, chapter 10 and chapter 13.

14.1 Exploring New Applications

To start with the most direct area for future research, while this dissertation was limited to two application domains (visual description and ASR), I believe that the lessons learned are widely applicable across a wide range of conditional natural language generation tasks. One of the most interesting application tasks would be in abstractive text summarization, which suffers from many of the same ambiguity issues that image captioning does. Applying the metrics from chapter 11 to the summarization task was effective in some of our preliminary experiments (which never made it into either the paper or this dissertation). Beyond summarization, several potential applications stand out as interesting for future work including digital storytelling, interactive media (such as games), and image/video generation.

14.2 Finding New Sources of Context

Going beyond new applications, in part II of this dissertation, we discussed including context from several key sources including co-occurring media (such as video and audio in chapter 7), text catalogs (in chapter 8), conversations (in chapter 9), and user identity in chapter 6). One of the key areas for future work is to uncover and utilize new areas of context that can be applied to both improve our models and to better align models with human preferences and goals.

One interesting possible area of context to explore in future work is socio-cultural sources of context: the social and cultural elements that influence a person's behavior, beliefs, experiences, and interactions within their community. Elements such as historical references, cultural practices, and societal norms can all influence how systems interpret the world. For ASR, this might be instantiated as being able to recognize specific cultural references, names of festivals, rituals, or historical events, while in visual description, it involves (but is not limited to) understanding the cultural significance of gestures, attire, or symbols. In addition to the culture that somebody is raised in, there may be other factors influencing how they see the world including socio-economic factors. For instance, certain brands, products, or lifestyle choices can convey different meanings depending on the socio-economic background of the audience. To generate good descriptions of an image, for example, a model may need to understand who took that image, and what they are trying to convey to generate a meaningful description. Beyond just cultural and socio-economic contexts, another interesting source of socio-cultural context is non-verbal cues. In visual description, recognizing non-verbal cues such as facial expressions, gestures, and body language, are not only important for visual understanding of the context but also differ between cultures and scenarios: for example, in many cultures, a shake of the head means "no", while a node of the head means yes, however in some parts of Bulgaria, Southern Italy, Greece and Turkey, the opposite is true. Building models which can understand these non-verbal cues, and the social contexts within which they occur will be a challenging, yet fruitful task. Overall, understanding cultural nuances can enable models to adjust their tone, style, and content to better align with the cultural background and expectations of the target audience, and it seems like a strong direction to explore when discovering new sources of context that impact our models.

Beyond socio-cultural factors, another source of context that I believe to be particularly interesting is emotional context. Models that are capable of detecting and responding to emotional cues can offer responses that are not just contextually relevant, but also emotionally intelligent. Wouldn't it be nice if your voice assistant realized that you were angry after repeating yourself for the thousandth time, and adjusted its handling of a situation based on that understanding? Beyond such superficial effects, emotional awareness is likely necessary in models designed for applications in mental health and customer service, as we have already seen evidence for in some of our group's work in health applications chat-bots (Figueroa et al., 2021). Beyond emotional context, general situational awareness could also provide an interesting source of potential context. Models that can adapt to and understand the immediate context of an interaction can enable responses that are aligned with a user's

current needs/circumstances. Take for example, a voice assistant that upon receiving an order for shampoo, realized that you had already ordered shampoo last week, and was able to adjust its response to confirm that such an action was desired automatically. In general, a level of sensitivity to emotion and situations has the potential to dramatically improve the effectiveness of models designed for automated customer service systems, educational content, global communication platforms, and more.

These two (or three) sources of context are only a potential set of possible sources that may be interesting. It remains useful and important for future work to continue uncovering and leveraging new sources of context in our models and designing models that are aware of these sources of context. Such an effort requires not only the pursuit of new data but also new techniques for integrating context into models – what is the most effective way to make models pay attention to socioeconomics? How can we represent user preferences? How can we tune models independently for each user? Such questions will undoubtedly be the goal of many future research projects.

14.3 Improving Grounding of Context-Aware Models

Once we have collected the context, it is a different matter to be certain that the model pays attention and leverages such context effectively. Building models that can strongly condition their output on contextual inputs is a challenging problem. This is already an issue we face in the form of hallucination in vision and language models (Liu et al., 2024a), and will only get worse as we introduce more complex and nuanced contexts and stimuli. Several key directions for future research in grounding stand out to me:

1. **Going Beyond Feature Concatenation:** Traditionally, models incorporate context through simple means: concatenating the latent spaces of the context or performing a simple cross-attention against the context to extract the most "relevant" features. While this does work, I believe it is an open question as to whether latent spaces are naturally additive/multiplicative, particularly across different modalities. Recent work from Nguyen et al. (2022) has shown that learning cross-modal manifolds which can be then aligned into a common latent space can improve overall model performance on specific multi-modal domains, and I wonder if such approaches will naturally be more effective in the long run.

2. **Understanding (and building better) pre-trained features:** Pre-trained features in vision/language and ASR models represent a challenging trade-off: in exchange for using less compute and fewer data, we force the model to take as input the pre-trained representations from a pre-existing model. Unfortunately, these pre-existing models have their own biases and training goals – information that may be required for a downstream task may not even be preserved in a particular feature set if that doesn't matter to the task (for example, localization information is often not preserved in classification models, as it doesn't matter where a pattern occurs, just "if" that pattern

occurs). We currently have a very weak understanding of how our visual features function. We know that they aren't the most effective (see `https://oatml.cs.ox.ac.uk/blog/2021/06/27/web-scraped-harmful.html` for an interesting discussion of this), but we don't yet know how to build robust feature sets that generalize to wide ranges of tasks. Perhaps this is even impossible (as there is no free lunch when it comes to representation learning). It remains interesting and challenging for future work to explore this.

3. **Building more interpretable models:** Beyond improving data integration techniques, enhancing the grounding of context-aware models also requires advancements in model interpretability. The word "interpretable" is a challenging one: what do we mean by "interpretable"? Here, I believe that "interpretable" means building models that can assign conditional probabilities in a factorizable way, where the factors are the context variables. Such a definition implies models which can "cite" their contextual sources, and generate reasonable explanations for their output in the language of a combination of reasoning over the context variables. As contexts grow, it will be more and more important for model designers to be able to get meaningful debugging output as to how their models are operating over the context space. One particularly interesting direction for exploration revolves around retrieval augmented generation (RAG, see Gao et al. (2023) for a survey), as such methods naturally have a two-stage process in which the applicable context is discovered (the data is grounded), and then a response is generated. Leveraging RAG techniques for more general-purpose grounding and interpretability represents, to me, another interesting direction for further exploration in the grounding space.

Again, such directions only represent a small segment of the possible problems in grounding, yet each is an eminently useful and challenging task that could be the subject of several PhD theses.

14.4 Learning to Understand What's Important

Beyond paying attention to the context, it is also important to pay attention to the right context and to relate the right information to a user. Determining which information in an image, video, or audio source is relevant (and which information is irrelevant) is an extremely challenging problem, and without a model that has a strong understanding of context, is a deeply under-specified one as well. As models are asked to consume and analyze increasing amounts of data, understanding what to relate, and how to relate the important details

One of the most interesting works that I have seen is work from MacLeod et al. (2017), which looks closely at what visually impaired users wish to know from an alt-text generating algorithm. This is a deeply insightful study that reveals a worrying trend: the field of visual description is generally departing from the field of alt-text generation in how images are described, and what information should be present in those images. Further, this study reveals

an insidious challenge: what is important to every person is different. Thus, determining the correct content will require a leap in models' ability to assess contextual cues, such as the focus of attention, emotional tone, or historical significance, to prioritize information that users desire (and deem relevant). Such models currently are far from a reality, and designing models and systems that are capable of such understanding will require research not only in model design and tuning, but also further studies of what humans find important in images, and how humans interact with model-generated descriptions.

Beyond vision and language tasks, it is also important for audio models to understand and relate important auditory clues. This involves not just accurately transcribing speech but understanding the intent behind words, the priority of tasks being communicated, and the emotional or situational context that surrounds the speech. Future advancements could include models that are capable of identifying the most critical parts of a conversation or instruction, based on the urgency, the speaker's emotional state, or the context in which the speech occurs. Imagine, for example, a model that can capitalize and underline stressed points in transcriptions, or produce translator/annotator notes (when such notes are required to augment the underlying transcription). Such models may require the integration of many different audio technologies and context clues including voice tone analysis, linguistic content evaluation, and contextual understanding to ensure that ASR systems can respond to or act upon what truly matters to the user.

This is not just an AI task, but one that spans several disciplines from AI and cognitive science to psychology and linguistics. By understanding how humans prioritize information, the subtle cues used to indicate importance, and how context influences perception and communication, future work may be able to design models that more closely mirror human processes of information selection and emphasis, leading to approaches which not only improve the efficiency and relevance of generated content but also enhance the naturalness and intuitiveness of human-model interactions.

14.5 Understanding and Evaluating how Models Interact with Humans

More generally than understanding what kinds of information are important to present to users of contextually aware systems/models, it is also interesting to understand how users interact with these systems as a whole. Further, it is important to not just understand but to evaluate the overall quality of these interactions. As these models become increasingly integrated into our daily lives, through devices, applications, and services, understanding and evaluating how they interact with humans becomes not just a technical challenge but a societal imperative.

While we introduce two new measures for the general evaluation of CNLG systems in this work, there is still more to be done. One particular area of improvement (which we touched on in chapter 13 is the introduction of new evaluation measures designed not just to evaluate

the accuracy, fluency, and speed of methods, but which take a multidisciplinary approach encompassing fields such as linguistics, psychology, cognitive science, and computer science to assess to what extent models are capable of nuanced interactions. Such measures could include those which measure how effectively a model can recognize and respond to the user's emotional state, or if a model can maintain a user's interest and adapt to their changing needs over time.

Beyond just simple evaluation measures, it will be important to understand the effects of human-model interactions on human behavior, cognition, and perception. Such a field is likely to offer a fertile ground for research, raising questions about dependency and trust, particularly important for contextually aware models that are integrating large datasets that humans can never hope to verify by hand. Moreover, the ethical dimensions of human-model interactions will require rigorous investigation, particularly as these technologies shift from human-in-the-loop to autonomous systems. Indeed, the design and deployment of language generation models must be guided by principles that prioritize user well-being, privacy, and autonomy, ensuring that these tools serve to enhance human capabilities rather than undermine them. It remains a challenging and open question to understand what principles are necessary, and how to develop systems to follow those principles as required.

Chapter 15

Conclusion

In this dissertation, we explored the question: *how can we understand, build, and evaluate models for contextual natural language generation*, in the context of several domains, including image and video description, and automatic speech recognition.

First, in chapter 3 we first looked at how linguistic diversity present in the dataset itself can impact models for conditional natural language generation in the image and video description domains. We examined several popular visual description datasets, demonstrating several dataset-specific linguistic patterns that models exploit to achieve strong performance. For example, at the token level, sample level, and dataset level, we found that caption diversity is a major driving factor behind the generation of generic and uninformative image and video descriptions. We further showed that state-of-the-art models even outperform held-out ground truth captions on modern metrics and that this effect is an artifact of linguistic diversity in datasets. Understanding that this linguistic diversity is key to building strong captioning models, we recommended several methods and approaches for maintaining diversity in the collection of new data and dealing with the consequences of limited diversity when using current models and metrics.

In chapter 4, we turned our attention to the collection of data for the video captioning task by asking: how can we exploit the inherent structure in the visual-linguistic data to reduce the amount of training data we need to collect. Here, we explored various active learning approaches for automatic video captioning and showed that a novel method based on cluster-regularized ensembling of models provides the best active learning approach to efficiently gather training sets for video captioning. We evaluated our approaches on the MSR-VTT and LSMDC datasets using both transformer and LSTM-based captioning models and showed that our strategy achieves high performance while using up to 60% fewer training data than the strong state-of-the-art baselines.

Then in chapter 6, we discussed how if you asked a human to describe an image, they might have done so in a thousand different ways, but image captioning models, on the other hand, were traditionally trained to generate a single "best" (most like a reference) caption. Unfortunately, this process encouraged captions that were informationally impoverished: Such captions often focused on only a subset of possible details, while ignoring other potentially

useful information in the scene. We then introduced a simple, yet novel, method: “Image Captioning by Committee Consensus" (IC^3), designed to generate a single caption that captures details from multiple viewpoints by sampling from the learned semantic space of a base captioning model, and carefully leveraging a large language model to synthesize these samples into a single comprehensive caption. Our evaluations showed that humans rated captions produced by IC^3 more helpful than those produced by SOTA models more than two-thirds of the time, and IC^3 improved the performance of SOTA automated recall systems by up to 84%, outperforming single human-generated reference captions and indicating significant improvements over SOTA approaches for visual description.

Turning our attention to automatic speech recognition (ASR), in chapter 7, we first looked at how we could learn to perform ASR tasks leveraging context drawn from video data. While traditionally, research in automated speech recognition had focused on local-first encoding of audio representations to predict the spoken phonemes in an utterance, such approaches relying on such hyper-local information tended to be vulnerable to both local-level corruption (such as audio-frame drops, or loud noises) and global-level noise (such as environmental noise, or background noise) that had not been seen during training. In this chapter, we introduced a novel approach that leveraged a self-supervised learning technique based on masked language modeling to compute a global, multi-modal encoding of the environment in which the utterance occurred. Then, using a new deep-fusion framework to integrate this global context into a traditional ASR method, we demonstrated that the resulting method could outperform baseline methods by up to 7% on Librispeech; gains on internal datasets ranged from 6% (on larger models) to 45% (on smaller models).

Next, in chapter 8, we investigated the potential of leveraging external knowledge through off-policy generated text-to-speech key-value stores, to allow for flexible post-training adaptation to new data distributions. In our approach, audio embeddings captured from text-to-speech were used, along with semantic text embeddings, to bias ASR via an approximate k-nearest-neighbor (KNN) based attentive fusion step. Our experiments on LibiriSpeech and Amazon Alexa voice assistant/search datasets showed that the proposed approach could reduce domain adaptation time by up to 1K GPU-hours while providing up to 3% WER improvement compared to a fine-tuning baseline, suggesting a promising approach for adapting production ASR systems in challenging zero and few-shot scenarios.

To wrap up our discussion on building models, in chapter 9, we introduced CLC: Contrastive Learning for Conversations, a family of methods for contrastive fine-tuning of models in a self-supervised fashion, making use of easily detectable artifacts in unsuccessful conversations with assistants. We demonstrated that our CLC family of approaches could improve the performance of ASR models on OD3, a new public large-scale semi-synthetic meta-dataset of audio task-oriented dialogues, by up to 19.2%. These gains transferred to real-world systems as well, where we showed that CLC could help to improve performance by up to 6.7% over baselines.

Turning our attention to evaluation, in chapter 11, we challenged the adequacy of existing metrics in semantically diverse contexts and introduced a novel approach for evaluating conditional language generation models, leveraging a family of meta-metrics that built on

existing pairwise distance functions. These meta-metrics assessed not just single samples, but distributions of reference and model-generated captions using small sample sets. We demonstrated our approach through a case study of visual description which revealed not only how current models prioritized single-description quality over diversity but also shed light on the impact of sampling methods and temperature settings on description quality and diversity.

To conclude the dissertation, in chapter 12, we explored how we could leverage the latest advances in large language models to judge the distance between captions while accounting for the variance that is generated by user contexts. In our evaluations, CLAIR, our novel method that leveraged the zero-shot language modeling capabilities of large language models (LLMs), demonstrated a stronger correlation with human judgments of caption quality compared to existing measures. Notably, on Flickr8K-Expert, CLAIR achieved relative correlation improvements over SPICE of 39.6% and over image-augmented methods such as RefCLIP-S of 18.3%. Moreover, CLAIR provided noisily interpretable results by allowing the language model to identify the underlying reasoning behind its assigned score.

From the inception of artificial intelligence, the ultimate objective has been to create agents that can effectively interact with our world. I strongly believe that the work presented here can provide a blueprint for building models that do not just process language as input but can understand and engage with the nuanced contexts and situations in which language occurs and generate responses to users that are both natural and relevant. Equipping such models with the acumen to perceive the world beyond their inputs and respond accordingly is a stride towards creating systems that understand the essence of human communication and interaction, and will lead to a new era where artificial intelligence is capable of mirroring the depth and complexity of human intellect and empathy.

Part V

References

References

Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *ArXiv preprint*, abs/1511.03292.

Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio. Association for Computational Linguistics.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24206–24221.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198.

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *ArXiv preprint*, abs/1707.02268.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander G. Schwing. 2019. Sequential latent spaces for modeling the intention during diverse image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4260–4269. IEEE.

Jake Archibald. 2023. Writing great alt text: Emotion matters.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE.

Deepak Baby, Pasquale D'Alterio, and Valentin Mendelev. 2022. Incremental learning for rnn-transducer based speech recognition models. In *Interspeech 2022*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Natã Miccael Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 543. ACM.

Michéle Basseville. 2013. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *ArXiv preprint*, abs/1912.00578.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. 2022. Retrieval-augmented diffusion models. *ArXiv preprint*, abs/2204.11824.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich

Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 59–66. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 839–850.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *ArXiv preprint*, abs/1808.01340.

Edresson Casanova, Julian Weber, Christopher Dane Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir A. Ponti. 2022. Yourtts: Towards zero-shot multi-speaker TTS and

zero-shot voice conversion for everyone. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2709–2720. PMLR.

David M. Chan and Shalini Ghosh. 2022a. Content-context factorized representations for automated speech recognition. In *Interspeech*.

David M. Chan and Shalini Ghosh. 2022b. Content-context factorized representations for automated speech recognition. In *Interspeech*.

David M. Chan, Shalini Ghosh, Debmalya Chakrabarty, and Bjorn Hoffmeister. 2022a. Multi-modal pre-training for automated speech recognition. In *ICASSP*.

David M Chan, Shalini Ghosh, Debmalya Chakrabarty, and Björn Hoffmeister. 2022b. Multi-modal pre-training for automated speech recognition. In *ICASSP*.

David M Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. 2023a. Domain adaptation with external off-policy acoustic catalogs for scalable contextual end-to-end automated speech recognition. In *ICASSP*.

David M Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. 2023b. Using external off-policy speech-to-text mappings in contextual end-to-end automated speech recognition. In *ICASSP*.

David M. Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A. Ross, Bryan Seybold, and John F. Canny. 2022c. What's in a caption? dataset-specific linguistic diversity and its effect on visual description models and metrics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 4739–4748. IEEE.

David M Chan, Yiming Ni, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. 2022d. Distribution aware metrics for conditional natural language generation. *ArXiv preprint*, abs/2209.07518.

Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. 2021. Context-aware transformer transducer for speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 503–510. IEEE.

Shuo-Yiin Chang et al. 2023. Context-aware end-to-end asr using self-attentive embedding and tensor fusion. In *ICASSP*.

Moitreya Chatterjee and Alexander G Schwing. 2018. Diverse and coherent paragraph generation from images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 729–744.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Ming Chen, Yingming Li, Zhongfei Zhang, and Siyu Huang. 2018. Tvt: Two-view transformer network for video captioning. In *Asian Conference on Machine Learning*, pages 847–862. PMLR.

Shaoxiang Chen and Yu-Gang Jiang. 2021. Motion guided region message passing for video captioning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1523–1532. IEEE.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *ArXiv preprint*, abs/2209.06794.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325.

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L. Seltzer, and Christian Fuegen. 2019. Joint grapheme and phoneme embeddings for contextual end-to-end ASR. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3490–3494. ISCA.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N Sainath, Patrick Nguyen, Liangliang Cao, et al. 2021. Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 873–880. IEEE.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Visual programming for text-to-image generation and evaluation. *ArXiv preprint*, abs/2305.15328.

Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. 2019. Why can't I dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in*

Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 851–863.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3031–3045, Online. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. 2008. Towards scalable dataset construction: An active learning approach. In *Computer Vision – ECCV 2008*, pages 86–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, page 746–751. AAAI Press.

Ido Dagan and Sean P. Engelson. 1995. Committee-based sampling for training probabilistic classifiers. In *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*, ICML'95, page 150–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2989–2998. IEEE Computer Society.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Yue Deng, KaWai Chen, Yilin Shen, and Hongxia Jin. 2018. Adversarial active learning for sequences labeling and generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4012–4018. ijcai.org.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David A. Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10695–10704. Computer Vision Foundation / IEEE.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv preprint*, abs/2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Saket Dingliwa, Ashish Shenoy, Sravan Bodapati, Ankur Gandhe, Ravi Teja Gadde, and Katrin Kirchhoff. 2022. Domain prompts: Towards memory and compute efficient domain adaptation of asr systems. In *Interspeech 2022*.

Ramanand Durga. 1980. Semantic distance and semantic judgment. *Outstanding Dissertations in Bilingual Education Recognized by the National Advisory Council on Bilingual Education, 1979*, page 15.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2022. Injecting semantic concepts into end-to-end image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17988–17998. IEEE.

Joshua Feinglass and Yezhou Yang. 2021. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260, Online. Association for Computational Linguistics.

Caroline A Figueroa, Tiffany C Luo, Andrea Jacobo, Alan Munoz, Minx Manuel, David Chan, John Canny, and Adrian Aguilera. 2021. Conversational physical activity coaches

for spanish and english speaking women: a user design study. *Frontiers in Digital Health*, 3:747153.

Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1141–1150. IEEE Computer Society.

Ankur Gandhe, Ariya Rastrow, and Bjorn Hoffmeister. 2018. Scalable language model adaptation for spoken dialogue systems. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 907–912. IEEE.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. 2013. Optimized product quantization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):744–755.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Mark E Glickman. 1995. The glicko system. *Boston University*, 16:16–17.

Anirudh Goyal, Abram L. Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adrià Puigdomènech Badia, Arthur Guez, Mehdi Mirza, Peter C. Humphreys, Ksenia Konyushkova, Michal Valko, Simon Osindero, Timothy P. Lillicrap, Nicolas Heess, and Charles Blundell. 2022a. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7740–7765. PMLR.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022b. News summarization and evaluation in the era of gpt-3. *ArXiv preprint*, abs/2209.12356.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

Chulaka Gunasekara et al. 2020. Noesis ii: Predicting responses, identifying success, and managing complexity in task-oriented dialogue. In *AAAI: Workshop on Dialog System Tech Challenges*.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3610–3614. ISCA.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021a. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021b. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Steven C. H. Hoi, Rong Jin, Jianke Zhu, and Michael R. Lyu. 2009. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Trans. Inf. Syst.*, 27(3).

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python, zenodo, 2020.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1419–1429. IEEE.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *ArXiv preprint*, abs/2211.09699.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *ArXiv preprint*, abs/2303.11897.

Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. 2016a. Active learning for speech recognition: the power of gradients. *ArXiv preprint*, abs/1612.03226.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016b. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Aaron Jaech and Mari Ostendorf. 2018. Personalized language model for query auto-completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705, Melbourne, Australia. Association for Computational Linguistics.

Dhruv Jhamb, David Chan, John F Canny, and Avideh Zakhor. 2022. Hallucination is all you need: Using generative models for test time data augmentation.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4565–4574. IEEE Computer Society.

Jungseock Joo and Kimmo Kärkkäinen. 2020. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 1–5.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, Seattle, United States. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *ArXiv preprint*, abs/2006.01997.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6271–6280. Computer Vision Foundation / IEEE.

Seokhwan Kim et al. 2021. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *ASRU*.

Suyoun Kim and Florian Metze. 2018. Dialog-context aware end-to-end speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 434–440. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Franz Klein, Shweta Mahajan, and Stefan Roth. 2022. Diverse image captioning with grounded style. In *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, September 28–October 1, 2021, Proceedings*, pages 421–436. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv preprint*, abs/2205.11916.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3337–3345. IEEE Computer Society.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *ArXiv preprint*, abs/1909.09577.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Wai-Chung Kwan et al. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intel. Res.*, 20(3).

Rosa Lafer-Sousa, Katherine L Hermann, and Bevil R Conway. 2015. Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, 25(13):R545–R546.

Tomer Levinboim, Ashish V. Thapliyal, Piyush Sharma, and Radu Soricut. 2021. Quality estimation for image captions based on large-scale human evaluations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3157–3166, Online. Association for Computational Linguistics.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. 2017. MMD GAN: towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2203–2213.

Dawei Li, Serafettin Tasci, Shalini Ghosh, Jingwen Zhu, Junting Zhang, and Larry Heck. 2019a. IMOD: Efficient incremental learning for mobile object detection. In *Proceedings of the ACM/IEEE Symposium on Edge Computing (SEC)*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023a. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Xiangyang Li, Shuqiang Jiang, and Jungong Han. 2019b. Learning object context for dense captioning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8650–8657. AAAI Press.

Yi Li and Nuno Vasconcelos. 2019. REPAIR: removing representation bias by dataset resampling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9572–9581. Computer Vision Foundation / IEEE.

Vladislav Lialin, Stephen Rawls, David Chan, Shalini Ghosh, Anna Rumshisky, and Wael Hamza. 2023. Scalable and accurate self-supervised multimodal representation learning without aligned video and text data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 390–400.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. 2017. Recurrent topic-transition GAN for visual paragraph generation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3382–3391. IEEE Computer Society.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering. In *Interspeech*.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. 2015. Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 899–907, Lisbon, Portugal. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Bing Liu and Ian Lane. 2017. Dialog context language modeling with recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 5715–5719. IEEE.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. 2021a. O2NA: An object-oriented non-autoregressive approach for controllable video captioning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 281–292, Online. Association for Computational Linguistics.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *ArXiv preprint*, abs/2103.10385.

Zhenyu Liu and Reza Modarres. 2011. A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, 23(3):605–615.

Yadan Luo, Zi Huang, Zheng Zhang, Ziwei Wang, Jingjing Li, and Yang Yang. 2019. Curiosity-driven reinforcement learning for diverse visual paragraph generation. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 2341–2350.

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, pages 5988–5999. ACM.

Sridhar Mahadevan, Bamdev Mishra, and Shalini Ghosh. 2018. A unified framework for domain adaptation using metric learning on manifolds. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML-PKDD*, volume 11052, pages 843–860.

Sridhar Mahadevan, Bamdev Mishra, and Shalini Ghosh. 2019. A unified framework for domain adaptation using metric learning on manifolds. In *ECML PKDD*.

Shweta Mahajan, Iryna Gurevych, and Stefan Roth. 2020. Latent normalizing flows for many-to-many cross-domain mappings. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shweta Mahajan and Stefan Roth. 2020. Diverse image captioning with context-object split latent spaces. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur. 2018. A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 250–257. IEEE.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. 2018. Show and tell more: Topic-oriented multi-sentence image captioning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4258–4264. ijcai.org.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Bonan Min et al. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Soumyajit Mitra, Swayambhu Nath Ray, Bharat Padi, Raghavendra Bilgi, Harish Arsikere, Shalini Ghosh, Ajay Srinivasamurthy, and Sri Garimella. 2023. Unified modeling of multi-domain multi-device ASR systems. In *TSD*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *ArXiv preprint*, abs/2111.09734.

Mathew Monfort, SouYoung Jin, Alexander H. Liu, David Harwath, Rogério Feris, James R. Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14871–14881. Computer Vision Foundation / IEEE.

Ladislav Mosner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Ken'ichi Kumatani, Shiva Sundaram, Roland Maas, and Björn Hoffmeister. 2019. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6475–6479. IEEE.

Tsendsuren Munkhdalai, Khe Chai Sim, Angad Chandorkar, Fan Gao, Mason Chua, Trevor Strohman, and Françoise Beaufays. 2022. Fast contextual adaptation with neural associative memory for on-device personalized speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6632–6636. IEEE.

Sharan Narang and Aakanksha Chowdhery. 2022. Pathways language model (palm): Scaling to 540 billion parameters for breakthrough performance.

Nam D Nguyen, Jiawei Huang, and Daifeng Wang. 2022. A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data. *Nature computational science*, 2(1):38–46.

Scott Novotney, Sreeparna Mukherjee, Zeeshan Ahmed, and Andreas Stolcke. 2022. CUE vectors: Modular training of language models conditioned on diverse contextual signals. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3368–3379, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2022a. Introducing chatgpt.

OpenAI. 2022b. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. 2021. Improving video captioning with temporal composition of a visual-syntactic embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3039–3049.

Hieu Pham et al. 2023. Combined scaling for zero-shot transfer learning. *Neurocomputing*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021c. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, abs/2204.06125.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv preprint*, abs/1910.01108.

Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8537–8541. IEEE.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, IDA '01, page 309–318, Berlin, Heidelberg. Springer-Verlag.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Ashish Shenoy, Sravan Bodapati, and Katrin Kirchhoff. 2021. Contextual biasing of language models for speech recognition in goal-oriented conversational agents. *ArXiv preprint*, abs/2103.10325.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4155–4164. IEEE Computer Society.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *ArXiv preprint*, abs/2305.13040.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. 2020. Don't judge an object by its context: Learning to overcome contextual bias. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11067–11075. IEEE.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5971–5980. IEEE.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

P Slade and Tamás D Gedeon. 1993. Bimodal distribution removal. In *International Workshop on Artificial Neural Networks*, pages 249–254. Springer.

Alan F Smeaton, Yvette Graham, Kevin McGuinness, Noel E O'Connor, Seán Quinn, and Eric Arazo Sanchez. 2019. Exploring the impact of training data bias on automatic generation of video captions. In *International Conference on Multimedia Modeling*, pages 178–190. Springer.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

StabilityAI. 2023. Stability ai launches the first of its stablelm suite of language models.

Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on image captioning. *ArXiv preprint*, abs/2107.06912.

Jonathan Stroud, David Ross, et al. 2020. D3d: Distilled 3d networks for video action recognition. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 625–634.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? *arXiv preprint arXiv:2109.09115*.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *ArXiv preprint*, abs/1911.08460.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pages 633–645.

Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. 2022. Improved vector quantized diffusion models. *ArXiv preprint*, abs/2205.16007.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

MosaicML NLP Team. 2023. Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

TJ Thomson, Daniel Angus, Paula Dootson, Edward Hurcombe, and Adam Smith. 2022. Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities. *Journalism Practice*, 16(5):938–962.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. 2022. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Grigorios Tsoumakas and Min-Ling Zhang. 2009. Learning from multi-label data.

Emiru Tsunoo, Yosuke Kashiwagi, and Shinji Watanabe. 2021. Streaming transformer asr with blockwise synchronous beam search. In *SLT*.

Iván Vallés-Pérez, Julian Roth, Grzegorz Beringer, Roberto Barra-Chicote, and Jasha Droppo. 2021. Improving multi-speaker tts prosody variance with a residual encoder and normalizing flow. In *Interspeech 2021*.

Aäron van den Oord and Joni Dambre. 2015. Locally-connected transformations for deep gmms. In *International Conference on Machine Learning (ICML): Deep learning Workshop*, pages 1–8.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Vijay V Vazirani. 2001. *Approximation algorithms*, volume 1. Springer.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4534–4542. IEEE Computer Society.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *ArXiv preprint*, abs/1610.02424.

Sudheendra Vijayanarasimhan and Kristen Grauman. 2009. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 2262–2269. IEEE Computer Society.

Sudheendra Vijayanarasimhan and Kristen Grauman. 2011. Large-scale live active learning: Training object detectors with crawled data and crowds. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1449–1456. IEEE Computer Society.

Sudheendra Vijayanarasimhan and Kristen Grauman. 2012. Active frame selection for label propagation in videos. In *Computer Vision – ECCV 2012*, pages 496–509, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sudheendra Vijayanarasimhan, Prateek Jain, and Kristen Grauman. 2010. Far-sighted active learning on a budget for image and video recognition. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3035–3042. IEEE Computer Society.

Pascal Wallisch. 2017. Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain:"the dress". *Journal of Vision*, 17(4):5–5.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2022a. On distinctive image captioning via comparing and reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5756–5766.

Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. 2021a. Multimodal self-supervised learning of general audio representations. *ArXiv preprint*, abs/2104.12807.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Qingzhong Wang and Antoni B. Chan. 2019. Describing like humans: On diversity in image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019,*

Long Beach, CA, USA, June 16-20, 2019, pages 4195–4203. Computer Vision Foundation / IEEE.

Qingzhong Wang, Jia Wan, and Antoni B Chan. 2020. On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021b. Faier: Fidelity and adequacy ensured image caption evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14050–14059. Computer Vision Foundation / IEEE.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2957–2964. IJCAI/AAAI Press.

Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R. Hershey. 2017. Student-teacher network learning with enhanced features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 5275–5279. IEEE.

Kai Wei et al. 2021. Attentive contextual carryover for multi-turn end-to-end spoken language understanding. In *2021 ASRU*, pages 837–844. IEEE.

Ian Williams, Anjuli Kannan, Petar S. Aleksic, David Rybach, and Tara N. Sainath. 2018. Contextual speech recognition in end-to-end neural network systems using beam search. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2227–2231. ISCA.

Daniel J Wilson. 2019. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. 2022. Visual clues: Bridging vision and language foundations for image paragraph captioning. *ArXiv preprint*, abs/2206.01843.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Rong Yan, Jie Yang, and Alexander Hauptmann. 2003. Automatically labeling video data using multi-class active learning. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, page 516, USA. IEEE Computer Society.

Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. 2021a. Non-autoregressive coarse-to-fine video captioning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3119–3127. AAAI Press.

Chao-Han Huck Yang, Yi-Le Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative asr error correction with large language models. In *ASRU*.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-sensitive conditional generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020a. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

Linjie Yang, Kevin D. Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense captioning with joint inference and visual context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1978–1987. IEEE Computer Society.

Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. 2020b. Hierarchical scene graph encoder-decoder for image paragraph captioning. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 4181–4189.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2021b. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *ArXiv preprint*, abs/2305.10400.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Context and attribute grounded dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6241–6250. Computer Vision Foundation / IEEE.

Ilmi Yoon, Umang Mathur, Brenna Gibson, Tirumalashetty Pooyan Fazli, and Joshua Miele. 2019. Video accessibility for the visually impaired. In *International Conference on Machine Learning AI for Social Good Workshop*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *ArXiv preprint*, abs/2205.01917.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *ArXiv preprint*, abs/2204.00598.

Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2019. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):710–722.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021a. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Jie Zhang, Junting Zhang, Shalini Ghosh, Dawei Li, Jingwen Zhu, Heming Zhang, and Yalin Wang. 2020a. Regularize, expand and compress: Nonexpansive continual learning. In *IEEE Winter Conference on Applications of Computer Vision, WACV*, pages 843–851.

Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry P. Heck, Heming Zhang, and C.-C. Jay Kuo. 2020b. Class-incremental learning via deep model consolidation. In *IEEE Winter Conference on Applications of Computer Vision, WACV*.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *ECCV*, pages 766–782.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020c. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7829–7833. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020d. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020e. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zi-qiang Zhang, Yan Song, Jian-Shu Zhang, Ian McLoughlin, and Li-Rong Dai. 2020f. Semi-supervised end-to-end ASR via teacher-student learning with conditional posterior distribution. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3580–3584. ISCA.

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020g. Object relational graph with teacher-recommended learning for video captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13275–13285. IEEE.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge. volume abs/2305.12091.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-fusion end-to-end contextual biasing. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1418–1422. ISCA.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *2021 IEEE/CVF International Conference on Computer*

Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 14810–14820. IEEE.

Kaiqi Zhao, Hieu Duy Nguyen, Animesh Jain, Nathan Susanj, Athanasios Mouchtaris, Lokesh Gupta, and Ming Zhao. 2022. Knowledge distillation via module replacing for automatic speech recognition with recurrent neural network transducer. In *Interspeech 2022*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2306.05685.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8739–8748. IEEE Computer Society.

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018c. End-to-end dense video captioning with masked transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8739–8748. IEEE Computer Society.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023a. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *ArXiv preprint*, abs/2303.06594.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

Part VI

Appendices

Appendix A

Appendix for Dataset-Specific Linguistic Diversity and Its Effect on Visual Description Models and Metrics

A.1 Datasets

We investigate four primary datasets for the work in chapter 3. An overview of the datasets is given in Table A.1.

MSR-VTT The MSR-VTT (MSR Video to Text) (Xu et al., 2016) dataset is a medium-scale open domain benchmark for visual description. It was originally collected using 257 YouTube search queries across 20 categories, with 118 videos collected for each query (41.2 Hours). The dataset is annotated with 20 captions per video by 1,327 Amazon Mechanical Turk workers. Each video has a duration between 10 and 30 seconds, with an average of two shots per clip.

VATEX The VATEX dataset (Wang et al., 2019) is a medium-scale open domain video description benchmark, based on a subset of the Kinetics-600 dataset for action recognition. VATEX consists of 41,269 video clips, and each clip is annotated with 10 unique descriptive captions by 2,159 Amazon Mechanical Turk workers.

MSVD The MSVD (Microsoft Video Description) dataset (Chen and Dolan, 2011) is a small-scale open domain benchmark for video description comprised of 1,970 YouTube clips of 4-10 seconds each, collected by asking Amazon Mechanical Turk workers to link a video, start time, and end time from YouTube that depicts a specific, short action. Each video is then annotated with an average of 41 ground truth descriptions by 835 Amazon Mechanical Turk workers.

Dataset	Domain	Categories	Videos	Avg. Length	Length (hrs)	Annotations/ Video	Annotation Method
MSR-VTT	open	20	10K	20s	41.2	20	AMT
VATEX	open	600	42K	-	-	10	AMT
MSVD	open	218	1970	10s	41	35.5	AMT
MS-COCO	open	-	120K	-	-	5	AMT

Table A.1: An overview of the datasets that we analyze in chapter 3. All of the datasets are open-domain, with a focus on video description. Additionally, each of the datasets include more than one ground truth description per video, which we use to validate the performance of ground truth data, without collecting additional human results. Notably, all of these methods use AMT as their annotation method.

MSCOCO The Microsoft Common Objects in Context (MS-COCO) (Lin et al., 2014) dataset is a large-scale open-domain benchmark for image description. MS-COCO consists of more than 120,000 images of complex scenes including people, animals, and common objects. Each image is annotated with five ground truth descriptions.

A.2 Experimental Details

In this section, we present detailed experimental details corresponding to our experiments. Along with these experimental details, we make the code for our work available at `https://github.com/CannyLab/vdtk`. Note that numbers may differ slightly between the released code, and our presented experiments due to the tokenization scheme. For our released code, we use the Spacy[1] tokenizer to compute all metrics, as it is significantly more efficient in practice than the Stanford tokenizer[2], however for academic purposes, we compute the metrics with the Stanford tokenizer to avoid tokenization shift. In most cases, the difference in the metrics between tokenization methods is negligible (or very small).

A.2.1 Motivation: Leave One Out Ground Truth Performance

To generate an estimate of human performance on the selected datasets, we use a procedure called "leave one out" performance. Let a dataset $\mathcal{D}$ be composed of N samples $S_0 \dots S_N$. For each sample S_i, there may be K_i possible reference captions, $C_0^i \dots C_{K_i}^i$. In order to compute the leave one out performance of human samples for the dataset, we first select a hypothesis caption $H_i \in \{C_0^i \dots C_{K_i}^i\}$. We then compute the updated reference set $R_i = \{C_0^i \dots C_{K_i}^i\}/\{H_i\}$. In the case that H_i is duplicated within R_i, *we allow the duplicate*

[1] `https://spacy.io/`

[2] `https://nlp.stanford.edu/software/tokenizer.html`

Dataset	BLEU@4	METEOR	ROUGE	CIDEr
MSVD	0.453 (0.644)	0.370 (0.419)	0.689 (0.795)	1.038 (1.115)
MSR-VTT	0.209 (0.472)	0.247 (0.312)	0.487 (0.648)	0.426 (0.600)
VATEX	0.234 (0.342)	0.249 (0.235)	0.478 (0.503)	0.611 (0.576)
MS-COCO	0.152 (0.410)	0.228 (0.311)	0.438 (0.609)	0.788 (1.409)

Table A.2: Raw leave-one-out score estimates for each of the datasets (SOTA in parentheses).

to remain to maximize the possible human score. In the case that there is only one (or fewer) captions for a video, we drop those captions from the computation. We then use the reference sets $R_0 \dots R_N$ and hypotheses $H_0 \dots H_N$ to compute the "leave-one-out" score for the dataset. Clearly, this is an estimate of the ground truth performance, as it is a random sample of the possible "leave-one-out" hypotheses sets.

Because some of the metrics (particularly CIDEr) are dataset dependent, it would be intractable to compute all possible hypotheses sets. Instead of computing all possible hypotheses sets, we perform 750 iterations of this sampling procedure and use the mean of the iterations to achieve our final "leave-one-out" estimates presented in chapter 3. We found empirically that 750 iterations were sufficient across all of the datasets to achieve a stable mean. The raw values of the "leave-one-out" estimates are presented in Table A.2, alongside the state of the art results.

A.2.2 Motivation: Semantically Masked Leave One Out performance

To test the performance of ground truths without semantic information, we devised an experiment based on the leave-one-out experiments above, however, focused on removing semantic information. To compute this value, we select hypotheses as in Appendix A.2.1, however for both the captions in the reference and the captions in the ground truth, we replace any token identified by the Spacy part of speech analysis as a noun, proper noun, or verb with a *unique* mask token. This means that this unique mask token will achieve a 0 in any associated token-based metric, as it will not match any semantic token in the ground truth. Table A.3 gives the full performance on each of the datasets in the masked setup.

A.2.3 Caption Diversity: Token Metrics

In this work, we compute several metrics based on token-level diversity, demonstrated in Table 3.1 in chapter 3. The number of unique tokens is equal to the number of tokens in the dataset as computed by the Stanford PTB tokenizer. This number does not do any lemmatizing or stemming, thus, is an upper bound for the vocabulary complexity. We then compute three additional metrics, the within-sample uniqueness, the between-sample uniqueness, and the 90% head of the vocabulary. The within-sample uniqueness corresponds

Dataset	BLEU@4	METEOR	ROUGE	CIDEr
MSVD	0.289 (0.453)	0.097 (0.370)	0.442 (0.689)	0.502 (1.038)
MSR-VTT	0.123 (0.209)	0.085 (0.247)	0.387 (0.487)	0.327 (0.426)
VATEX	0.132 (0.234)	0.201 (0.249)	0.391 (0.478)	0.511 (0.611)
MS-COCO	0.079 (0.152)	0.198 (0.228)	0.396 (0.438)	0.684 (0.788)

Table A.3: Raw leave-one-out score estimates under semantic masking for each of the datasets (Non-masked in parentheses).

Dataset	Unique	BS-Unique	WS-Unique	Head
MSVD	9455	1.21%	11.8%	944
MSR-VTT	22780	0.76%	21.55%	1636
VATEX	31364	0.33 %	24.87%	1363
MS-COCO	35341	0.22%	33.76%	824

Table A.4: Vocabulary metrics for each of the datasets. Unique: The number of unique tokens. BS-Unique: Average percent of tokens per description that are unique. WS-Unique: Average percent of of tokens that are unique within a sample. Head: The number of unique tokens comprising 90% of the total tokens.

to the percentage of tokens that are unique within a sample - i.e. the percentage of tokens that appear exactly once among the references for any particular image or video. We then average this number over all of the samples to get the number presented in Table 3.1. The between-sample uniqueness is a measure of the percentage of tokens in each sample that are unique at the *dataset* level, i.e. the percentage of tokens among the tokens in the reference set of a single sample that do not appear in any other caption in the dataset. These per-sample numbers are then averaged across the dataset to get the number presented in Table A.4. Finally, the 90% head corresponds to the number of tokens that make up 90% of the mass of the total number of tokens in the dataset. This is an approximate measure of how long-tailed the distribution is. The 90% number is selected empirically (further analysis could look at the full cumulative distribution of the token counts). Table A.4 replicates Table 3.1 in chapter 3, however includes between-sample token uniqueness.

We also compute many of the same metrics restricted to counting nouns and verbs (as identified by the Spacy POS tagger). Each of the above metrics is computed the same way, however instead of considering all tokens, we consider only tokens that are tagged as either nouns or verbs during the computation of the metrics. Table A.5 demonstrates the full results of this experiment, plus an additional metric: the average number of tokens per caption which also appears in Table 3.2 in chapter 3.

Dataset	WSNU	BSNU	WSVU	BSVU	NC	VC	NH	VH	NPC	VPC	TPC
MSVD	12.6%	1.9%	14.8%	1.5%	4985	1773	755	229	2.39	1.10	7.03
MSR-VTT	23.1%	1.2%	29.4%	0.8%	12697	3639	1512	293	3.28	1.32	9.32
VATEX	26.9%	0.67%	35.7%	0.3%	16670	4975	1161	338	4.37	2.10	15.29
MS-COCO	34.9%	0.41%	55.8%	0.2%	20155	4200	723	184	3.71	1.02	11.33

Table A.5: Part of speech distributions for each of the datasets. DS: Dataset. WSNU: Within sample noun uniqueness. BSNU: Between sample noun uniqueness. WSVU: Within sample verb uniqueness. BSVU: Between sample verb uniqueness. NC: Unique noun count. VC: Unique verb count. NH: Noun head (90% of mass). V: Verb Head (90% of mass). VPC: Average number of verbs per caption. NPC: Average number of nouns per caption. TPC: Average number of tokens per caption.

A.2.4 Caption Diversity: N-Gram Metrics

To explore the diversity of samples at an n-gram level, we introduce two novel metrics, the Expected Vocab Size @ N (EVS@N), and the Expected Number of Decisions @ N (ED@N). Both of these metrics measure the diversity of the language at an n-gram level by exploring the properties of an n-gram language model trained on the dataset. In this section, we discuss the explicit definition of these metrics. For all n-grams, we use an n-gram language model based on tokens extracted with the Stanford PTB tokenizer. In all cases, we pad the references with $[BOS]$ and $[EOS]$ tokens to allow the model to handle the beginning and end of the sequences. For WikiText-103, we create individual reference sentences by splitting on '.' tokens, and pad each of these references individually with $[BOS]$ and $[EOS]$ tokens.

A.2.4.1 Expected Vocab Size @ N

The EVS@N metric is a measure of how many n-grams *do not* act as 1-grams in practice in the dataset. This measure is computed by looking at the entropy of the next-token distribution of an n-gram language model. For a sequence of words $w_0, \dots w_{n-1}$, we first compute the distribution $P(w_n|w_0, \dots, w_{n-1})$. If this distribution has 0 entropy (i.e. it assigns all of the probability mass to a single next token), then we consider this n-gram a "static n-gram". If the entropy is non-zero, then we consider it a "dynamic n-gram". The EVS@N can then be computed as the proportion of dynamic n-grams

$$\text{EVS@N} = \frac{|\text{dynamic n-grams}|}{|\text{static n-grams}| + |\text{dynamic n-grams}|}$$

This measures a set of effective n-grams in the data (i.e. the size of the n-gram vocab), as it coalesces n-grams where no decisions are made into a single logical unit.

A.2.4.2 Expected Decisions @ N

The ED@N metric is a measure of how many decisions an n-gram language model has to make for a sequence of N tokens. ED@N is a counting measure of the EVS@N - i.e. how many dynamic n-grams are expected in a sequence of length n. For a $K-gram$ language model, this measure is explicitly computed as:

$$\text{ED@N} = 1 + \sum_{i=1}^{N-1}(1 - EVS@K)(0) + (EVS@K)(1)$$

In this work, for the first token we use a 2-gram language model ($K = 2$), for the second token we use a 3-gram language model ($K = 3$), and for any additional tokens, we use a 4-gram language model ($K = 4$).

A.2.5 Sample Diversity: Within Sample Diversity

We use several techniques to measure the within-sample semantic diversity of the data. In all of these cases, the notion of semantics is somewhat subjective. In this work, we use a BERT-style embedding trained for sentence similarity, called MP-Net (Song et al., 2020) to embed each reference description as a 384-dimensional vector. We leverage the implementation in Sentence Transformers[3], which is pre-trained on over 1 billion sentence pairs.

Figure 3.2 measures the minimum within-sample distances, i.e. it looks for the closest pair of references in each sample, and plots the distance between them. Thus, for a dataset of length N with a set of samples $S_0 \dots S_N$ and captions $S_i^0 \dots S_i^{K_i}$, this histogram plots the distribution over all descriptions of

$$H_{ij} = \min_{k \neq n} ||S_i^k - S_i^j||$$

In order to avoid obvious issues with repetition in the semantics, we use only the unique set of captions in a sample, as opposed to allowing for duplicates, which would force H_i to zero for any sample with repeated captions (actually exaggerating the effect in Figure 3.2. We don't allow this in order to avoid biasing our experiments to datasets such as VATEX, which explicitly remove exact duplicates. Close duplicates are not affected, as can clearly be seen by MSVD, which contains a lot of semantic redundancy. Note that this is a distribution over all references (as opposed to samples).

Another method of measuring semantic diversity is by looking at the spread of the semantics in the sample. While we use the literal variance of the within-sample pairwise distance distribution in Figure 3.3, we can also look at other measures of spread. Figure A.1 demonstrates the difference (as a percent of the mean) between the mean of the inter-sample distances and the closest inter-sample distance. When this percentage is high, the descriptions

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

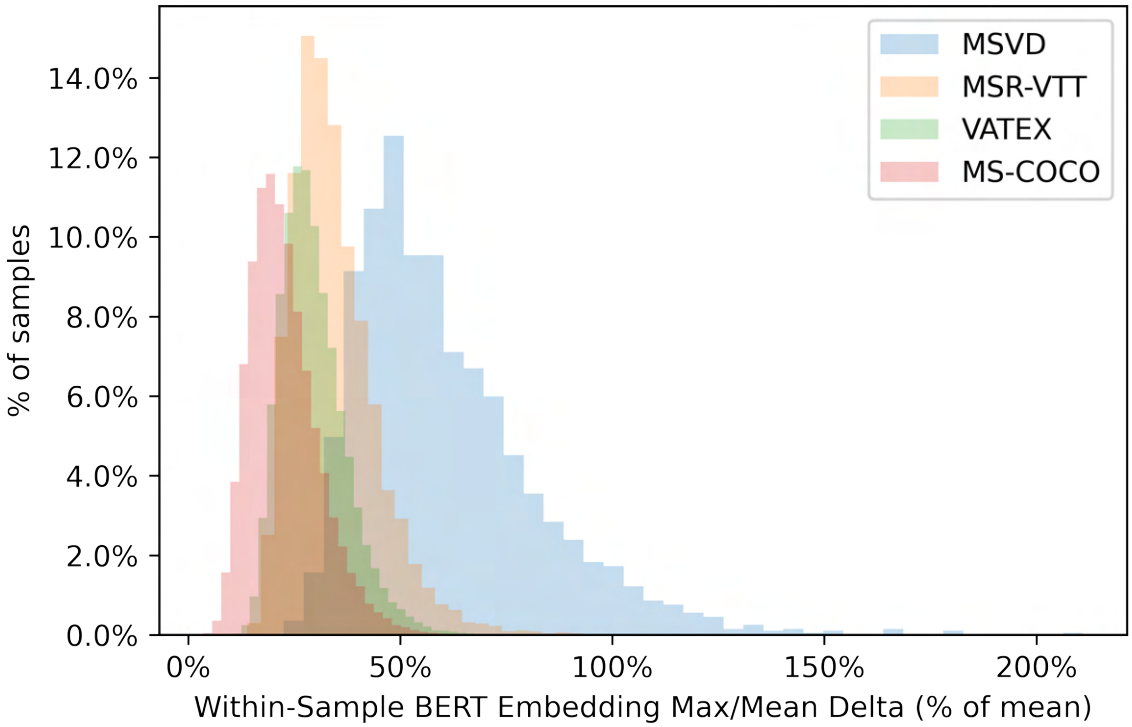


Figure A.1: Plot demonstrates the difference between the closest semantic vector, and the mean of the semantic vectors. In all cases, the mean will always be further than the closest sample, however, a low delta suggests a more equal spread of references, while a high delta represents highly redundant samples.

are relatively spread out for a sample, with clusters of descriptions that are close together in semantic space. If the percentage is low, the descriptions for a sample are well-distributed (mostly equidistant) in the semantic space.

Figure A.2 gives a general overview for the video description datasets of the exact-duplicate distribution of the descriptions. While most of the samples have high within-sample uniqueness, there are some samples that are highly redundant (and in the case of MSVD, have exact-redundancy of as much as $\sim 50\%$.

A.2.6 Dataset Diversity: Number of Ground Truths

To investigate how the number of ground truth metrics impacts the computation of the metrics, we performed several leave one out experiments as in Appendix A.2.1 where we restricted the size of R_i for each sample to a certain number of references r by randomly

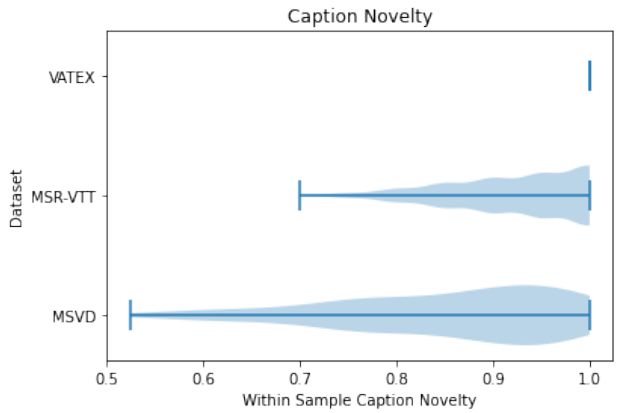


Figure A.2: Violin plot demonstrating the distribution of caption novelty - i.e. how many captions in each sample are not exact matches in the text space. As we can see, while the vast majority of captions are novel in some datasets, in datasets like MSVD, there some samples which have high *exact* redundancy.

sampling r elements without replacement from the original reference set. This allows us to measure the approximate performance of the methods if the number of ground truths was reduced. The results of this experiment are given in Figure A.3. We can see from Figure A.3 that except for CIDEr, increasing the number of ground truths increases the leave one out performance of the metrics. In fact, we can see that in most cases, the performance is nowhere near saturated, and collecting more ground truths will allow metrics to better capture the semantic variance of a scene. The standout among the group is CIDEr, in which the score does not increase as we increase the number of ground truths. This is primarily due to the IDF component of the CIDEr score, which penalizes increasing the number of tokens harshly. We can see that here, as we increase the number of ground truths, the CIDEr score *decreases*! This suggests that CIDEr is relatively robust to adding more ground truth, however cannot capture as much semantic variance as the other metrics, as the CIDEr score does not materially account for new information from the ground truth samples.

A.2.7 Concept-Diversity: Captions Required for BLEU Score

One of the key experiments we perform is designed to measure the minimum number of captions from the training set that are required to "solve" the test set of the dataset for a particular BLEU score. We first compute a set of all hypothesis descriptions from the training set. Then, for each sample in the test set, we compute the BLEU@4 score using that hypothesis for every sample in the test set. In the case of large datasets such as MS-COCO, which contains $591,435$ unique hypothesis captions, this can be time-consuming, even for the (relatively quick) BLEU@4 metrics. Each hypothesis thus has a score for each sample in the test dataset. Finding the minimal core-set of captions that covers this test dataset

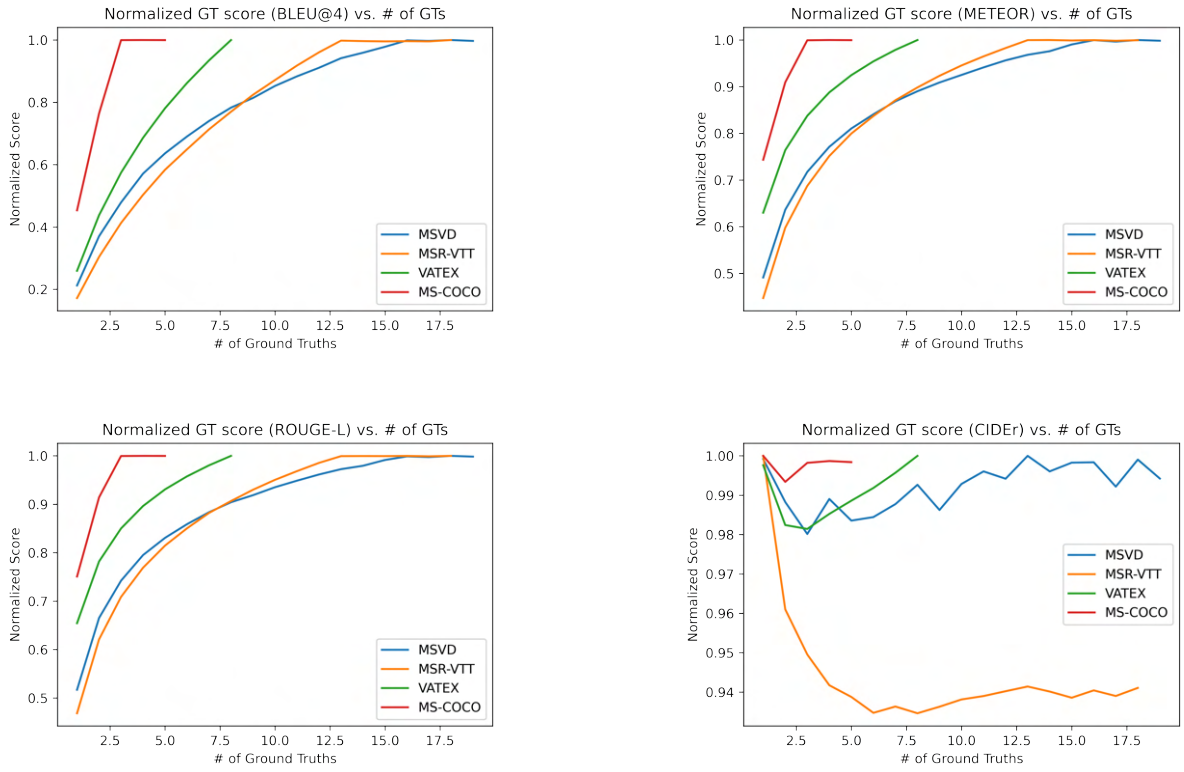


Figure A.3: Performance of different metrics with respect to the number of ground truths considered in leave-one-out experiments. Raw scores are normalized to a maximum of 1, so we can compare the different datasets on the same plot.

to a specified BLEU threshold is a weighted set-cover problem, which can be solved to an $O(\log N)$ approximation with a randomized rounding algorithm (Vazirani, 2001), however, we found that it was sufficient to use the greedy approximation algorithm for set cover, which selects the caption which covers the largest number of new samples at each iteration. Thus, the results in Figure 3.5 provide an upper bound on the possible number of captions required.

Figure 3.5 plots the required number of captions to achieve a BLEU@4 score of X (the value on the X-axis) on every sample. Note that this requirement is *more restrictive* than the plotted SOTA scores, which achieve a mean of X. Thus, the effect of this figure may be even more dramatic than is pictured. The reason for this discrepancy is we compute the core-set using a greedy set cover, and due to our implementation details, it is difficult to terminate the cover efficiently when a mean score is reached.

While our work only computes the core-set for BLEU@4, we believe it would be interesting to see the numbers for other metrics, however, with current implementations, it may be intractable, as the computations require a full pairwise computation of the metrics between

the hypotheses and the test-set samples. Additionally, metrics such as CIDEr which have dataset-wide effects would have to be estimated, requiring several hundred iterations of this experiment to achieve high-quality estimates of the performance. It thus remains interesting (and important) future work to explore how many captions are required to perform well on any given dataset for other metrics.

A.2.8 Concept-Diversity: Feature Sets

To measure the diversity of the datasets at a concept level, we look at how the ground truth captions overlap with the label sets from common feature extractors. If we find that this overlap is high, it suggests that features may have the ability to bias the model along the classification lines of the feature-extractor label set (since a lot of the time, the information extracted by the features is useful primarily for segmenting data along feature class boundaries).

A.2.8.1 Computing Label-Set overlap

We discuss two methods for computing label-set overlap in chapter 3: exact match and fuzzy matching. Exact match is implemented as a string substring: i.e. does the label string appear as a direct substring of the caption. This method provides a lower bound on the true conceptual overlap, as it does not account for misspellings (which are surprisingly common in datasets such as MSR-VTT, and others collected using AMT without additional review steps), and other close matches. While this is a lower bound, it has the benefit of not introducing false-positive matches (as any match is guaranteed to be label overlap). We also discuss the use of fuzzy matching, which we implement using the fuzzywuzzy[4] library for approximate string matching with a threshold of 90. This library uses Levenshtein distance to compute approximate matching, however introduces false-positives which makes it difficult to analyze the overlap. In all cases, the numbers in Table 3.3 represent the percentage of samples that have at least one reference description that has exact overlap with a label from the dataset.

We explore overlap on four common datasets for feature extraction:

ImageNet-1K (Deng et al., 2009) is a popular image classification dataset consisting of 1K labels for object classification ranging across a very wide variety of objects. We can see this from the overlap scores in Table 3.3, which are relatively high on almost all of the datasets. MSR-VTT is relatively low, suggesting that it is one of the most open-domain datasets among the datasets we explore.

[4] `https://github.com/seatgeek/thefuzz`

Kinetics-600 (Carreira et al., 2018) is a popular dataset for action recognition, which contains 600 activities. We can see that the video datasets have a much higher overlap with kinetics, but even though MS-COCO is an image dataset only, there is still some overlap, suggesting that captions of static data still contain human inferences about motion and activity.

MS-COCO (Lin et al., 2014) is a dataset for object detection (and also for visual description) containing object-detection labels over 80 object classes from everyday life. Even though COCO has a relatively restricted object set, we can see that it consists of a set of very popular objects, as the overlap is more than 50% for all captions. Additionally, it's interesting that the object labels for MS-COCO don't always appear in the captions themselves (as the self-overlap is only 92%).

Places-365 (Zhou et al., 2017) is a dataset for scene recognition, consisting of 365 labels of scenes or settings for an image. We find empirically that the overlap for places is likely low, not due to a lack of descriptions of setting, but rather a lack of wide coverage of the variance of settings in Places.

A.2.8.2 Feature-Set Core-Sets and BLEU@4 performance

To directly measure how transferable descriptions are along feature-extractor label axes, we explore the leave-one-out performance of captions sharing the same feature label, but from different samples in the dataset. The results of this experiment using BLEU@4 scores are given in Table 3.4. In order to compute the leave-one-out performance, we begin by computing a set of reference captions R_c for each label in each feature-extractor label set, drawing from the training dataset. These concept-level reference sets consist of all captions containing that label as an exact sub-string. Then, for each sample S_i with references R_i, we compute the set of all concepts overlapping that sample's references C_i. We then compute the hypothesis set for sample S_i as

$$H_i = \left[\bigcup_{c \in C_i} R_c \right] / \{R_i\}$$

Next, for each hypothesis in H_i, we compute the BLEU@4 score for that hypothesis using ground truths R_i. Table 3.4 reports the mean over all samples of the maximum across H_i for each sample in the test set. The results of this metric are clear - when you use the best caption from another sample along feature boundaries, then these captions are relatively transferable (and almost always outperform samples from even the same sample).

A.2.9 Tools & Hardware

The experiments in chapter 3 are computed using the metric implementations provided by the MSCOCO evaluation toolkit in order to compute numeric metric values that are comparable with state of the art methods. In the experiments in chapter 3, we use the Stanford PTB[5] tokenizer provided as part of the toolkit for tokenization and standardization. Unfortunately, because the MSCOCO toolkit does not explicitly specify a tokenization scheme and most works in video description do not subscribe to a standard tokenization tool, we are unable to be certain that the metric is consistent between our work, and the work presented in the state of the art papers.

The experiments are run in parallel on a machine with 96 AMD EPYC 7B12 cores and 378 GB of RAM running on Google Cloud Platform. Notably, the caption concept-overlap experiments require a very large amount of compute, with this machine requiring almost 10 hours to compute the BLEU score for the core-set concept overlap. We found scores such as METEOR (Agarwal and Lavie, 2008) and SPICE (Anderson et al., 2016) to be computationally prohibitive (requiring several months of sustained compute) for some of these experiments, thus, we do not include those scores in this work. We also do not report several modern metrics for this reason - as a major downside to many of the automated metrics that have recently been developed is their forward inference speed (up to 1000s of times slower than the computation of the BLEU score). A key area of future work is improving the computational performance of metrics, as this will allow such metrics to not only be used for more detailed analysis but will allow such metrics to be optimized directly using techniques such as self-critical sequence training (Rennie et al., 2017).

[5] https://nlp.stanford.edu/software/tokenizer.html

A.3 Additional Qualitative Examples

Additional qualitative examples are selected at random from the datasets using a random number generator over the length of each dataset. Some randomly selected samples are omitted due to explicit content in the visual data or descriptions (which is an additional cause for concern, but out of scope of the current research).

- a person mixing potatoe salad in a bowl (Dist: 0.10) (BLEU@4: 0.0000)
- someone mixes a potato salad (Dist: 0.13) (BLEU@4: 0.0001)
- this is someone stirring a bowl of potato salad (Dist: 0.13) (BLEU@4: 0.3156)
- a person in a kitchen is mixing a bowl of potato salad with a spatula in a white bowl on a cutting board (Dist: 0.14) (BLEU@4: 0.1613)
- the bowl of potato salad is stirred (Dist: 0.18) (BLEU@4: 0.3267)
- someone is mashing potato salad together (Dist: 0.19) (BLEU@4: 0.0000)
- someone is making food (Dist: 0.22) (BLEU@4: 0.0000)
- a women stirring and mixing some egg salad (Dist: 0.23) (BLEU@4: 0.0000)
- a person is cooking (Dist: 0.24) (BLEU@4: 0.0000)
- a bowl is being stirred (Dist: 0.26) (BLEU@4: 0.0000)
- a person stirring food (Dist: 0.26) (BLEU@4: 0.0000)
- someone mixing pieces of boiled eggs and yogurt in a bowl (Dist: 0.27) (BLEU@4: 0.0000)
- mixing up food in a bowl in the kitchen (Dist: 0.30) (BLEU@4: 0.0000)
- making some potato salad (Dist: 0.30) (BLEU@4: 0.0000)
- a great way to make potato salad (Dist: 0.36) (BLEU@4: 0.0000)
- how to prepare a potato salad (Dist: 0.36) (BLEU@4: 0.0001)
- preparation of some egg dish (Dist: 0.44) (BLEU@4: 0.0000)

msrvtt+train+video481
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.10,0.24,0.44]
Mean Leave One Out BLEU@4 score: 0.0473

Figure A.4: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- baby muppets are talking to a judge kermit (Dist: 0.18) (BLEU@4: 0.3457)
- cartoon muppets are having a court proceeding (Dist: 0.18) (BLEU@4: 0.0000)
- kermit is a judge and is talking to 2 other muppet babies (Dist: 0.20) (BLEU@4: 0.1499)
- some muppet characters are interacting with each other (Dist: 0.22) (BLEU@4: 0.0000)
- a scene from the muppet babies where baby kermit is a judge in a court room is playing in a movie theater (Dist: 0.24) (BLEU@4: 0.1336)
- a cartoon clip of baby muffets in a court room (Dist: 0.24) (BLEU@4: 0.0876)
- muppets are talking to each other (Dist: 0.25) (BLEU@4: 0.3641)
- the scene is from baby muppets the green frog character kermit is a judge and two other characters are standing before him (Dist: 0.26) (BLEU@4: 0.1759)
- muppet babies playing in a theater (Dist: 0.29) (BLEU@4: 0.0000)
- kermit judges gonzo and another muppet baby (Dist: 0.31) (BLEU@4: 0.0000)
- some cartoon characters speak to a judge (Dist: 0.32) (BLEU@4: 0.0000)
- cartoon babies standing in front of a judge (Dist: 0.34) (BLEU@4: 0.0000)
- the muppets are being shown in a theater (Dist: 0.35) (BLEU@4: 0.0000)
- the muppet babies are on a theater screen (Dist: 0.36) (BLEU@4: 0.0000)
- cartoon characters are talking in a tv show (Dist: 0.38) (BLEU@4: 0.0000)
- a cartoon frog is talking to other cartoon characters (Dist: 0.41) (BLEU@4: 0.0000)
- a clip taken from the cartoon the muphets (Dist: 0.42) (BLEU@4: 0.0000)
- cartoons are being displayed (Dist: 0.43) (BLEU@4: 0.0000)

msrvtt+train+video1902
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.18,0.30,0.43]
Mean Leave One Out BLEU@4 score: 0.0698

Figure A.5: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- a man is playing guitar and singing (Dist: 0.07) (BLEU@4: 0.4468)
- a man playing the guitar and singing (Dist: 0.07) (BLEU@4: 0.5314)
- a man is playing the guitar and singing outside (Dist: 0.09) (BLEU@4: 0.5874)
- a man is singing while playing his guitar (Dist: 0.09) (BLEU@4: 0.7071)
- a man is singing while playing the guitar (Dist: 0.09) (BLEU@4: 0.7071)
- a man playing a guitar in various locations singing a song (Dist: 0.13) (BLEU@4: 0.0000)
- a man plays a guitar and sings a song outside (Dist: 0.13) (BLEU@4: 0.4111)
- a man sings while playing the acoustic guitar (Dist: 0.16) (BLEU@4: 0.4214)
- a man is singing on a doorstep and playing his guitar (Dist: 0.18) (BLEU@4: 0.3393)
- a man singing a song (Dist: 0.19) (BLEU@4: 0.0000)
- man is playing a guitar (Dist: 0.22) (BLEU@4: 0.0001)
- a man sings while playing a guitar throughout a town (Dist: 0.22) (BLEU@4: 0.4111)
- a man is playing the guitar (Dist: 0.23) (BLEU@4: 0.6065)
- a man plays a guitar outside (Dist: 0.25) (BLEU@4: 0.4083)
- this is a video of a man playing guitar and singing (Dist: 0.25) (BLEU@4: 0.2900)
- a man is singing a melodious song (Dist: 0.29) (BLEU@4: 0.3564)
- guy performing music on building (Dist: 0.40) (BLEU@4: 0.0000)

msrvtt+train+video5073
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.18, 0.40]
Mean Leave One Out BLEU@4 score: 0.3661

Figure A.6: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- a football coach talking to his players on the sideline (Dist: 0.16) (BLEU@4: 0.0001)
- a person talking to football players (Dist: 0.18) (BLEU@4: 0.0000)
- a coach is talking to his team of football players (Dist: 0.19) (BLEU@4: 0.0000)
- a football coach talks on the sideline (Dist: 0.23) (BLEU@4: 0.0001)
- a football coach watches the players (Dist: 0.26) (BLEU@4: 0.0001)
- football players are on the field (Dist: 0.27) (BLEU@4: 0.0001)
- football players on the field (Dist: 0.28) (BLEU@4: 0.0001)
- a coach is speaking with his team (Dist: 0.29) (BLEU@4: 0.0000)
- a man on the sidelines at a football game is talking continuously (Dist: 0.29) (BLEU@4: 0.0000)
- a football coach is shown losing his patience with his players (Dist: 0.36) (BLEU@4: 0.0000)
- a football player is on the sidelines (Dist: 0.37) (BLEU@4: 0.0000)
- people are playing sports (Dist: 0.38) (BLEU@4: 0.0000)
- footage of a football game (Dist: 0.39) (BLEU@4: 0.0000)
- there are men at a sports match (Dist: 0.44) (BLEU@4: 0.0000)
- players playing soccer game (Dist: 0.49) (BLEU@4: 0.0000)
- a man is chewing and watching football players (Dist: 0.50) (BLEU@4: 0.0000)
- a trainer guide a player (Dist: 0.52) (BLEU@4: 0.0000)
- a man is talking (Dist: 0.54) (BLEU@4: 0.0000)
- some guys are geting ready to play rugby in a ground (Dist: 0.57) (BLEU@4: 0.0000)
- a man talking or yelling with his lid backward (Dist: 0.71) (BLEU@4: 0.0000)

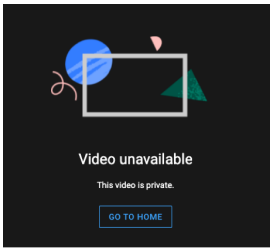
msrvtt+train+video63
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.16, 0.37, 0.71]
Mean Leave One Out BLEU@4 score: 0.0000

Figure A.7: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- While outside, a man is using a chainsaw to cut through a wood log. (Dist: 0.07) (BLEU@4: 0.2804)
- A man cuts into a large log with a chainsaw. (Dist: 0.10) (BLEU@4: 0.0000)
- A man using a chain saw to cut a board out of a log. (Dist: 0.11) (BLEU@4: 0.4547)
- A man uses a chainsaw to cut a log into smaller segments. (Dist: 0.12) (BLEU@4: 0.3787)
- A man holds a chainsaw and carves a log with the saw. (Dist: 0.14) (BLEU@4: 0.0000)
- A man is outside using a chainsaw to shape a large tree trunk. (Dist: 0.15) (BLEU@4: 0.4002)
- A man is using a chain saw to cut a large tree trunk. (Dist: 0.17) (BLEU@4: 0.4575)
- A man uses a chainsaw to carve a fallen tree in the woods. (Dist: 0.19) (BLEU@4: 0.3760)
- A man is using an electric saw to cut a block of wood. (Dist: 0.22) (BLEU@4: 0.3662)
- A man is taking a video showing us how to cut a good lumber. (Dist: 0.37) (BLEU@4: 0.0000)

vatex+train+c7R_RDH6s_A_000122_000132
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.16, 0.37]
Mean Leave One Out BLEU@4 score: 0.2714

Figure A.8: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).



- A man demonstrates a lunge with a barbell with weights on the end. (Dist: 0.14) (BLEU@4: 0.0000)
- A man is using barbells to teach how to do a lunge with them. (Dist: 0.16) (BLEU@4: 0.0000)
- A man is in the gym performing a barbell single side lunge. (Dist: 0.18) (BLEU@4: 0.1623)
- a man in a gym is demonstrating how to properly do an exercise (Dist: 0.18) (BLEU@4: 0.0000)
- A person demonstrates the proper technique for a barbell exercise while text explains the technique. (Dist: 0.19) (BLEU@4: 0.0000)
- A man place a set of barbells on his shoulder and steps forward. (Dist: 0.21) (BLEU@4: 0.0000)
- The guy is in the gym lifting some barbel on the back of his shoulders. (Dist: 0.25) (BLEU@4: 0.1956)
- A man is holding a barbell on his shoulders as he steps forward. (Dist: 0.26) (BLEU@4: 0.0000)
- A person doing exercises and telling how to do them the right way. (Dist: 0.29) (BLEU@4: 0.0000)
- A man is standing in the gym lifting weights from the ground to his shoulders (Dist: 0.36) (BLEU@4: 0.1967)

vatex+train+j53UDMQ8mxo_000000_000010
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.14, 0.22, 0.36]
Mean Leave One Out BLEU@4 score: 0.0555

Figure A.9: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4). The visual content of this video is missing (as the video has become private since the collection of the dataset), however we include the video as it is one of the randomly sampled instances.

- A man is pouring tea from a tea kettle into two mugs while keeping his other hand in his pocket. (Dist: 0.17) (BLEU@4: 0.1573)
- Man places hand in pocket before pouring water from kettle into two mugs. (Dist: 0.18) (BLEU@4: 0.1492)
- A man in a kitchen puts his hand in his pocket while pouring a beverage and explains why. (Dist: 0.23) (BLEU@4: 0.1579)
- The man is pouring something from the kettle into the two coffee mugs on the counter. (Dist: 0.26) (BLEU@4: 0.0000)
- He places his hand in his pant pocket then pours coffee into the cups. (Dist: 0.29) (BLEU@4: 0.1395)
- A man wearing a gray tank top is pouring tea into a cup. (Dist: 0.29) (BLEU@4: 0.1880)
- A man wearing a tee shirt pours coffee in two cups. (Dist: 0.29) (BLEU@4: 0.1898)
- A man in his kitchen has a kettle pouring himself hot water (Dist: 0.30) (BLEU@4: 0.0000)
- There is a man making tea and he is explaining where to put your arms so you don't burn them while pouring the water. (Dist: 0.31) (BLEU@4: 0.0000)
- A guy is talking about a cooking video that he saw a couple weeks ago. (Dist: 0.53) (BLEU@4: 0.0000)

vatex+train+DbJWd2K2Hw0_000229_000239
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.17, 0.29, 0.53]
Mean Leave One Out BLEU@4 score: 0.0982

Figure A.10: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- A person peels and slices an apple using a knife onto a plate. (Dist: 0.07) (BLEU@4: 0.3161)
- A person peels an apple and chops it to tiny pieces. (Dist: 0.09) (BLEU@4: 0.0000)
- A person is cutting up an apple with a knife onto a green plate. (Dist: 0.11) (BLEU@4: 0.2460)
- A person is peeling apples and then cutting out the cores. (Dist: 0.11) (BLEU@4: 0.0000)
- A person is cutting a peeled and cored red apple into slices. (Dist: 0.12) (BLEU@4: 0.1907)
- A person is taking the skin off of some apples using a small knife. (Dist: 0.12) (BLEU@4: 0.0000)
- A person peeling an apple using a knife while sitting at a table (Dist: 0.13) (BLEU@4: 0.3184)
- A man uses a knife to peel of the side of apple. (Dist: 0.13) (BLEU@4: 0.0000)
- A person inside sitting at a table peels an apple on a green plate. (Dist: 0.18) (BLEU@4: 0.2289)
- A person cuts an onion into pieces using a knife. (Dist: 0.30) (BLEU@4: 0.0000)

vatex+train+W3orrcZAb2w_000200_000210
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.14, 0.30]
Mean Leave One Out BLEU@4 score: 0.1300

Figure A.11: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- On a beach three men are doing cartwheels after another (Dist: 0.07) (BLEU@4: 0.1949)
- Three young men are doing cartwheels across the beach, towards the surf. (Dist: 0.09) (BLEU@4: 0.1967)
- Three males are doing cartwheels down the beach towards the ocean. (Dist: 0.09) (BLEU@4: 0.4125)
- Three guys do cartwheels on a beach toward the water until the stop and fall. (Dist: 0.11) (BLEU@4: 0.0000)
- Three men wearing shorts are doing cartwheels on the sand on the beach. (Dist: 0.12) (BLEU@4: 0.1967)
- Group of men doing numerous cartwheels down the beach towards the water. (Dist: 0.13) (BLEU@4: 0.4090)
- Three kids do cartwheels down the beach together. (Dist: 0.17) (BLEU@4: 0.2367)
- Three men doing cart wheels across a beach, stop then laughing. (Dist: 0.20) (BLEU@4: 0.0000)
- A group of people are successively doing cart wheels on the beach. (Dist: 0.23) (BLEU@4: 0.1982)
- Three boys are doing cart wheeling on the beach all by themselves. (Dist: 0.33) (BLEU@4: 0.0000)

vatex+train+qyHurkZFOp0_000002_000012
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.07, 0.15, 0.33]
Mean Leave One Out BLEU@4 score: 0.1845

Figure A.12: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- A man is using specialized ice climbing gear and metal cleats on his shoes to climb an icy cliff. (Dist: 0.10) (BLEU@4: 0.1532)
- A man using proper equipment and shoe's to ice climb. (Dist: 0.10) (BLEU@4: 0.0000)
- A man is climbing up the side of an icy mountains using climbing shoes and tools. (Dist: 0.11) (BLEU@4: 0.4925)
- A man is wearing shoes with spikes and using a ice pick to climb a wall of ice. (Dist: 0.12) (BLEU@4: 0.3709)
- A man is attached to a harness as he is climbing a wall of snow and ice. (Dist: 0.13) (BLEU@4: 0.0000)
- A man is climbing up the side of an ice wall. (Dist: 0.14) (BLEU@4: 0.4808)
- A man wearing a harness and spiked shoes climbs up a snow covered wall. (Dist: 0.14) (BLEU@4: 0.0000)
- A man uses a pick and a harness to scale an icy cliff. (Dist: 0.18) (BLEU@4: 0.1480)
- a person tries their best to climb up a snowy mountain (Dist: 0.30) (BLEU@4: 0.0000)
- A man cheers as another man is using crampons to climb a wall of ice. (Dist: 0.30) (BLEU@4: 0.3714)

vatex+train+keBAoE5tC44_000011_000021
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.10, 0.16, 0.30]
Mean Leave One Out BLEU@4 score: 0.2017

Figure A.13: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

msvd+train+_WRC7HXBJpU_360_370
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.09, 0.30, 0.74]
Mean Leave One Out BLEU@4 score: 0.2693

- someone is melting butter in a pan (Dist: 0.09) (BLEU@4: 0.7014)
- butter is being melted in a pan (Dist: 0.09) (BLEU@4: 0.4889)
- the butter is melting in the pan (Dist: 0.10) (BLEU@4: 0.5329)
- a man melts a piece of butter in a pan (Dist: 0.10) (BLEU@4: 0.5170)
- a person melts butter on a pan (Dist: 0.10) (BLEU@4: 0.0000)
- butter is melted in a pan (Dist: 0.11) (BLEU@4: 0.5115)
- a man is heating butter in a pan (Dist: 0.12) (BLEU@4: 0.5000)
- a man is melting butter in a pan (Dist: 0.12) (BLEU@4: 0.8409)
- someone is melting a piece of butter in a skillet (Dist: 0.13) (BLEU@4: 0.5170)
- butter is melting in a pan (Dist: 0.14) (BLEU@4: 0.6732)
- a man is stirring melting butter in a pan (Dist: 0.14) (BLEU@4: 0.5969)
- butter melts in a pan (Dist: 0.14) (BLEU@4: 0.0001)
- butter is melting in a frying pan (Dist: 0.15) (BLEU@4: 0.6435)
- melting the butter on the frying pan (Dist: 0.15) (BLEU@4: 0.0001)
- some butter is melting in a hot skillet (Dist: 0.16) (BLEU@4: 0.5170)
- butter melted in the pan (Dist: 0.16) (BLEU@4: 0.8187)
- the butter melted in the pan (Dist: 0.16) (BLEU@4: 0.7598)
- a man is melting butter in a skillet (Dist: 0.17) (BLEU@4: 0.8409)
- butter is melting on a skillet (Dist: 0.17) (BLEU@4: 0.0001)
- a cook spreads melted butter around a pan (Dist: 0.18) (BLEU@4: 0.0000)
- a man is spreading some butter in a pan (Dist: 0.19) (BLEU@4: 0.4317)
- the man mmelted butter in the frying pan (Dist: 0.19) (BLEU@4: 0.0001)
- a man demonstrates how to prepare a pan with butter (Dist: 0.21) (BLEU@4: 0.0000)
- the man is meliting butter in the pan (Dist: 0.22) (BLEU@4: 0.0001)
- a man is pressing a butter with mixing utensil in a black pan (Dist: 0.26) (BLEU@4: 0.0000)
- somebody is cooking (Dist: 0.36) (BLEU@4: 0.0000)
- a man is cooking (Dist: 0.39) (BLEU@4: 0.0001)
- a cheese cub is melting in the pan (Dist: 0.39) (BLEU@4: 0.5170)
- he is cooking on the plate (Dist: 0.48) (BLEU@4: 0.0000)
- a man is melting the cheese (Dist: 0.50) (BLEU@4: 0.3641)
- a man coking pork chops (Dist: 0.52) (BLEU@4: 0.0000)
- a man makes food for him self (Dist: 0.54) (BLEU@4: 0.0000)
- cheese in the oil (Dist: 0.55) (BLEU@4: 0.0000)
- a man cooking his kichen (Dist: 0.58) (BLEU@4: 0.0000)
- a person is making pork chops (Dist: 0.59) (BLEU@4: 0.0000)
- cooking the pork chops (Dist: 0.60) (BLEU@4: 0.0000)
- anyone is frying (Dist: 0.61) (BLEU@4: 0.0000)
- a man in a blue shirt and yellow cap preparing pork chops (Dist: 0.67) (BLEU@4: 0.0000)
- clear cooking details for pork chops (Dist: 0.70) (BLEU@4: 0.0001)
- racipe for pork chops (Dist: 0.74) (BLEU@4: 0.0001)

Figure A.14: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- a woman is hugging a little boy (Dist: 0.06) (BLEU@4: 0.6435)
- a woman is hugging a boy (Dist: 0.08) (BLEU@4: 0.7598)
- a woman hugs a little boy (Dist: 0.08) (BLEU@4: 0.6732)
- a woman is hugging a child (Dist: 0.10) (BLEU@4: 0.7598)
- a young woman hugs a little boy (Dist: 0.11) (BLEU@4: 0.6435)
- the woman is hugging the boy (Dist: 0.11) (BLEU@4: 0.0001)
- a woman is hugging her child (Dist: 0.12) (BLEU@4: 0.7598)
- the lady hugged her little boy close to her (Dist: 0.16) (BLEU@4: 0.0000)
- a woman is hugging her son (Dist: 0.18) (BLEU@4: 0.7598)
- a woman is hugging a toddler (Dist: 0.21) (BLEU@4: 0.7598)
- a girl hugs a little boy (Dist: 0.23) (BLEU@4: 0.5373)
- a girls is hugging a little boy (Dist: 0.23) (BLEU@4: 0.6435)
- a woman hugs a small boy then sets him down on the floor (Dist: 0.25) (BLEU@4: 0.2045)
- the mother is hugging the baby (Dist: 0.28) (BLEU@4: 0.5373)
- a teenage girl is hugging a little boy (Dist: 0.29) (BLEU@4: 0.5411)
- two people are hugging (Dist: 0.33) (BLEU@4: 0.0000)
- the girl hugged the little boy tightly (Dist: 0.34) (BLEU@4: 0.0000)
- the girl is hugging the baby (Dist: 0.44) (BLEU@4: 0.5373)
- whitney is hugging her son in her house (Dist: 0.44) (BLEU@4: 0.3457)
- a lady is playing with child (Dist: 0.47) (BLEU@4: 0.0000)
- one lady and child are standing together (Dist: 0.50) (BLEU@4: 0.0000)
- the baby is doing the something (Dist: 0.75) (BLEU@4: 0.0000)

msvd+test+pzq5fPfsPZg_51_57
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.06, 0.26, 0.75]
Mean Leave One Out BLEU@4 score: 0.4139

Figure A.15: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

- Young boy leaping into air with tennis racket with dog below him. (Dist: 0.05) (BLEU@4: 0.0000)
- A boy jumping into the air with a tennis racket. (Dist: 0.10) (BLEU@4: 0.3352)
- A boy with a tennis racket jumping in the air. (Dist: 0.10) (BLEU@4: 0.3352)
- A young man with a tennis racket jumping high into the air near a dog on a street side. (Dist: 0.19) (BLEU@4: 0.2195)
- A kid is in the air with a dog looking on (Dist: 0.32) (BLEU@4: 0.2481)

mscoco+train+174229
Within-Sample mean BERT Embedding distances [Min, Mean, Max]: [0.05, 0.15, 0.32]
Mean Leave One Out BLEU@4 score: 0.2276

Figure A.16: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

Figure A.17: Qualitative example of metrics presented in chapter 3. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 3.4).

Appendix B

Appendix for IC3: Image Captioning by Committee Consensus

Appendix B.1 describes the code release, and links to available released resources associated with the chapter.

Appendix B.2 describes several explorations of the hyperparameters, including the language model, number of candidate samples, and prompts.

Appendix B.3 describes additional expeirmental details including the image captioning models and datasets.

Appendix B.4 describes an ELO scoring system which we use in some additional appendix experiments.

Appendix B.5 describes the human studies, and analysis of the human studies in detail.

Appendix B.6 gives additional qualitative results for the method.

Appendix B.7 explores how IC^3 can be used in zero-shot style and language transfer situations.

Appendix B.8 describes some additional failure modes and limitations of IC^3 in detail.

B.1 Code Release

Our code is available at `https://github.com/davidmchan/caption-by-committee`, and is made publicly available on Github with an MIT license, and contains the implementation, as well as the validation results for each of the models, the evaluation server/framework, and other necessary artifacts, to encourage further research in the domain of diverse/summarized image captioning.

B.2 Hyperparameter Exploration

In this section, we provide additional experimental details regarding the choice of the hyperparameters for our method discussed in section 6.2.

B.2.1 Choice of Summarization Model

The choice of the summarization model $\mathcal{S}$ is a key decision when implementing IC^3. Table B.1 demonstrates the performance of IC^3 with several models, both using prompting for large language models, and using summarization of the captions directly. Generally we found that models from OpenAI (Such as GPT-3 and GPT-4) were strong performers, however models from Anthropic (such as CLAUDE), have strong summarization performance as well. The strongest open-source models are Koala and Vicuna, both Chat-style models, with LLama and StableLM following. While Table B.1 seems to imply that T-5 is a strong model (and it likely is in terms of content-coverage in recall), T-5 often copy-pastes several of the candidate sentences instead of generating a strong abstractive summary, leading to decreased fluency.

B.2.2 Prompts

In this section, we present several explorations of possible prompts. First, Table B.2, we present an exploration of the prompt with and without language which encourages the model to produce uncertainty-specific language (the green text in subsection 6.2.3). To evaluate this, we use two approaches: a head-to-head experiment where captions generated by the two prompts are evaluated directly by human raters for helpfulness and correctness (following subsection 6.2.4), and an automated measure of "likely-language occurrence", LLOP. To compute LLOP, we compute the number of captions containing words that indicate some uncertainty including "likely", "probably", "possibly" and others. We find that without explicitly encouraging the model to produce uncertain language, the model seldom does so, while doing so improves both the helpfulness and correctness when measured by human raters.

In the second exploration in Table B.3, we explore the question of using a prefix similar to one explored in Kojima et al. (2022). We find that while the prefix does help, it is not a key component of our method, and increases automated measures by small, but perceivable, levels.

B.2.3 Choosing the number of candidate samples

Choosing the number of captions to summarize is highly dependent on both the abilities of the language model, and the tolerance for execution cost of the method. Adding more captions can increase the amount of information discovered by the visual model, and generate better representative samples of the input data distribution, however it can increase the context

Table B.1: Exploration of the choice of language model, when holding the prompt and candidate captions stable, using BLIP on a 200-element randomly sampled subset of the MS-COCO dataset.

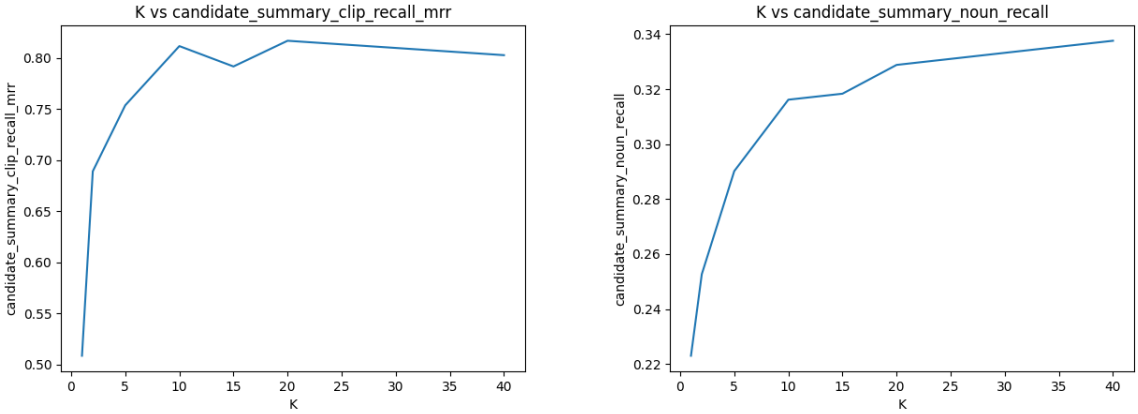
Model	Exact Noun	Exact Verb	Fuzzy Noun	Fuzzy Verb	CLIP Recall MRR	CLIP Recall R@1	CLIP Recall R@5	CLIP Recall R@10
Language Models								
IC^3 + bloom (Scao et al., 2022)	0.248	0.16	0.551	0.402	0.834	0.725	0.98	0.995
IC^3 + distilgpt2 (Sanh et al., 2019)	0.208	0.146	0.517	0.488	0.643	0.535	0.795	0.825
IC^3 + gpt2 (Radford et al., 2019)	0.272	0.159	0.602	0.542	0.638	0.51	0.79	0.83
IC^3 + gpt2 lg (Radford et al., 2019)	0.28	0.164	0.583	0.486	0.735	0.64	0.85	0.89
IC^3 + gpt2 med (Radford et al., 2019)	0.299	0.187	0.606	0.531	0.79	0.705	0.9	0.935
IC^3 + gpt2 xl (Radford et al., 2019)	0.28	0.18	0.58	0.473	0.849	0.755	0.975	0.99
IC^3 + gpt3 (Brown et al., 2020)								
+ ada	0.282	0.18	0.585	0.463	0.866	0.78	0.975	0.985
+ babbage	0.199	0.115	0.504	0.341	0.83	0.735	0.97	**1.0**
+ curie	0.218	0.111	0.519	0.319	0.827	0.71	0.975	0.995
+ davinci2	0.321	0.207	0.622	0.491	0.939	0.895	**1.0**	**1.0**
+ davinci3	0.381	0.251	0.675	0.547	0.958	**0.925**	**1.0**	**1.0**
IC^3 + gptneo 125M (Black et al., 2021)	0.235	0.157	0.521	0.447	0.777	0.69	0.895	0.915
IC^3 + gptneo 1B (Black et al., 2021)	0.253	0.155	0.546	0.403	0.844	0.75	0.985	0.995
IC^3 + gptneo 2B (Black et al., 2021)	0.242	0.15	0.536	0.393	0.844	0.74	0.98	**1.0**
IC^3 + llama7B (Touvron et al., 2023)	0.224	0.128	0.517	0.324	0.777	0.65	0.945	0.995
IC^3 + llama13B (Touvron et al., 2023)	0.257	0.175	0.554	0.419	0.834	0.725	0.99	**1.0**
IC^3 + stable lm 3B (StabilityAI, 2023)	0.247	0.184	0.552	0.454	0.873	0.785	0.985	0.995
Chat Models								
IC^3 + alpaca 7B (Taori et al., 2023)	0.324	0.216	0.63	0.503	0.912	0.85	**1.0**	**1.0**
IC^3 + chatgpt (OpenAI, 2022b)	0.401	0.27	0.692	0.595	0.954	0.920	**1.0**	**1.0**
IC^3 + claude (Bai et al., 2022)	0.38	0.262	0.677	0.583	**0.962**	**0.935**	**1.0**	**1.0**
IC^3 + gpt4 (OpenAI, 2023)	**0.42**	0.29	**0.713**	0.609	0.96	0.925	**1.0**	**1.0**
IC^3 + koala 7B (Geng et al., 2023)	0.284	0.178	0.586	0.455	0.899	0.825	0.99	**1.0**
IC^3 + koala 13B v1 (Geng et al., 2023)	0.418	**0.323**	0.692	**0.637**	0.916	0.865	0.985	0.985
IC^3 + koala 13B v2 (Geng et al., 2023)	0.376	0.264	0.67	0.592	0.923	0.87	0.995	0.995
IC^3 + stable lm 3B (StabilityAI, 2023)	0.077	0.06	0.299	0.144	0.265	0.23	0.28	0.295
IC^3 + stable lm 7B (StabilityAI, 2023)	0.263	0.197	0.564	0.489	0.873	0.785	0.995	**1.0**
IC^3 + vicuna 7B (Chiang et al., 2023)	0.361	0.247	0.658	0.548	0.938	0.89	**1.0**	**1.0**
IC^3 + vicuna 13B (Chiang et al., 2023)	0.384	0.272	0.676	0.584	0.927	0.87	0.995	0.995
Summary Models								
IC^3 + t5 base (Raffel et al., 2020)	0.353	0.25	0.646	0.587	0.903	0.845	0.98	0.99
IC^3 + t5 small (Raffel et al., 2020)	0.402	0.289	0.681	0.609	0.944	0.9	0.995	**1.0**
Baselines								
IC^3 + references	0.434	0.305	0.684	0.564	0.939	0.895	0.995	1.0
blip baseline (Li et al., 2022)	0.266	0.196	0.567	0.491	0.865	0.77	0.995	1.0
Chat Captioner (t5-xxl + chatgpt) (Zhu et al., 2023a)	0.361	0.207	0.669	0.564	0.947	0.905	**1.0**	**1.0**

Table B.2: Exploration of "uncertainty-encouraging" language in the prompt, using BLIP and GPT-3 on a 200 element randomly sampled subset of the MS-COCO dataset. See Appendix B.2.2 for a discussion of LLOP, the "likely-language occurrence percentage". Helpfulness and correctness are given as head-to-head win percentage following subsection 6.2.4.

Model	LLOP	Helpfulness	Correctness
Candidates With	**62.5**%	**52.01**%	**72.32**%
Candidates Without	4.0%	43.63%	18.16%
References With	**52.0**%	**34.65**%	**53.00**%
References Without	0%	28.79%	26.20%

Table B.3: Content coverage and CLIP recall demonstrating the use of "This is a hard problem" in the prompt, using BLIP on a 200 element randomly sampled subset of the MS-COCO dataset.

	Exact		Fuzzy		CLIP Recall			
Model	Noun	Verb	Noun	Verb	MRR	R@1	R@5	R@10
Candidates With	**0.322**	**0.216**	**0.647**	**0.503**	**0.770**	**0.645**	**0.96**	**0.99**
Candidates Without	0.316	0.208	0.638	0.496	0.765	0.635	0.94	**0.99**
References With	0.516	**0.308**	0.744	**0.560**	**0.833**	**0.745**	**0.97**	**0.995**
References Without	**0.518**	0.305	**0.745**	0.558	0.830	**0.745**	**0.97**	0.99



(a) K vs. CLIP MRR

(b) K vs. Noun Recall

Figure B.1: Exploration of the number of candidate set captions vs CLIP MRR and Noun Recall for the GPT-3 language model.

length passed to the large language model, straining the summarization capabilities of the model, and leading to increased cost for the LLM. We ablate the choice in Figure B.1. Here, we can see that increasing the number of captions can lead to increases in automated measure performance (as more captions will capture more information), however more captions can also increase execution time linearly with the number of candidate captions. We can see here that much of the benefit is captured at 10 candidate captions, which we chose for this work, since it represents a good trade-off between execution time, and caption quality.

Is performance just due to LLMs correcting caption errors? Table Table B.4 shows both the automated performance of the GPT-3 model, and the human performance of the GPT-3 model for several values of K. We can see that while GPT can help to improve on single captions through error correction (as evidenced by slightly higher scores for GPT-3 (K=1) vs. the baseline), the best scores are achieved with higher values of K.

B.2.4 BLIP + OFA

Because the caption summarization process is independent of the caption generation process, it is a natural question to ask if multiple different sources of caption generation could be used during the generation phase. The results of combining the sampled candidates from both BLIP and OFA are shown in Table B.5

B.2.5 How is caption diversity related to IC3 outputs?

One reasonable question to ask is: does the diversity of the input captions impact the quality of the output summarized caption? In Figure B.2, we plot the Self-BLEU (Zhu et al.,

Table B.4: Exploration of the choice of K for the GPT-3 language model and the BLIP-2 captioning engine on a randomly sampled 200 element MS-COCO subset. Human results are given as Glicko-2 scores (See Appendix B.4).

	Exact		Fuzzy		CLIP Recall			Human
K	Noun	Verb	Noun	Verb	MRR	R@1	R@5	Glicko
Baseline	0.264	0.162	0.564	0.423	0.885	0.805	0.985	1367.28
1	0.258	0.161	0.562	0.456	0.872	0.790	0.985	1534.48
10	0.346	0.212	0.646	0.516	**0.956**	**0.920**	**1.000**	1674.51
100	**0.368**	**0.223**	**0.665**	**0.526**	0.948	0.905	**1.000**	1623.22

Table B.5: Content coverage and CLIP recall demonstrating the combination of caption engines on a 200 element randomly sampled subset of the MS-COCO dataset.

	Exact		Fuzzy		CLIP Recall			
Model	Noun	Verb	Noun	Verb	MRR	R@1	R@5	R@10
ofa + blip + ic^3	0.341	0.204	0.648	0.485	0.796	0.685	0.945	0.985
refs + ic^3	0.517	0.308	0.744	0.561	0.833	0.745	0.970	0.995
Blip + ic^3	0.313	0.206	0.637	0.493	0.770	0.645	0.960	0.99
Ofa + ic^3	0.300	0.184	0.623	0.474	0.770	0.660	0.935	0.97
Blip	0.230	0.178	0.542	0.439	0.551	0.400	0.760	0.91
Ofa	0.212	0.150	0.531	0.387	0.341	0.115	0.630	0.89
refs	0.214	0.537	0.099	0.337	0.683	0.540	0.877	0.965

2018) of the candidate captions (a measure of caption-set diversity), against the automated evaluation measures. We find that in general, there are very weak correlations between the CLIP MRR and the diversity of the candidate set (OFA, $r = 0.079$, BLIP, $r = 0.094$, BLIP-2, $r = 0.059$): when more diversity is needed to express the content to high specificity, the model is including it. When less diversity is required, the model does not include it. We do however see correlation between the content recall scores of the model, and the diversity of the input candidates (Noun Recall: OFA, $r = 0.233$, BLIP $r = 0.238$, BLIP-2, $r = 0.252$, Verb Recall: OFA, $r = 0.199$, BLIP $r = 0.193$, BLIP-2, $r = 0.185$). This suggests that when the candidates are more diverse, this information is captured in the output summary sentence.

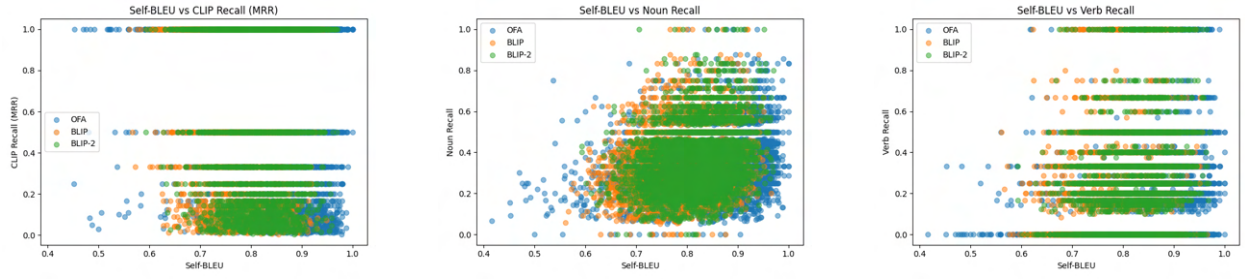


Figure B.2: Plots showing diversity of candidate captions plotted against automated evaluation measures when using 10 candidate captions, and GPT-3 (Davinci v3) as a LM.

B.3 Additional Experimental Details

B.3.1 Image Captioning Methods

In this work, we explore two image captioning models as seed models for IC^3: BLIP (Li et al., 2022) and OFA (Wang et al., 2022b).

BLIP: BLIP (Li et al., 2022) is a vision-language pre-training framework designed to effectively use noisy web data at scale for pre-training. The model operates by using a large dataset of synthetic image-caption pairs, generated by a seed captioning model, and a filter to remove low-quality synthetic captions. BLIP has demonstrated strong transfer performance to many vision-language tasks, and performs particularly well when transferred to image captioning in zero-shot and fine-tuning scenarios. The BLIP (Large) model that we use is fine-tuned on MS-COCO for image captioning, and unless otherwise specified, we generate baseline captions using beam search with 16 beams, and generate candidate captions for IC3 using temperature sampling as described in subsection 6.2.1.

OFA: OFA (Wang et al., 2022b) is a unified paradigm for multimodal pre-training, which is both task and modality agnostic. For pre-training, OFA unifies several vision and language tasks including image generation, visual grounding, image captioning, image classification and language modeling among others, and is pre-trained using 20M publicly available image-text pairs. The OFA (Huge) model that we use is fine-tuned on MS-COCO for image captioning, and unless otherwise specified, we generate baseline captions using beam search with 16 beams, and generate candidate captions for IC3 using temperature sampling as described in subsection 6.2.1.

B.3.2 Datasets

We explore image captioning across several datasets in chapter 6.

MS-COCO Dataset: The MS-COCO dataset (Lin et al., 2014) is a dataset for image description containing 328K images, each with 5 ground truth descriptions. MS-COCO is licensed under a Creative Commons Attribution 4.0 license. All of the results in this work are presented on the test-split of the Karpathy splits of the COCO-2014 dataset.

Flickr-30K Dataset: The Flickr-30K dataset (Young et al., 2014) is an image description dataset containing 30K images and 150K captions (5 ground truth captions per image), and is licensed under a custom non-commercial (for research use only) license. All of the results are presented on the test-split of the Karpathy splits of the Flickr-30K dataset.

Hard MRR Splits: In some situations, we want to be able to explore the performance of our model vs. baselines on the most challenging captions. We call these splits the "Hard MRR" splits, and they consist of the set of 200 samples for which the MRR of the underlying captioning model is lowest. Thus, HARD MRR - BLIP contains the 200 samples minimizing MRR for the baseline BLIP model, with a caption generated using beam search (16 beams), and similarly HARD MRR - OFA contains the 200 samples minimizing MRR for the baseline OFA model with a caption generated using beam search (16 beams).

B.3.3 N-Gram Metric Scores

The performance of the model on traditional n-gram measures is demonstrated in Table B.6. In this work, IC^3 models are designed to produce captions which are the combination of all of the viewpoints presented by each of the individual captions, suggesting that they contain more information on average than any single reference sentence. Because of this, often the n-gram performance of the model is significantly worse, as while the overlap of content n-grams may be higher (suggested by Table 6.7 and Table 6.8 in chapter 6), there are a lot of extra n-grams per caption, which will decrease metric scores. We explore four n-gram measure: BLEU (Papineni et al., 2002), CIDEr: (Vedantam et al., 2015), METEOR (Agarwal and Lavie, 2008) and ROUGE-L (Lin, 2004). We also compute the MAUVE (Pillutla et al., 2021) score between *all* samples generated and the reference samples, which measures the deviation in the language space, and notice that the MAUVE score is extremely low, suggesting that we have succeeded in producing a language distribution which is significantly different from the reference distribution.

B.4 ELO Scoring for Human Ratings

The results shown in subsection 6.3.1 indicate a challenging reality: humans can often find it difficult to calibrate to the quality of image captions when viewing single image captions alone, but can find it much easier to understand any differences in quality when presented with two pairs of captions, in a head to head fashion. Unfortunately, since human caption ratings can be expensive, it is tricky to perform all head to head caption evaluations across

Table B.6: Performance of models augmented with IC^3 on traditional N-gram measures.

Model	BLEU@4 ↑	CIDER ↑	ROUGE-L ↑	MAUVE ↑
MS-COCO (Karpathy Split)				
OFA + IC^3	0.159	0.495	0.483	0.091
BLIP + IC^3	0.118	0.325	0.445	0.074
OFA	0.292	1.323	0.598	0.254
BLIP	0.292	1.315	0.595	0.158
Flickr-30K (Test Split)				
OFA + IC^3	0.132	0.392	0.449	0.004
BLIP + IC^3	0.092	0.277	0.401	0.004
OFA	0.212	0.872	0.541	0.004
BLIP	0.160	0.501	0.727	0.004

values of K, language models, captions, etc. In order to compensate for this, in some situations instead of running the full head to head experiment, we instead use a tournament, which measures the quality of a model through an Glicko-2 score-based rating system (Glickman, 1995). In some cases, we report the Glicko-2 scores of each of the models in our human-rating tournament, as a proxy for the overall quality of the model.

B.5 Human Studies

In our work, we run two different human rating studies, a head-to-head comparison between methods, and a context-free method which generates mean opinion scores. A screenshot of our evaluation tool for mean opinion scores is given in Figure B.11, and a screenshot of the evaluation tool for head-to-head rating is given in Figure B.12. Both of these experiments have been approved as *exempt* under category 2 by the UC Berkeley IRB, protocol ID 2022-11-15846. For any questions, contact `ophs@berkeley.edu`.

Significant prior work has explored the collection of human judgments of the quality of visual descriptions. Human judgment is considered the gold standard for visual description evaluation, and previous studies typically rely on human annotators to rate caption quality on one or multiple axes (Levinboim et al., 2021; Kasai et al., 2022). While automated methods exist for the evaluation of caption quality (Agarwal and Lavie, 2008; Vedantam et al., 2015; Papineni et al., 2002), recent work including THUMB (Kasai et al., 2022), which has run human evaluations on captions produced by models based on "Precision", "Recall", "Fluency", "Conciseness" and "Inclusive Language", has shown that humans produce captions which score significantly higher when judged by human raters, than when judged by existing measures

(and further, that human judgments of quality correlate poorly with existing measures), necessitating the need for human evaluation as opposed to evaluation of captioning methods using automated measures for caption quality. Our model quality evaluation method is closely aligned with the work in (Levinboim et al., 2021), where we use similar questions to determine the "helpfulness" and "correctness" of visual descriptions. Our work differs in that instead of collecting a set of human ratings of captions for the purpose of training statistical models, we aim to evaluate the quality of both human and machine generated captions, in an effort to determine if the machine generated captions from recently proposed methods in our group outperform existing human and machine generated captions for the images.

In this study, subjects participate in sessions of reviewing visual media with corresponding visual descriptions, e.g. pairs of images and captions. These sessions consist of sequences of rating tasks, with a session consisting of not more than ten tasks. The types of tasks, which we call activities, are as follows:

- Caption Rating - Viewing a given image and caption pair, and rating the quality of the caption on several axes (described below).
- head-to-head Caption Rating - Viewing a given image, and a pair of captions, and deciding which caption better describes the image.

While there are several possible activities, each session consists of sequences of the same type of activity, and each task is presented in randomized order for each subject.

Subjects are linked to the data collection interface on our server developed by us in a frame directly from an Amazon Mechanical Turk internal HIT using the ExternalQuestion API which allows external web content to be displayed within the internal HIT. No third-party software is used with the HITs and no reviewing data is collected by Amazon or any third-parties with the use of this API.

The subjects are shown a consent form on the Amazon Mechanical Turk HIT prior to entering our data collection interface. Subjects are then required to click the "I Accept" button to confirm their agreement with the consent information of the study. They are then redirected to the data collection interface. For each image, users are presented with an image, and an associated image description. Images are drawn from the MSCOCO dataset (Lin et al., 2014). Human generated captions are drawn from the references collected by the authors of (Lin et al., 2014).

For task A (Caption Rating), users are first asked to rate the "helpfulness" of the caption, with the prompt: "Does the caption provide useful information for a person who cannot see the image (and who is trying to understand what is in the image)?", among the options: Not helpful at all", "Slightly helpful", "Moderately helpful", "Helpful", "Very helpful".. The user can also select the option "I can't tell". The user is then asked to rate the "Correctness" of the caption with the prompt "How correct is the caption?", among the options "Completely wrong", "Mostly wrong" "Slightly wrong", "Slightly right", "Mostly right", "Completely right". The user can also select the option "I can't tell".

The user is then asked to select the "submit" button, to move to the next task in the HIT, which is composed of 10 tasks. The user can also skip the task by selecting the "Image/Caption not visible button". If the user selects "I can't tell" or "Image/Caption not visible" for any option, the tasks remaining are not decreased, but if the user selects submit, and passes a valid rating, then the number of tasks remaining are reduced by 1.

For task B (head-to-head Caption Rating), the user is presented with two image captions instead of one image caption, and asked to select "Caption A better represents the content in the image" or "Caption B better represents the image". The user can also select "I can't tell". The user is then asked to select the "submit" button, to move to the next task in the HIT, which is composed of 10 tasks. The user can also skip the task by selecting the "Image/Caption not visible button". If the user selects "I can't tell" or "Image/Caption not visible" for any option, the tasks remaining is not decreased, but if the user selects submit, and passes a valid rating, then the number of tasks remaining is reduced by 1.

After completing all of the tasks in the session, users are given a randomly generated code, which is entered in the Amazon MTurk HIT page, and links the user's survey results to the Amazon worker ID. We collect these linkings to perform analysis on inter-rater agreement, as while the session itself is anonymous, users may complete multiple sessions, and some method is required to maintain identity between the sessions.

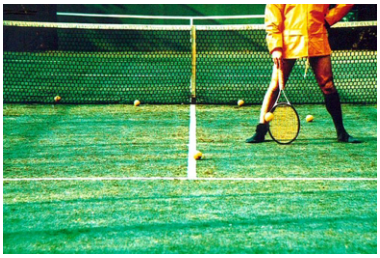
After each of these sessions, subjects are given a brief survey regarding the task difficulty (Select from the options: "Very Easy", "Easy", "Normal", "Hard", "Very Hard") and prompted for any additional comments on the session in general for each session in an (optional) open-response format. Users are also encouraged to protect their privacy with the prompt: "After submitting your responses, you can protect your privacy by clearing your browser's history, cache, cookies, and other browsing data. (Warning: This will log you out of online services.)" Subjects were compensated with $0.18 USD per session (based on the recommended Amazon wage (federal minimum wage, $7.25/Hr), with an expected completion time of 1.5 minutes per session), and should be able to complete the session in under one and half minutes (based on several pilot examples). Subjects can participate in the task a maximum of 100 times. The maximum time commitment for each subject over two months of our study is 2 hours.

To compute p-values, we first aggregate each users' session scores for each model (in the case of MOS, we take the mean for each model, and in the case of head-to-head, we assign a $+1$ value for a win, and a -1 value for a loss, and take the mean). For MOS, we compute a 1-sided t-test on the aggregated samples (which should be independent) to the baseline, while for the head-to-head scores, we compute a 1-sided single-sample t-test against a mean of zero.

B.6 Additional Qualitative Examples

Additional qualitative examples are given in Figure B.3, Figure B.4, Figure B.5, Figure B.6 and Figure B.7. From these examples, it is clear that IC^3 outperforms the baseline in many situations. Examples in this section are randomly selected from the test set when indicated.

(a) BLIP+IC^3: A plate with an orange, crackers, lettuce, and possibly other items such as nuts or a book.
BLIP: A close up of a plate of food on a table.



(b) BLIP+IC^3: A man standing on a tennis court holding a tennis racquet, possibly wearing an orange outfit or raincoat.
BLIP: A man standing on top of a tennis court holding a racquet.

(c) BLIP+IC^3: A woman riding a brown horse and jumping over hurdles in a competition, with other people watching.
BLIP: A woman riding on the back of a brown horse.

Figure B.3: Additional qualitative examples of BLIP + IC^3 on the MS-COCO dataset.

(a) OFA+IC^3: A woman in a bikini jumping in the air to hit a volleyball on a beach, possibly while playing a game of beach volleyball.
OFA: A woman in a bikini is jumping in the air to hit a volleyball.

(b) OFA+IC^3: Two women in kimonos standing in front of an information board with umbrellas, possibly in the rain.
OFA: Two women with umbrellas standing in front of an information board.

(c) OFA+IC^3: A lacrosse player in a white jersey running down the field with the ball during a game or match.
OFA: A lacrosse player runs with the ball.

Figure B.4: Randomly selected qualitative examples of OFA + IC^3 on the Flickr30K dataset.

(a) BLIP+IC^3: A man in striped trunks riding a surfboard on a large wave near a group of people.
BLIP: A man riding a wave on top of a surfboard.

(b) BLIP+IC^3: A young boy standing outside of a building, possibly in front of a window or doorway, holding a cell phone to his ear and wearing a red shirt.
BLIP: A little boy standing outside of a building talking on a cell phone.

(c) BLIP+IC^3: Two people, possibly hikers, sitting on top of a mountain, possibly icy or rocky, overlooking a snowy valley.
BLIP: A couple of people sitting on top of a mountain.

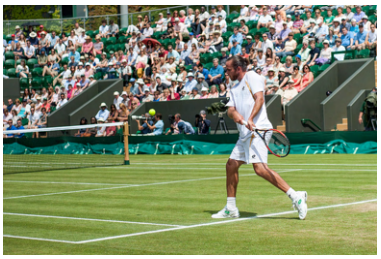
Figure B.5: Randomly selected qualitative examples of BLIP + IC^3 on the Flickr30K dataset.

(a) OFA+IC^3: A group of people sitting on a bench under a tree, with four green street signs hanging from it.
OFA: A group of people sitting on a bench under a tree.

(b) OFA+IC^3: A plate of food on a table with rice, beans, and possibly a meat dish, such as chicken or mashed potatoes.
OFA: A plate of food on a table.

(c) OFA+IC^3: A plate with two items of food on it, possibly a sandwich and a burrito, or two empanadas, sitting on a table or counter.
OFA: A plate of food on a table.

Figure B.6: Randomly selected qualitative examples of OFA + IC^3 on the MS-COCO dataset.



(a) BLIP+IC^3: A male tennis player wearing all white, walking across a tennis court while holding a racquet, possibly after losing a match.
BLIP: A man walking across a tennis court holding a racquet.

(b) BLIP+IC^3: A banana, bowl of cereal, and cup of coffee sitting on a table or counter.
BLIP: A banana sitting next to a bowl of cereal and a cup of coffee.

(c) BLIP+IC^3: A man wearing either red or green and white, holding a tennis racquet and swinging it at a tennis ball on a tennis court.
BLIP: A man holding a tennis racquet on a tennis court.

Figure B.7: Randomly selected qualitative examples of BLIP + IC^3 on the MS-COCO dataset.

B.7 Zero-Shot Style and Language Transfer

It is well known that models such as GPT 3 (Radford et al., 2021b) are capable of many zero-shot tasks, such as language style transfer and translation. By modifying the prompt in the summarization approach, IC^3 can be used to generate captions in different styles and languages. For example, we can modify the prompt to generate captions in different languages, for example, to generate captions in Japanese, we could use the prompt:

```
This is a hard problem. Carefully summarize in Japanese in ONE detailed
sentence the following captions by different (possibly incorrect) people
describing the same scene. Be sure to describe everything, and identify when
you're not sure. For example:
Captions: {formatted captions}.
Summary (in Japanese): 写真はおそらく
```

We can see the performance of the model for such prompts in Figure B.8. Such captions represent an easy way to transfer knowledge to different languages, however may not outperform translating the English caption alone.

Figure B.8: Examples of the generated caption (BLIP + IC^3) for different languages:
English: A woman riding a brown horse and jumping over hurdles in a com- petition, with other people watching.
Spanish: Una mujer montando a caballo un caballo marrón mientras salta un obstáculo en un campo verde, posiblemente en una competición con espectadores mirando.
French: Quelqu'un qui monte à cheval sur le dos d'un cheval brun et saute par-dessus un obstacle dans un champ, avec des gens en arrière-plan.
Japanese: 女性が茶色の馬の背中に乗って障害物を跳び越える様子を捉えたものであるが、確実ではない。

(a) Generated BLIP+IC^3 Caption: A man sitting at a desk in front of at least one computer, possibly two, **with other details such as clothing and accessories varying**.

(b) Generated BLIP+IC^3 Caption: A small bathroom with a white toilet, sink, counter, and possibly a marble tile floor, **and there may be a cat present**.

(c) Generated BLIP+IC^3 Caption: A white plate topped with a sandwich and a cup of coffee, **possibly accompanied by other food items such as french toast and/or meat**.

Figure B.9: Examples of BLIP + IC^3 failure modes (MS-COCO Dataset).

B.8 Failure modes & Limitations

In this section of the appendix, we explore some of the limitations of the method, and provide some insight into how the model could be improved.

B.8.1 Hallucination

Some examples of failure cases are shown in Figure B.9. In Figure B.9a, we can see the effect of “4th-wall breaking," one of the key failure modes of the method. Because the prompt suggests that that the model should combine several underlying captions, the output caption references the fact of this combination in a hidden way, when it says “other details varying". In some cases, the model might produce captions that end with words such as “... as stated by the captions" or “... but the captions differ." which both reference the prompt, and interfere with the flow of the caption.

In Figure B.9b, we can see a situation where the model passes through a hallucination from the underlying captioning model. Because 3 of the 10 captions in the candidate set K mention a cat: “A cat in a bathroom staring into a sink", “A bathroom with a toilet and sink and a cat", and “A cat walking around in a bathroom", and the LLM is not visually aware, there is no reason to doubt the existence of the cat, and it is included in the caption.

Luckily, in this failure case the model prefaces the existence of the cat with a “may", however there are situations where this is not the case.

In Figure B.9c, we can see the third major failure case of the model: treating uncertainty as multiple objects. Because the captioning model is not aware of the visual content of the image, when there is a high amount of noise in the captions, such as here, where the actual contents of the plate are unclear, the model often ascribes the noise to several objects in the scene, instead of a single uncertain object. This can sometimes be automatically detected by counting the number of commas in the caption, and we have found empirically that re-generating any caption with more than 7 commas can reduce or eliminate these effects (though we do not use this post-processing step in chapter 6).

B.8.2 Controllable Alt-Text Generation

While our model is capable of generating high-fidelity descriptions of the image, as discussed in section 6.4, the model can struggle when asked to describe background and contextual details that differ significantly from the reference dataset distribution. To demonstrate this, we perform a case study with the image in Figure B.10.

Figure B.10: Photo by Zhang Kaiyv[1]. Alt-text is often contextual. From the reference: “If [the image] is used in an article about a public bicycle hire scheme named Hire-a-bike, then the bike is the focus, and the alt text could be "A woman rides a Hire-a-bike along a city road". However, if it’s used in an article about a dispute between the café and the restaurant, the alt text might be "The storefronts of the ’Café Bar Hotel’ and ’Alpen Hotel Restaurant’". Although if the dispute is about color of the buildings, the guttering, or the window frames, then those details should be part of the alt text."

In the case study, we take the prompt:

`This is a hard problem. Carefully summarize in ONE detailed sentence the following captions by different (possibly incorrect) people describing the same scene. Be sure to describe everything, and identify when you’re not sure. For`

```
example:
Captions:  {formatted captions}.
Summary:  I'm not sure, but the image is likely of...
```

and replace the colored prompt with a set of different prompts, to generate potential alt-text for the image in Figure B.10. From these prompts, we can see that in many cases, the model fails to be controllable, and only repeats the key information in the scene. While in some cases the prompts can elicit additional information, we believe there is significant work to be done to explore how we can sample enough relevant information from the base captioning model (it seems that increasing the number of samples is not always sufficient), and surface that information in the generated summary (simple modifications to the prompt seem ineffective).
The results for BLIP are below:
Prompt: With original prompt
Result: A woman riding a bicycle past a building with a cafe, bar, hotel, and/or restaurant, possibly with a fence nearby.
Prompt: "Focus on describing what the person is doing"
Result: A person riding a bicycle past a building or establishment of some kind.
Prompt: "Focus on describing the buildings"
Result: A woman riding a bicycle past some buildings, possibly a hotel, with a gate in the background.
Prompt: "Only describe the buildings"
Result: A person riding a bicycle past a tall building.
Prompt: "do not mention any people in the summary"
Result: A person riding a bicycle down a street past buildings.
Prompt: "Focus on describing the color and architecture of the buildings"
Result: A person riding a bike past a hotel with colorful wooden shutters and other buildings nearby.
Prompt: "Focus on describing what the woman is wearing"
Result: A woman wearing casual clothing riding a bike past two colorful buildings.
Prompt: "Make sure to describe the appearance of the buildings in detail"
Result: A person riding a bicycle in front of a building, which could be an old hotel, cafe, apartment building, restaurant, or other type of building with many windows and possibly a fence or gate.
Prompt: "Make sure to describe the person's appearance in detail"
Result: A person, possibly a woman, riding a bicycle past buildings, a fence, a restaurant, and a hotel.
Prompt: "Make sure to describe the person's appearance in detail, such as what they are wearing and doing"
Result: A woman in a floral dress riding a red bike past an old, historic building with a restaurant across the street, while an older man is riding a bicycle past several other older-fashioned buildings.

The results for OFA are below:
Prompt: With original prompt
Result: A woman riding a bicycle in front of a building, fence, or restaurant in a variety of cities in Europe.
Prompt: "do not mention any people in the summary"
Result: A person riding a bicycle in front of a building, fence, or restaurant.
Prompt: "Focus on describing what the person is doing"
Result: A person riding a bicycle in front of a building, fence, or restaurant.
Prompt: "Focus on describing what the person is doing" + "I'm not sure, but the image is likely of a person, who is"
Result: Possibly a woman, riding a bicycle on a street in front of a building, fence, metal gate, restaurant, hotel, or other structures in various cities such as Tallinn, Vienna, Albany, Alphen aan den Rijn, Berlin, Strasbourg, or Frankfurt am Main.
Prompt: "Make sure to include details about what the person is wearing and doing"
Result: A woman riding her bicycle past a tall fence, possibly with a restaurant behind it, in front of a building or hotel in a city such as Kitzingen, Albéliyah, Breckenridge, Frankfurt, Innbruck, or Bad Schwartau..
Prompt: "Make sure to only include details about what the person is wearing and doing"
Result: A woman riding a bicycle on a city street in front of a building, fence, or restaurant.

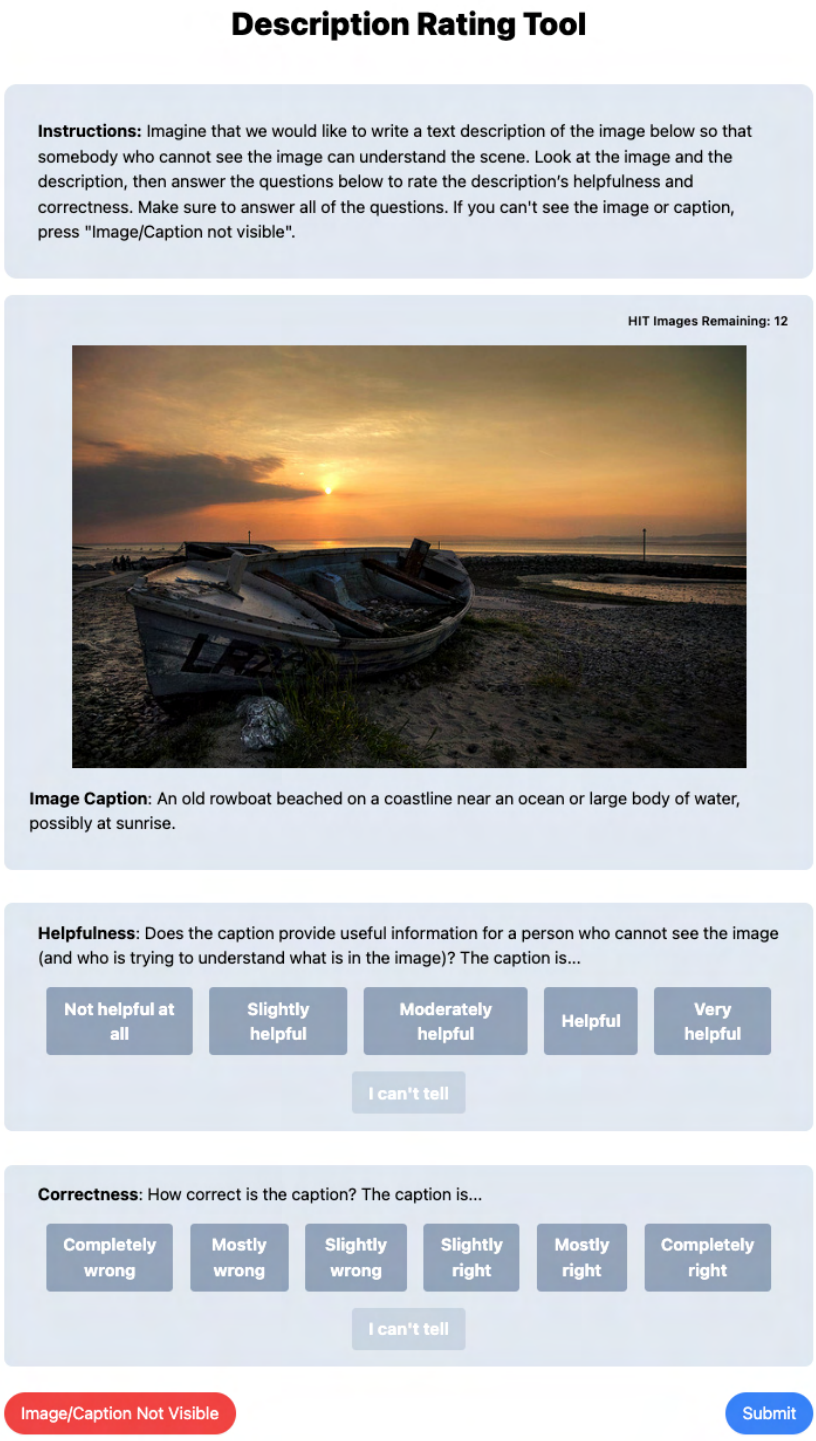


Figure B.11: Description rating tool (Mean Opinion Scores).

Figure B.12: Description rating tool (head-to-head).

Appendix C

Appendix for Triangle-Rank Metrics for Distribution Aware Conditional Natural Language Generation

Appendix

C.1 Additional Experimental Details

In this section, we discuss additional experimental details for interested readers.

C.1.1 Code

We make all code/data publicly available for use at `https://github.com/CannyLab/vdtk`. We hope that releasing our code, along with the JSON files containing test-set predictions for the models in question will help inspire further research and examination into the evaluation of models for visual description.

C.1.2 Datasets

MSR-VTT Dataset: The MSR-VTT dataset (Xu et al., 2016) is a dataset for video description consisting of 10,000 videos, with 20 reference ground truth descriptions for each video. It was collected by downloading 118 videos for each of 257 queries from a popular video sharing website. MSR-VTT contains 41.2 hours of video, with an average clip length lying between 10 to 30 seconds. It has a vocabulary size of 21,913. For more details about the diversity of the language present in the dataset, we refer readers to Chan et al. (2022c).

MS-COCO Dataset: The MS-COCO dataset (Lin et al., 2014) is a large-scale dataset for image description, object detection and segmentation. MS-COCO contains 328K images, each with 5 ground truth descriptions generated by human AMT workers. For more details about the diversity of the language present in the dataset, we refer readers to Chan et al. (2022c). MS-COCO is licensed under a Creative Commons Attribution 4.0 license.

C.1.3 Models

In chapter 11, we explore the performance of our metrics over several models: two video captioning models, and two image captioning models.

TVT The Two-View Transformer (Chen et al., 2018) is a baseline method for video description, which consists of a transformer encoder/decoder structure. While we did not have access to the original code, we trained our own version of the model on the MSR-VTT dataset (standard splits), leveraging features from Perez-Martin et al. (2021). The model was trained for 300 epochs, with a batch size of 64, model hidden dimension of 512, 4 transformer encoder and decoder layers with 8 heads each, and dropout of 0.5. For optimization, we

leveraged the Adam optimizer with a learning rate of $3e^{-4}$ and weight decay of $1e^{-5}$ with exponential learning rate decay with gamma 0.99. This model achieves a $CIDEr$ score of 56.39 on the test dataset. The model was trained using a Titan RTX-8000 GPU over the course of several hours.

O2NA O2NA (Liu et al., 2021a) is a recent approach for non-auto-regressive generation of video captions. While the method had available code and checkpoints which we used for this experiment, the method is not designed to sample more than one candidate caption at any given time. To adjust the model to sample multiple candidate captions, we made several adjustments. First, the model was modified to sample a length according to a softmax distribution over the length likelihoods (instead of using a greedy choice of length, or beam search over lengths, as proposed in chapter 11). Second, the model was modified to sample tokens at each non-autoregressive step from a temperature-adjusted softmax distribution instead of greedily sampling tokens. We make our modified code available as a patch to the original repository, in the hopes that other users will continue to build on these alterations.

CLIPCap CLIPCap (Mokady et al., 2021) is a recent model for image description based on using the CLIP (Radford et al., 2021c) model for large vision and language pre-training as a feature encoder, and GPT (Brown et al., 2020) as a natural language decoder. CLIPCap code and MS-COCO trained model checkpoints are publicly available from the authors, however we made some alterations to support temperature-based and nucleus sampling. We make our modified code available as a patch to the original repository, in the hopes that other users will continue to build on these alterations. CLIPCap is licensed under the MIT license.

VLP VLP (Zhou et al., 2020) is a unified vision and language pre-training model, designed to perform both image captioning and visual question answering. The model is pre-trained on the Conceptual Captions (Sharma et al., 2018) dataset, and fine-tuned on the MS-COCO captions dataset for image description. The authors make code and pre-trained models publicly available, however we modified the code somewhat to support additional sampling methods. We make our modified code available as a patch to the original repository, in the hopes that other users will continue to build on these alterations. VLP is licensed under the Apache License 2.0.

C.1.4 Distance Metrics

In chapter 11, we explore three base semantic metrics as distance underlying our TRM methods, CIDEr-D (Vedantam et al., 2015), METEOR (Agarwal and Lavie, 2008), and BERT Distance (Zhang et al., 2020d).

CIDEr-D CIDEr-D (Vedantam et al., 2015) is a n-gram-based metric designed for visual description, and based on the idea that common words are less useful in practice than

uncommon words. In practice, this takes the form of a cosine similarity between TF-IDF weighted vectors representing the sentences. Because CIDEr-D is a score, and not a distance, we create a distance function: $d(c, r) = 10 - C(c, r)$, which works as CIDEr-D is bounded by 10. Note that because CIDEr-D is 10 if and only if and only if the two sentences are equal, this fulfills the TRM requirements.

METEOR METEOR (Agarwal and Lavie, 2008) is a score which evaluates the semantic distance between two text utterances based on one-to-one matches between tokens in the candidate and reference text. The score first computes an alignment between the reference and candidate, and computes a score based on the quality of the alignment. Because METEOR is a score, and not a distance function, we use the distance $d(c, r) = 1 - M(c, r)$, where M is the METEOR score of the reference. Because METEOR is bounded at 1 if and only if the two utterances are identical, this simple transformation satisfies the requirements of the TRM adjustment. While we could explore other ways of deriving a distance from METEOR, we found that this simple approach was sufficient to demonstrate the performance of our methods.

BERT Distance A recent method for determining the semantic distance between two samples is to leverage a pre-trained BERT embedding model to create a semantic embedding of the text, and computing the cosine distance between the test samples. In our work, we leverage the `MiniLM-L6-v2` model from the sentence-transformers package by Reimers and Gurevych (2019) to embed our descriptions. Because cosine distance is already a distance function, no additional transformation is necessary.

C.1.5 P-value Computations

For our experiments, our null hypothesis is that *the candidate samples and the ground truth samples are drawn from the same distribution.* Because most of the methods do not have an analytical way to compute the p-values (in fact, the TRMs are the only method which has an analytic p-value computation given in Liu and Modarres (2011)), we instead must compute the p-values though sampling. We thus enumerate the value of the statistic across all of the possible candidate/reference partitions given the joint set of candidates and references, and determine the probability of observing the sampled value, or some value more extreme.

The values in Table 11.1 represent the p-value obtained with a single candidate sentence, and 4 ground truth candidates for MS-COCO, or 19 ground truth candidates for MSR-VTT. We reserve one gorund truth description in both datasets to serve as the “Human" performance description. For TVT, CLIPCap and VLP, we sample the descriptions using beam search with 16 beams. For O2NA, which is a non-autoregressive model, we sample according to the method suggested in the original work (see Liu et al. (2021a)). Because there are several thousand videos per dataset, computing all possible combinations across the dataset would

be far from tractable. Thus, the p-values were computed on a per-visual-input basis, and then aggregated across videos using the harmonic mean, as suggested by Wilson (2019). Such an aggregation method is valid when the experiments are not independent (which they are not), unlike Fischer's method (Fisher, 1992).

Figure 11.3 demonstrates the log p-values for the proposed methods across several candidate samples. For MS-COCO, we use all five reference captions, and between one and ten candidate captions sampled from CLIPCap using Nucleus Sampling (Holtzman et al., 2020) with a temperature of 1.0, top-p of 0.9 and top-k of 20. The caption set is generated once, meaning that the two-candidate set consists of the one-candidate set and one more additional caption. For MSR-VTT, we use 10 reference captions, and between one and seven candidate captions sampled from O2NA as described in appendix C.1.3 with a temperature of 1.0 for both the length and token samples. We do not go to the full 10 candidate captions for MSR-VTT due to tractability concerns, since adding an additional caption forces twice the number of partitions to be evaluated when computing p-values.

The above experiments were performed on several n2d-standard-32 cloud GCP instances, containing 32vCPUs and 128GB of RAM.

C.1.6 Frechet BERT Distance

The Frechet Inception Distance, originally proposed in Salimans et al. (2016), has often been used for the evaluation of the distance between samples of images generated by GANs. Images are first embedded in a latent space using a pre-trained inception network, and then the Frechet distance between the generated samples and the reference samples is computed. In our work, we replace the images with text, and the inception network with a pre-trained BERT embedding network (Devlin et al., 2019). For a set of candidate samples $(c_1, \ldots, c_n) = C$, a set of reference samples $(r_1, \ldots, r_m) \in R$, and a BERT embedding function $\phi_{\text{BERT}} : C \cup R \to \mathbb{R}^k$, we compute the Frechet BERT Distance as:

$$\begin{aligned} d^2 = &\left|\left| \frac{1}{n}\sum_{i=1}^{n} \phi_{BERT}(c_i) - \frac{1}{m}\sum_{i=1}^{n} \phi_{BERT}(r_i) \right|\right| \\ &+ \text{Tr}\left(C_C + C_R - 2\sqrt{C_C C_R}\right) \end{aligned} \tag{C.1}$$

where C_C andC_R are the covariance matrices of the C and R sets embedded with ϕ_{BERT} respectively.

To get the BERT embedding, we leverage the CLS token of a large pre-trained model, in this case, the `MiniLM-L6-v2` model from the sentence-transformers package by Reimers and Gurevych (2019).

The computation of p-values for the Frechet-BERT distance is largely bottle-necked by the slow performance of the `sqrtm` function, which, because the matrices are not symmetric, has no efficient algorithm for computation. Additionally, unlike the feature computation, this operation must occur for every partition, leading to significantly reduced efficiency compared to the other measures presented in chapter 11.

C.1.7 MMD-BERT

Another common metric in the GAN literature is the computation of a maximum-mean discrepancy between kernel-estimates of the samples introduced by Li et al. (2017). For a set of candidate samples $(c_1, \ldots, c_n) = C$, a set of reference samples $(r_1, \ldots, r_m) \in R$, and a BERT embedding function $\phi_{\text{BERT}} : C \cup R \to \mathbb{R}^k$, we compute the MMD-BERT distance as:

$$\begin{aligned} M\hat{M}D = &\sum_{i=1}^{N}\sum_{j=1}^{N} K(\phi_{\text{BERT}}(c_i), \phi_{\text{BERT}}(c_j)) \\ &+ \sum_{i=1}^{M}\sum_{j=1}^{M} K(\phi_{\text{BERT}}(r_i), \phi_{\text{BERT}}(r_j)) \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{M} K(\phi_{\text{BERT}}(c_i), \phi_{\text{BERT}}(r_j)) \end{aligned} \tag{C.2}$$

where K is a kernel function. In our experiments, we use an RBF kernel function with σ equal to the median distance pairwise distance divided by two.

C.1.8 Search Techniques

In section 11.2, Figure 11.6, we explore the performance of several different search techniques for our two-view transformer model on the MSR-VTT dataset. In this figure, we explore four decoding search techniques: Greedy Search, Beam Search, Temperature-Based Sampling, and Nucleus Sampling. For each method, and for each video in the test set, we sample 10 descriptions. For Greedy Search, we sample 10 repeated sentences. For beam search we sample the top beam search candidate, and repeat this ten times. While we did explore using the top 10 results from a larger beam search, we found that a smaller beam search and repeated values produced better METEOR scores, so we chose to compare against this. Wider beam searches did produce higher $\text{TRM}_{\text{METEOR}}$ scores, but because optimizing for METEOR would be the current paradigm, we decided to include that in the referenced figure. For standard temperature based sampling, we sampled 10 results at each temperature. For Nucleus sampling, we sample 10 results at each temperature, however we freeze they hyper-paramters of top-p at 0.9 and top-k at 20, as we found these values to generate the best scores under the standard pairwise metrics. It remains relevant future work to perform a deep-dive into the different generative methods with respect to TRMs, as there are likely many interesting lessons that can be learned.

C.1.9 Correlation with Human Judgement

In our work, we run a human correlation experiment to determine how well human ratings correlate with our metric's judgements. A screenshot of our evaluation tool for mean opinion

scores is given in Figure C.1. In each HIT, raters from Mechanical Turk were presented with the reference captions, along with two sets of candidate captions. These candidate captions were sampled from two models: OFA (Wang et al., 2022b) and BLIP (Li et al., 2022), at 11 different temperate settings: $0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. We then query the subjects with two questions, both of which can be evaluated on a scale of $\{-2, 2\}$, with 0 indicating a tie:

- Which group of candidate captions (as a whole) provides more useful information about the reference group for a person who cannot see the reference group?
- Which group of candidate captions (as a whole) matches best to the reference group factually?

Subjects are linked to the data collection interface on our server developed by us in a frame directly from an Amazon Mechanical Turk internal HIT using the ExternalQuestion API which allows external web content to be displayed within the internal HIT. No third-party software is used with the HITs and no reviewing data is collected by Amazon or any third-parties with the use of this API. The subjects are shown a consent form on the Amazon Mechanical Turk HIT prior to entering our data collection interface. Subjects are then required to click the "I Accept" button to confirm their agreement with the consent information of the study. They are then redirected to the data collection interface. For each image, users are presented with an image, and an associated image description. Images are drawn from the MSCOCO dataset (Lin et al., 2014). Human generated captions are drawn from the references collected by the authors of (Lin et al., 2014).

After completing all of the tasks in the session, users are given a randomly generated code, which is entered in the Amazon MTurk HIT page, and links the user's survey results to the Amazon worker ID. We collect these linkings to perform analysis on inter-rater agreement, as while the session itself is anonymous, users may complete multiple sessions, and some method is required to maintain identity between the sessions.

After each of these sessions, subjects will be given a brief survey regarding the task difficulty (Select from the options: "Very Easy", "Easy", "Normal", "Hard", "Very Hard") and prompted for any additional comments on the session in general for each session in an (optional) open-response format. Users are also encouraged to protect their privacy with the prompt: "After submitting your responses, you can protect your privacy by clearing your browser's history, cache, cookies, and other browsing data. (Warning: This will log you out of online services.)" Subjects were compensated with $0.18 USD per session (based on the recommended Amazon wage (federal minimum wage, $7.25/Hr), with an expected completion time of 1.5 minutes per session), and should be able to complete the session in under one and half minutes (based on several pilot examples). Subjects can participate in the task a maximum of 100 times. The maximum time commitment for each subject over two months of our study is 2 hours.

We analyze the experiments by first collecting all human ratings, and taking the mean of each score per image. We collect 5 ratings each for 794 images in the dataset, using 397

unique Mechanical Turk workers. We then compute the Pearson correlation for the standard max-aggregate scores, and for each of our methods against the mean of the human ratings. To compute the human-human correlation, we compute first the leave-one-out mean for each human rating, and compute the correlation of the leave-one-out mean with the existing images.

C.2 Additional Results

In this section we present several additional interesting results to augment those in the main discussion.

C.2.1 Embedding Methods for KBMs

In the main work, we primarily explore a BERT-based embedding method for the kernel-based methods. Such an exploration does not preclude the use of other embedding methods, each of which has different trade-offs, when looking at the quality of the resulting metric, what the resulting metric measures, the time required to compute the embedding, and the performance when the reference distribution is limited to small numbers of human samples (such as happens in practice). Figure Figure C.2 shows a quick look at several possible choices for embedding methods in the MMD-* family, including Bag of words (with a 5K vocab), GLoVe (Pennington et al., 2014), FastText (Bojanowski et al., 2017), and CLIP (Radford et al., 2021b).

While we can see that some of the methods are more sensitive to deviations in the image distributions, such methods come with additional trade-offs. CLIP-style embeddings are the most sensitive to human versus generated captions with fewer captions created, but are significantly slower to evaluate at test time (almost 4x slower) than MMD-BERT, and also produce a higher p-value when computing the leave-one scores on the human captions (which is less desirable, as the human captions are drawn from the same distribution).

C.2.2 Unique vs. Correct Descriptions

In Figure C.3, we explicitly demonstrate how TRMs enable evaluation of both caption diversity and quality. We artificially generate candidates for the MSR-VTT dataset by mixing human-generated exact descriptions with human-generated descriptions from other videos. On one axis we have the number of unique descriptions and on the other axis we have the number of correct (exactly-matching) descriptions. Clearly, unlike METEOR alone, $\text{TRM}_{\text{METEOR}}$ scores are affected by both correctness and diversity.

Each experiment consisted of 10 candidate captions from the MSR-VTT dataset, and 10 reference captions from the MSR-VTT dataset. We first split the 20 MSR-VTT reference captions into two sets of 10. One set of 10 captions formed the references. To select the candidate captions, we first sampled k unique captions from the remaining reference set

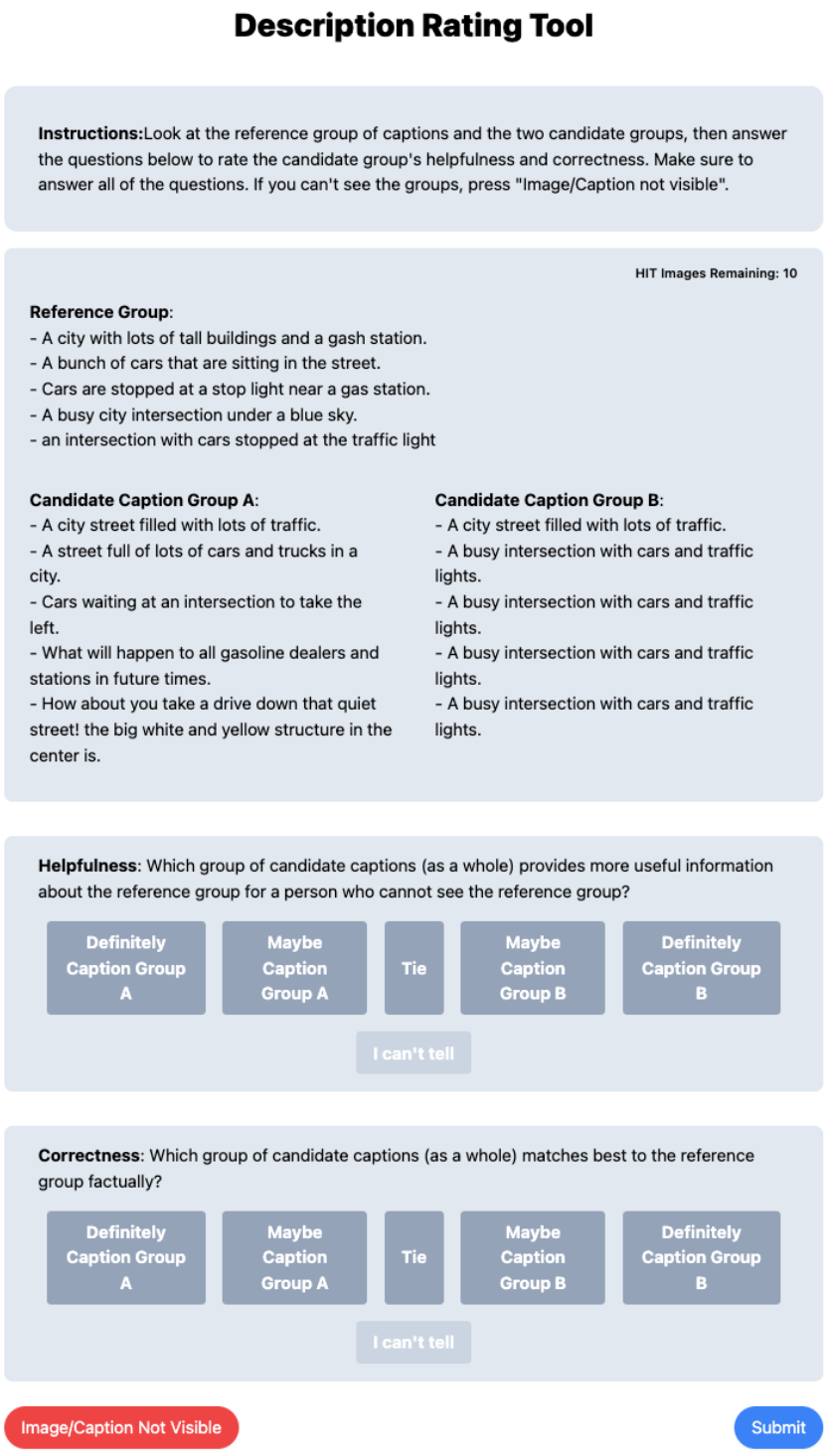
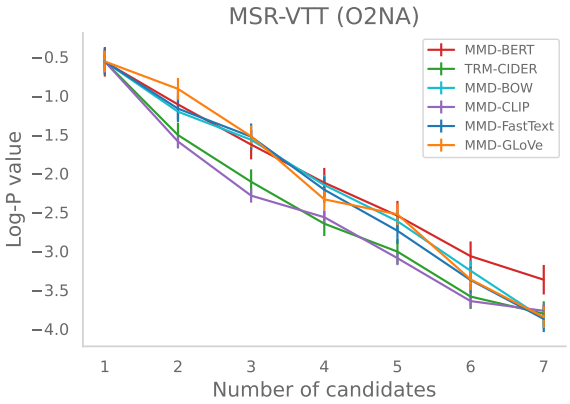


Figure C.1: A screenshot of our human rating interface.



Sensitivity and performance on Human-Generated Captions

Method	Log-P	Samples/Sec
TRM-CIDEr	-1.596	88.93
MMD-BERT	-1.786	56.68
MMD-CLIP	-1.887	14.41
MMD-GLoVe	-1.952	54.8
MMD-FastText	-1.954	57.45
MMD-BOW	-2.022	49.41

Figure C.2: Performance of several different embedding functions for the MMD-* family of metrics. Left: Sensitivity when evaluated on the MSR-VTT dataset with ten reference captions and between one and seven candidate captions generated by O2NA. Right: Sensitivity and speed when evaluated on human reference samples with 5 references and 5 candidates.

(which formed the “correct pool"), and k unique captions from other videos in the dataset at random (forming the “incorrect pool"). We then selected m correct captions, from the correct pool (at random) and $10 - m$ captions from the incorrect pool (at random). This was then plotted with m on the x-axis, and k on the y-axis, as a heat-map, where lighter colors represent better scores (higher METEOR, or lower TRM-METEOR), and darker colors represent poor scores.

We also explored the performance of the CIDEr metric across the same axes, the results of which are shown in Figure C.4. We can see that they are largely similar to those from the METEOR metric, suggesting that regardless of the underlying metric, we are still making similar trade-offs between diversity and correctness.

C.2.3 Human p-values

Strong metrics for distributional comparison will have high sensitivity to samples coming from distinct distributions, and will produce high p-values for samples which come from the same distribution. To check that such a relationship holds, we also perform leave-one-out experiments using human-generated captions from the reference set for both MSR-VTT and MS-COCO. For MSR-VTT, we split the reference data into sets of 10 candidate samples and 10 reference samples, and compute the deviations using this partitioning. For MS-COCO, we leverage the c40 split which has 40 reference descriptions for 5000 samples of the ground truth. We partition the references for each video into groups of ten descriptions, and compute the p-values from pairs of these partitions. Table C.1 gives the performance of the metrics on this human data.

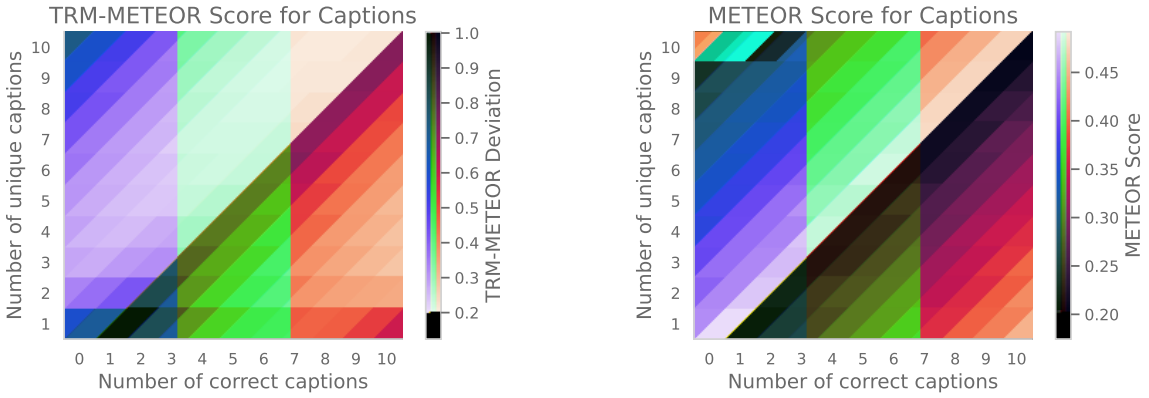


Figure C.3: Plots showing how TRMs evaluate both diversity and quality. Left: $\text{TRM}_{\text{METEOR}}$, Right: METEOR. Lighter colors represent better scores. While $\text{TRM}_{\text{METEOR}}$ trades off between diversity and quality, METEOR focuses only on quality not diversity.

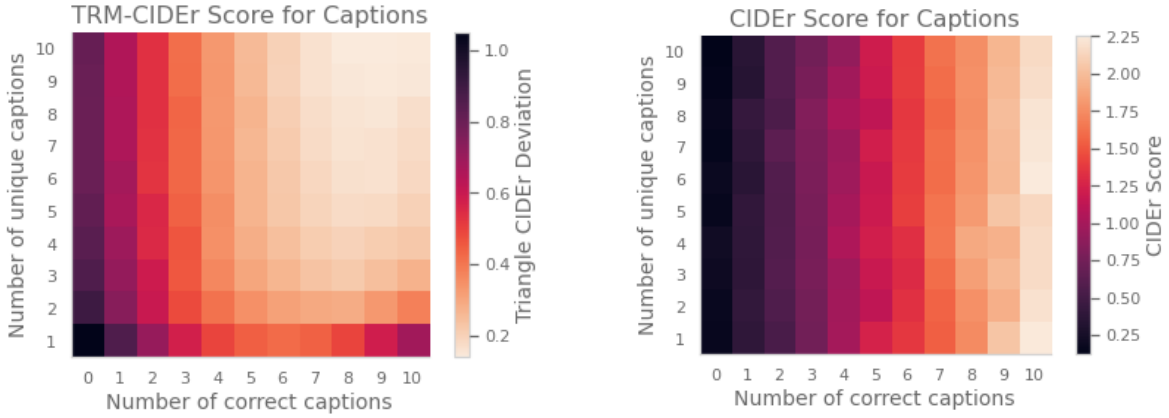


Figure C.4: Plots showing diversity vs. quality tradeoffs. Left: $\text{TRM}_{\text{CIDEr}}$, Right: CIDEr. Lighter colors represent better scores. While $\text{TRM}_{\text{CIDEr}}$ trades off between diversity and quality, CIDEr focuses only on quality not diversity.

C.2.4 MAUVE performance

In the main work, we found that MAUVE was prohibitively slow to use to compute p-values for the training data. Because our p-values were computed with 10 reference sentences, and up to 10 candidate sentences, at the existing rate, it could take several years to compute the MAUVE p-values for the 50,000 sample MS-COCO dataset. In Table C.2, we present several high-variance estimates of the MAUVE p-values (computed using only 100 samples).

	METEOR	$\text{TRM}_{\text{METEOR}}$	CIDEr	$\text{TRM}_{\text{CIDEr}}$	BERT	TRM_{BERT}	MMD-BERT
MSCOCO	-0.6303	-0.5941	-0.5957	-0.4742	-0.6230	-0.5633	-0.6550
MSR-VTT	-1.0046	-0.9613	-1.0224	-0.9777	-1.0172	-1.040	-1.0374

Table C.1: Log P-Values on human leave-one our samples. We can see that, surprisingly, none of the methods (even the standard aggregations) produce statistically signficant differences. That being said, TRMs often produce higher p-values, indicating that they may be more robust to noise in human caption sets. We do not compute the Frechet-BERT values for humans here, as it was prohibitively expensive.

Dataset	MAUVE Log p-value	METEOR Log p-value
MSR-VTT (O2NA)	-0.4414	-1.7881
MSR-VTT (Human Captions)	-0.1441	-0.6037
MS-COCO (CLIPCap)	-0.3980	-2.5585
MS-COCO (VLP)	-0.3234	-2.8609
MS-COCO (Human Captions)	-0.2189	-0.7233

Table C.2: Log p-value estimates for MAUVE using five candidates, five references, and 100 samples (at nucleus sampling temperature 1.0 for O2NA, CLIPCap and VLP models). We can see that Log p-values for MSR-VTT and MS-COCO are signficantly worse than METEOR even with aggregation, likely due to the method using k-means to approximate the text distributions with only 5 samples.

C.2.5 Visualizing Central Descriptions

We have found that descriptions which minimize the expected distance to the ground truth distribution are relatively sparse in detail compared to other descriptions. Figures C.5, C.6, C.7 and C.8 show qualitative examples of such descriptions for the MS-COCO dataset. Each plot shows qualitative examples of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions, and the other captions are the additional references in the MS-COCO dataset. Images are selected at random, and do not represent cherry-picked samples from MS-COCO.

C.2.6 Additional Qualitative Samples

Two hot dogs sitting side by side with condiments.
Two hot dogs are laden with relish, ketchup, and mustard.
>>> two hot dogs on a plate loaded with condiments
Two hot dogs covered with ketchup and relish on a plate.
Two hot dogs in buns are smothered with condiments.

The meal is ready on the tray to be eaten.
A breakfast was delivered to a hotel room on a tray.
a bunch of food and stuff is laying on a tray
>>> Bananas, cereal, juice and other breakfast foods on a tray.
This tray includes several different items for a full breakfast.

Figure C.5: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

A photo taken from a boat with a long bridge in the background.
A view of the coast from within a boat
The side of a boat and a bridge going over the ocean.
>>> A view of the lake, taken from a boat.
A boat flies its flag while sailing just off a pier.

a microwave on a kitchen counter above a dishwasher
this micro wave is black and silver and is on the counter
>>> A microwave oven sitting on top of a counter.
A microwave sitting on a counter, its stainless steel.
a silver microwave oven on a tan counter and a window

Figure C.6: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

A narrow city street has a leaning one way sign.
>>> a street with a line of cars parked on the side
Cars are parked alongside the road and a man is standing next to a sign.
A man is standing next to a road sign with a line of parked cars across the street in an urban area
A crooked one way sign pointing into the ground

A person pressing a button on a Wii controller.
A hand holds a remote that operates a video game.
There are no image to describe on this page..
>>> A person is holding a white Wii control
someone that is holding a wii remote in their hand

Figure C.7: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

Mom gives her daughter a lesson in using her baseball glove.
>>> Mother and her son playing in a few
two girls in red shirts grass and a baseball glove
A woman playing catch with her young child.
The mom is teaching her daughter to play baseball

a blue truck and a male in a purple shirt and a tree
Blue pickup truck filled with scrap pieces of household items.
A man has filled his truck with wheelchairs.
>>> A blue truck parked next to a tree and a man.
a man standing next to a truck full of bikes and a wheel chair

Figure C.8: Qualitative example of “central" captions. The caption marked with arrows is the ground truth caption which minimizes the expected METEOR distance to the other reference captions.

MSCOCO Image 421060

Candidate Set 1
METEOR (↑): 0.236
TRM-METEOR (↓): 0.912

A person on a snowboard does a trick in the air.
A person on a snowboard does a trick in the air.
A person on a snowboard does a trick in the air.
A person on a snowboard does a trick in the air.
...

Candidate Set 2
METEOR (↑): 0.264
TRM-METEOR (↓): 0.362

A person in mid air on a snowboard in front of a TV.
A snowboarder getting some air after a jump.
A man is performing a ski jump on a green slope.
A person on skis going down a ramp.
...

References

Competitive spirit during a competition in mid air
A skier races down the track at a competition.
A person is skiing on a slope covered in snow.
The skier is jumping into the air in a half pipe.
Skier performing aerial jump during outdoor competition.

Figure C.9: A qualitative sample from CLIPcap. Candidate set one uses beam search (8 beams), while candidate set two uses nucleus sampling (with temperature one, top-k of 20 and top-p of 0.9).

Appendix D

Appendix for CLAIR: Evaluating Image Captions with Large Language Models

D.1 Additional Experimental Details

In this section, we provide several additional details for the experiments in section 12.2 run with the CLAIR measure.

D.1.1 Input Prompt Formatting

The CLAIR prompt is given in its entirety in Figure 12.1. During run-time, candidate and reference captions are prefixed with a "- " and inserted into the prompt, one per line. The resulting query is passed to the large language model. In addition, for models which were not RLHF-tuned to perform conversation (such as PaLM), we found that it was helpful to append an additional prefix `{"score":` to the beginning of the output, to encourage the correct output formatting. CLAIR is surprisingly simple: it uses no in-context examples (is entirely zero-shot), and default inference parameters for the APIs. The model checkpoint metadata is generally unknown (as the APIs are somewhat fluid and evolving).

D.1.2 LLM Output Post-Processing

Because CLAIR relies on an LLM to produce output, there is no guarantee that the output will be in the format we expect (i.e. valid, parsable JSON). To extract both the score and the reason, we first extract the first set of paired braces from the output of the LLM and attempt to parse the result as JSON. In most cases (99.997% for GPT-3, 99.991% for Claude, and 99.94% for PaLM during the course of our experiments), this is successful, and the score and reason are returned. In the case that the JSON output is malformed, we attempt to extract any sequence of digits from the LLM to use as a score, and set the reason to "Unknown." When this fails, as can be the case when the models produce an output such as "As an AI language model, I cannot see, and thus, cannot determine if the image captions

match the references", we retry the prompt at a higher temperature ($t = 1.0$) several times. Failing this (which occurred only three times in the entire evaluation of this thesis, across several hundred thousand calls), we set the score to 0 for the caption.

D.1.3 Datasets

In this section, we provide additional detail regarding the datasets used in the evaluations in section 12.2.

COMPOSITE: The COMPOSITE dataset (Aditya et al., 2015) contains machine-generated test captions for 3995 images spread across the MS-COCO (Xu et al., 2016), Flickr8K (Mao et al., 2014) and Flickr30k (Young et al., 2014) datasets. Each image has three test captions, one written by a human, and two that are model generated. The candidate captions are graded by annotators on Amazon Mechanical Turk (AMT) on a scale of 1 (not relevant) to 5 (very relevant). Inter-human correlations are not available for this dataset.

Flickr8K-Expert: The Flickr8K-Expert dataset (Hodosh et al., 2013) contains 5822 captions associated with 1000 images. The dataset is annotated with expert human judgments of quality, where images are rated from 1 (caption is unrelated to the image) to 4 (caption describes the image without errors). Unlike the composite and MS-COCO datasets, the captions here are selected using an image retrieval system, instead of generated using a learned image captioning model. Following Jiang et al. (2019), we exclude any candidate captions that overlap the reference set.

MS-COCO: Following experiments by Rohrbach et al. (2018), we compute the sample-level correlation between our method and human ratings on a 500-image subset of the MS-COCO Karpathy test set. Each image in the subset contains candidate captions generated by 5 models, and each caption is labeled with the average three human ratings generated by AMT workers which range from 1 (very bad) to 5 (very good). Inter-human correlations are not available for this dataset.

PASCAL-50S: PASCAL-50S contains 1000 images drawn from the PASCAL sentence dataset. Each image is associated with at least 50 (and as many as 120) reference captions. In addition to the reference captions, PASCAL-50S contains a set of 4000 human annotated image/caption pairs containing an image, and two candidate captions. The caption pairs fall into four groups:

1. HC: In the HC group, both captions in the pair are human written, and describe the content of the target image correctly.

2. HI: In the HI group, both captions in the pair are human written, but one caption correctly describes the content of the image, and the other caption describes the content of a different image.

3. HM: In the HM group, one caption is written by a human, and one caption is written by a machine, but both correctly describe the content of the image.

4. MM: In the MM group, both captions are written by a machine, and both correctly describe the image content.

In PASCAL-50S, the task is to decide which caption in the pair humans prefer more (a subjective task, hopefully indicating caption quality). Following previous work (Jiang et al., 2019; Hessel et al., 2021b), we limit the number of reference sentences to five during evaluation.

COCO-Sets: The COCO-Sets dataset (Chan et al., 2022d) is a set of samples that are designed to evaluate the correlation of distribution-aware image captioning measures with human judgments of distributional distance. In this dataset, humans were presented with two candidate caption sets (two image captioning models, OFA (Wang et al., 2022b) and BLIP (Li et al., 2022) using different temperatures), and asked which candidate caption set correlated better with a reference caption set on two measures: how much they overlapped factually (correctness), and how much information they provided about the references (coverage). It consists of 794 AMT worker-generated judgments of caption quality for images in the MS-COCO dataset.