

Addressing and Understanding Shortcomings in Vision and Language

Kaylee Burns



Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-93

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-93.html>

May 22, 2019

Copyright © 2019, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

I'd like to thank my advisor, Professor Trevor Darrell, the other talented faculty members -- Professors Tom Griffiths, Kate Saenko, and Alison Gopnik, graduate students -- Lisa Anne Hendricks, Erin Grant, Eric Tzeng, and postdocs -- Anna Rohrbach and Aida Nematzadeh -- who have made this all possible.

Professors Paul Hilfinger, Josh Hug, Satish Rao, and Kannan Ramchandran taught me the importance of sharing your knowledge with others. I'm extremely grateful to them, my fellow TAs, my early mentors -- Sarah Kim and Alan Yao, and the amazing Student Affairs Staff.

Thank you Adora, Michelle, Lyn, Gary, and the many other friends who made college unforgettable.

Finally, I'm extremely thankful for my family's persistent love and support.

Addressing and Understanding Shortcomings in Vision and Language

by Kaylee Burns

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences,
University of California at Berkeley, in partial satisfaction of the requirements for the
degree of **Master of Science, Plan II.**

Approval for the Report and Comprehensive Examination:

Committee:



Professor Trevor Darrell
Research Advisor

5/22/19

(Date)



Professor Alexei Efros
Second Reader

5/22/19

(Date)

Addressing and Understanding Shortcomings in Vision and Language

by

Kaylee Burns

A thesis submitted in partial satisfaction of the
requirements for the degree of

Masters of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Trevor Darrell, Chair
Professor Alyosha Efros

Spring 2019

Addressing and Understanding Shortcomings in Vision and Language

Copyright 2019
by
Kaylee Burns

Abstract

Addressing and Understanding Shortcomings in Vision and Language

by

Kaylee Burns

Masters of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Trevor Darrell, Chair

Aligning vision and language is an important step in many applications; whether it's enabling the visually impaired to navigate the world through natural language or providing a familiar interface to otherwise opaque computational systems, the field is ripe with promise. Some of the largest roadblocks to realizing integrated vision and language systems, such as image captioning, are prediction artifacts from the training process and data. This report will discuss two weaknesses of captioning systems: the exaggeration of dataset bias related to gender presentation and the “hallucination” of objects that are not visually present in the scene.

The first chapter focuses on correcting the salient issue of gender bias in image captioning models. By introducing loss terms that encourage equal gender probability when gender evidence is occluded in a scene and confident predictions when gender evidence is present, we can enforce that the predictions are not only less error prone, but also more grounded in the image input.

In the second chapter, we broaden the lens of our analysis by developing a new image relevance metric to investigate “hallucinations”. With this tool, we will analyze how captioning model architectures and learning objectives contribute to object hallucination, explore when hallucination is likely due to image misclassification or language priors, and assess how well current sentence metrics capture object hallucination.

To my family

Contents

Contents	ii
List of Figures	iii
List of Tables	v
1 Introduction	1
2 Overcoming Bias in Captioning Models	2
2.1 Motivation	3
2.2 Related Work	5
2.3 Equalizer: Overcoming Bias in Description Models	8
2.4 Experiments	12
2.5 Discussion	24
2.6 Continuing the Conversation	25
3 Object Hallucination in Image Captioning	26
3.1 Motivation	27
3.2 Method: Caption Hallucination Assessment	29
3.3 Exploring Caption Hallucination Results	31
3.4 Discussion	41
Bibliography	42

List of Figures

2.1	Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender word correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.	3
2.2	Equalizer includes two novel loss terms: the Confident Loss on images with men or women (top) and the Appearance Confusion Loss on images where men and women are occluded (bottom). Together these losses encourage our model to make correct predictions when evidence of gender is present, and be cautious in its absence. We also include the Caption Correctness Loss (cross entropy loss) for both image types.	8
2.3	Accuracy across man, woman, and gender neutral terms for different models as a function of annotator confidence. When only one annotator describes an image with a gendered word, Equalizer has a low accuracy as it more likely predicts gender neutral words but when more annotations mention gendered words, Equalizer has higher accuracy than other models.	16
2.4	Qualitative comparison of multiple baselines and our model. In the top example, being conservative (“person”) is better than being wrong (“man”) as the gender is not obvious. In the bottom example the baselines are looking at the wrong visual evidence.	19
2.5	Qualitative comparison of baselines and our model when our model predicts “person” rather than “woman” or “man”.	22
2.6	Qualitative comparison of baselines and our model. At the top we show success cases where our model predicts the right gender for the right reasons. At the bottom we show failure cases with incorrectly predicted gender and the wrong gender evidence.	23

3.1	Image captioning models often “hallucinate” objects that may appear in a given context, like e.g. a <i>bench</i> here. Moreover, the sentence metrics do not always appropriately penalize such hallucination. Our proposed metrics (CHAIRs and CHAIRi) reflect hallucination. For CHAIR <i>lower is better</i>	27
3.2	Example of image and language consistency. The hallucination error (“fork”) is more consistent with the Language Model.	30
3.3	Examples of object hallucination from two state-of-the-art captioning models, TopDown and NBT, see Section 3.3.	34
3.4	Image and Language model consistency (IM, LM) and CHAIRi (instance-level, CHi) on deconstructed TopDown models. Images with less hallucination tend to make errors consistent with the image model, whereas models with more hallucination tend to make errors consistent with the language model, see Section 3.3.	35
3.5	Examples of how TopDown (TD) sentences change when we enforce that objects cannot be hallucinated: SPICE (S), Meteor (M), CIDEr (C), see Section 3.3.	37
3.6	Difference in percentage of sentences with <i>no</i> hallucination for TopDown and FC models when SPICE scores fall into specific ranges. For sentences with low SPICE scores, the hallucination is generally larger for the FC model, even though the SPICE scores are similar, see Section 3.3.	38

List of Tables

2.1	Evaluation of predicted gender words based on error rate and ratio of generated sentences which include the “woman” words to sentences which include the “man” words. Equalizer achieves the lowest error rate and predicts sentences with a gender ratio most similar to the corresponding ground truth captions (Ratio Δ), even when the test set has a different distribution of gender words than the training set, as is the case for the MSCOCO-Balanced dataset.	14
2.2	Accuracy per class for MSCOCO-Bias dataset. Though UpWeight achieves the highest recall for both men and women images, it also has a high error, especially for women. One criterion of a “fair” system is that it has similar outcomes across classes. We measure outcome similarity by computing the Jensen-Shannon divergence between Correct/Incorrect/Other sentences for men and women images (lower is better) and observe that Equalizer performs best on this metric. . . .	15
2.3	<i>Pointing game</i> evaluation that measures whether the visual explanations for “man” / “woman” words fall in the person segmentation ground-truth. Evaluation is done for ground-truth captions on the MSCOCO-Balanced.	18
2.4	Breakdown of error rate and difference to ground-truth woman:man ratio over images with specific biased words. We see that the full Equalizer generally outperforms the Baseline-FT. On error, Equalizer w/o ACL performs best, followed by Equalizer. Equalizer performs best when considering predicted gender ratio. .	21
3.1	Hallucination analysis on the Karpathy Test set: Spice (S), CIDEr (C) and METEOR (M) scores across different image captioning models as well as CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). All models are generated with beam search (beam size=5). * are trained/evaluated within the same implementation [34], † are trained/evaluated with implementation publicly released with corresponding papers, and ‡ sentences obtained directly from the author. For discussion see Section 3.3. . .	32
3.2	Hallucination Analysis on the Robust Test set: Spice (S), CIDEr (C) and METEOR (M) scores across different image captioning models as well as CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). * are trained/evaluated within the same implementation [34], † are trained/evaluated with implementation publicly released with corresponding papers. All models trained with cross-entropy loss. See Section 3.3. .	34

3.3	Hallucination analysis on deconstructed TopDown models with sentence metrics SPICE (S), METEOR (M), and CIDEr (C), CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). See Section 3.3	36
3.4	Pearson correlation coefficients between 1-CHs and CIDEr, METEOR, and SPICE scores, see Section 3.3	38
3.5	Pearson correlation coefficients between individual/combined metrics and human scores. See Section 3.3	39

Acknowledgments

I'm extremely grateful to all of my teachers, friends, classmates, and mentors.

None of this would have been possible without Berkeley's vibrant research community. I'd like to thank my advisor, Professor Trevor Darrell, for enabling my first excursions in research and for helping me chart new paths as I move forward in my career. I'd also like to thank the other talented faculty members who have helped me: Professors Tom Griffiths, Kate Saenko, and Alison Gopnik. The guidance of graduate students – like Lisa Anne Hendricks, Erin Grant, and Eric Tzeng – as well as postdocs – like Anna Rohrbach and Aida Nematzadeh – has been absolutely essential to my progress. Thanks to all of you and to the Berkeley AI research community for championing my success.

I was also extremely fortunate to be a part of the unique teaching community at Berkeley. Professors Paul Hilfinger, Josh Hug, Satish Rao, and Kannan Ramchandran taught me the importance of sharing your knowledge with others. I'm extremely grateful to them, my fellow TAs, my early mentors – Sarah Kim and Alan Yao, and the amazing Student Affairs Staff who make the teaching process run smoothly.

And of course, I was lucky to have great companions by my side. I'd like to thank Adora for teaching me how to laugh at myself, Michelle for demonstrating the power of resilience, Lyn for being so forthcoming with her culinary expertise, Gary for his constant support, and the many other friends who made my college experience unforgettable.

Finally, I'm extremely thankful for my family's persistent love and support. This report is dedicated to you.

Chapter 1


Introduction

Aligning vision and language is an important step in many applications; whether it's enabling the visually impaired to navigate the world through natural language or providing a familiar interface to otherwise opaque computational systems, the field is ripe with promise. Some of the largest roadblocks to realizing integrated vision and language systems, such as image captioning, are prediction artifacts from the training process and data. In the next two chapters we will discuss two weaknesses of captioning systems: the exaggeration of dataset bias related to gender presentation and the “hallucination” of objects that are not visually present in the scene. The first chapter will focus on a solution to the issue of bias related to gender presentation and the second will analyze possible causes of “hallucinations”.

Both of the following chapters were joint works with Lisa Anne Hendricks, Anna Rohrbach, Kate Saenko, and Trevor Darrell [23, 45].

Chapter 2

Overcoming Bias in Captioning Models

Most machine learning methods are known to capture and exploit biases of the training data. While some biases are beneficial for learning, others are harmful. Specifically, image captioning models tend to exaggerate biases present in training data (e.g., if a word is present in 60% of training sentences, it might be predicted in 70% of sentences at test time). This can lead to incorrect captions in domains where unbiased captions are desired, or required, due to over-reliance on the learned prior and image context. In this work we investigate generation of gender-specific caption words (e.g. man, woman) based on the person’s appearance or the image context. We introduce a new *Equalizer* model that encourages equal gender probability when gender evidence is occluded in a scene and confident predictions when gender evidence is present. The resulting model is forced to look at a person rather than use contextual cues to make a gender-specific prediction. The losses that comprise our model, the *Appearance Confusion Loss* and the *Confident Loss*, are general, and can be added to any description model in order to mitigate impacts of unwanted bias in a description dataset. Our proposed model has lower error than prior work when describing images with people and mentioning their gender and more closely matches the ground truth ratio of sentences including women to sentences including men. Finally, we show that our model more often looks at people when predicting their gender. 

¹This chapter is based on joint work with Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach presented at ECCV 2018 [\[23\]](#). Lisa Anne Hendricks also led the paper. https://people.eecs.berkeley.edu/~lisa_anne/snowboard.html

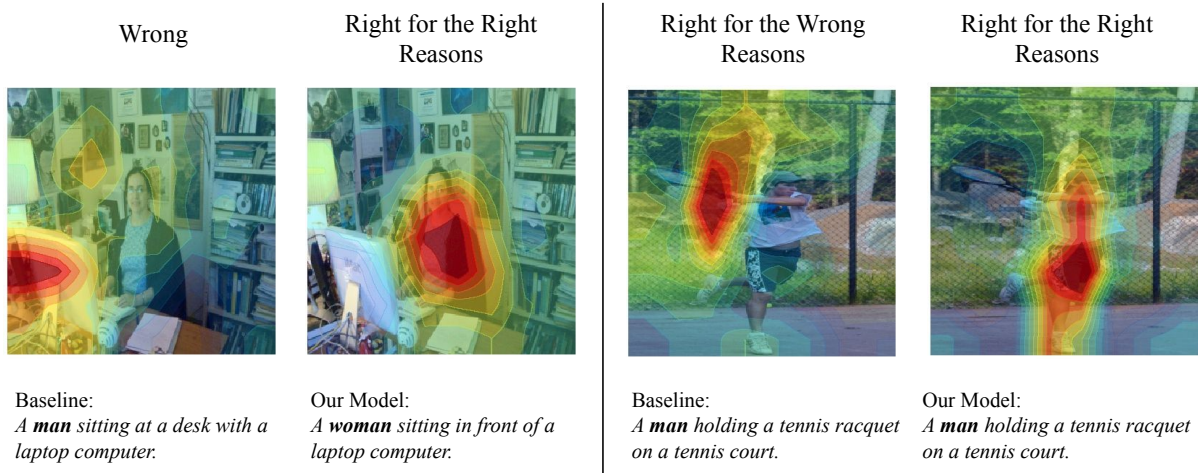


Figure 2.1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender word correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

2.1 Motivation

Exploiting contextual cues can frequently lead to better performance on computer vision tasks [55, 54, 17]. For example, in the visual description task, predicting a “mouse” might be easier given that a computer is also in the image. However, in some cases making decisions based on context can lead to incorrect, and perhaps even offensive, predictions. In this work, we consider one such scenario: generating captions about men and women. We posit that when description models predict gendered words such as “man” or “woman”, they should consider visual evidence associated with the described person, and not contextual cues like location (e.g., “kitchen”) or other objects in a scene (e.g., “snowboard”). Not only is it important for description systems to avoid egregious errors (e.g., always predicting the word “man” in snowboarding scenes), but it is also important for predictions to be right for the right reason. For example, Figure 2.1 (left) shows a case where prior work predicts the incorrect gender, while our model accurately predicts the gender presentation by considering the correct gender evidence. Figure 2.1 (right) shows an example where both models predict the correct gender, but prior work does not look at the person when describing the image (it is right for the wrong reasons).

Bias in image captioning is particularly challenging to overcome because of the multi-modal nature of the task; predicted words are not only influenced by an image, but also

biased by the learned language model. Though [68] studied bias for structured prediction tasks (e.g., semantic role labeling), they did not consider the task of image captioning. Furthermore, the solution proposed in [68] requires access to the entire test set in order to rebalance gender predictions to reflect the distribution in the training set. Consequently, [68] relies on the assumption that the distribution of genders is the same at training and test time. We make no such assumptions; we consider a more realistic scenario in which captions are generated for images independent of other test images.

In order to encourage description models to generate less biased captions, we introduce the *Equalizer* Model. Our model includes two complementary loss terms: the *Appearance Confusion Loss (ACL)* and the *Confident Loss (Conf)*. The Appearance Confusion Loss is based on the intuition that, given an image in which evidence of gender is absent, description models should be unable to accurately predict a gendered word. However, it is not enough to confuse the model when gender evidence is absent; we must also encourage the model to consider gender evidence when it is present. Our Confident Loss helps to increase the model’s confidence when gender is in the image. These complementary losses allow the Equalizer model to be cautious in the absence of gender information and discriminative in its presence.

Our proposed Equalizer model leads to less biased captions: not only does it lead to lower error when predicting gendered words, but it also performs well when the distribution of gender presentation in the test set is not aligned with the training set. Additionally, we observe that Equalizer generates gender neutral words (like “person”) when it is not confident of the gender. Furthermore, we demonstrate that Equalizer focuses on humans when predicting gender words, as opposed to focusing on other image context.

2.2 Related Work

Unwanted Dataset Bias.

Unwanted dataset biases (e.g., gender, ethnic biases) have been studied across a wide variety of AI domains [48, 51, 6, 7, 5, 38]. One common theme is the notion of *bias amplification*, in which bias is not only learned, but amplified [68, 6, 51]. For example, in the image captioning scenario, if 70% of images with umbrellas include a woman and 30% include a man, at test time the model might amplify this bias to 85% and 15%. Eliminating bias amplification is not as simple as balancing across attributes for a specific category. [51] study bias in classification and find that even though white and black people appear in “basketball” images with similar frequency, models learn to classify images as “basketball” based on the presence of a black person. One explanation is that though the data is balanced in regard to the class “basketball”, there are many more white people in the dataset. Consequently, to perfectly balance a dataset, one would have to balance across all possible co-occurrences which is infeasible.

Natural language data is subject to *reporting bias* [6, 18, 37, 36] in which people over-report less common co-occurrences, such as “male nurse” [6] or “green banana” [37]. [36] also discuss how visual descriptions reflect cultural biases (e.g., assuming a woman with a child is a mother, even though this cannot be confirmed in an image). We observe that annotators specify gender even when gender cannot be confirmed in an image (e.g., a snowboarder might be labeled as “man” even if gender evidence is occluded).

Our work is most similar to [68] who consider bias in semantic role labeling and multilabel classification (as opposed to image captioning). To avoid bias amplification, [68] rebalance the test time predictions to more accurately reflect the training time word ratios. This solution is unsatisfactory because (i) it requires access to the entire test set and (ii) it assumes that the distribution of objects at test time is the same as at training time. We consider a more realistic scenario in our experiments, and show that the ratio of woman to man in our predicted sentences closely resembles the ratio in ground truth sentences, even when the test distribution is different from the training distribution.

Fairness.

Building AI systems which treat *protected attributes* (e.g., age, gender, sexual orientation) in a fair manner is increasingly important [20, 13, 64, 40]. In the machine learning literature, “fairness” generally requires that systems do not use information such as gender or age in a way that disadvantages one group over another. We consider a different scenario as we are trying to *predict* protected attributes.

Distribution matching has been used to build fair systems [40] by encouraging the distribution of decisions to be similar across different protected classes, as well as for other applications such as domain adaption [56, 67] and transduction learning [39]. Our Appear-

ance Confusion Loss is similar as it encourages the distribution of predictions to be similar for man and woman classes when gender information is not available.

Right for the Right Reasons.

Assuring models are “right for the right reasons,” or consider similar evidence as humans when making decisions, helps researchers understand how models will perform in real world applications (e.g., when predicting outcomes for pneumonia patients in [9]) or discover underlying dataset bias [53]. We hypothesize that models which look at appropriate gender evidence will perform better in new scenarios, specifically when the gender distribution at test and training time are different.

Recently, [47] develop a loss function which compares explanations for a decision to ground truth explanations. However, [47] generating explanations for visual decisions is a difficult and active area of research [41, 49, 15, 43, 69, 63]. Instead of relying on our model to accurately explain itself during training, we verify that our formulation encourages models to be right for the right reason at test time.

Visual Description.

Most visual description work (e.g., [58, 12, 25, 60, 1]) focuses on improving overall sentence quality, without regard to captured biases. Though we pay special attention to gender in this work, all captioning models trained on visual description data (MSCOCO [32], Flickr30k [62], MSR-VTT [59] to name a few) implicitly learn to classify gender. However current captioning models do not discuss gender the way humans do, but *amplify* gender bias; our intent is to generate descriptions which more accurately reflect human descriptions when discussing this important category.

Gender Classification.

Gender classification models frequently focus on facial features [30, 66, 14]. In contrast, we are mainly concerned about whether contextual clues in complex scenes bias the production of gendered words during sentence generation. Gender classification has also been studied in natural language processing ([3, 61], [8]).

Ethical Considerations.

Frequently, gender classification is seen as a binary task: data points are labeled as either “man” or “woman”. However, AI practitioners, both in industrial² and academic³ settings, are increasingly concerned that gender classification systems should be inclusive.

²<https://clarifai.com/blog/socially-responsible-pixels-a-look-inside-clarifais-new-demographics-recognition-model>

³<https://www.media.mit.edu/projects/gender-shades/faq>

Our captioning model predicts three gender categories: male, female, and gender neutral (e.g., person) based on visual appearance. When designing gender classification systems, it is important to understand where labels are sourced from [28]. We determine gender labels using a previously collected, publicly released dataset in which annotators describe images [32]. Importantly, people in the images are not asked to identify their gender. Thus, we emphasize that we are not classifying biological sex or gender identity, but rather gender expression.

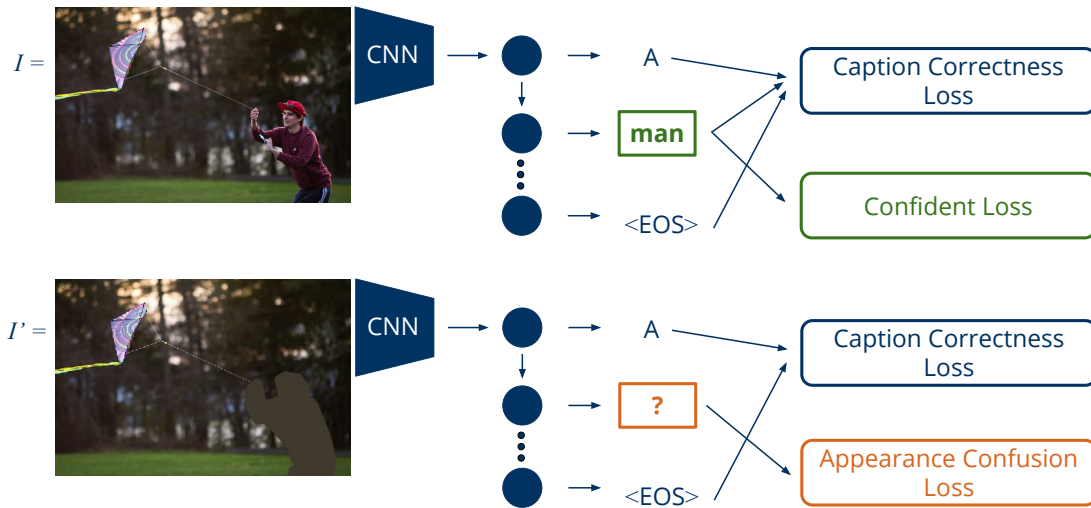


Figure 2.2: Equalizer includes two novel loss terms: the Confident Loss on images with men or women (top) and the Appearance Confusion Loss on images where men and women are occluded (bottom). Together these losses encourage our model to make correct predictions when evidence of gender is present, and be cautious in its absence. We also include the Caption Correctness Loss (cross entropy loss) for both image types.

2.3 Equalizer: Overcoming Bias in Description Models

Equalizer is based on the following intuitions: if evidence to support a specific gender decision is not present in an image, the model should be *confused* about which gender to predict (enforced by an Appearance Confusion Loss term), and if evidence to support a gender decision is in an image, the model should be *confident* in its prediction (enforced by a Confident Loss term). To train our model we require not only pairs of images, I , and sentences, S , but also annotation masks M which indicate which evidence in an image is appropriate for determining gender. Though we use [58] as our base network, Equalizer is general and can be integrated into any deep description frameworks.

Background: Description Framework

To generate a description, high level image features are first extracted from the InceptionV3 [52] model. The image features are then used to initialize an LSTM hidden state. To begin sentence generation, a start of sentence token is input into the LSTM. For each subsequent time step during training, the ground truth word w_t is input into the LSTM. At test time, the previously predicted word w_{t-1} is input into the LSTM at each time step. Generation concludes when an end of sequence token is generated. Like [58], we include the standard

cross entropy loss (\mathcal{L}^{CE}) during training:

$$\mathcal{L}^{CE} = -\frac{1}{N} \sum_{n=0}^N \sum_{t=0}^T \log(p(w_t|w_{0:t-1}, I)), \quad (2.1)$$

where N is the batch size, T is the number of words in the sentence, w_t is a ground truth word at time t , and I is an image.

Appearance Confusion Loss

Our Appearance Confusion Loss encourages the underlying description model to be *confused* when making gender decisions if the input image does not contain appropriate evidence for the decision. To optimize the Appearance Confusion Loss, we require ground truth rationales indicating which evidence is appropriate for a particular gender decision. We expect the resulting rationales to be masks, M , which are 1 for pixels which should not contribute to a gender decision and 0 for pixels which are appropriate to consider when determining gender. The Hadamard product of the mask and the original image, $I \odot M$, yields a new image, I' , with gender information that the implementer deems appropriate for classification removed. Intuitively, for an image devoid of gender information, the probability of predicting man or woman should be equal. The Appearance Confusion Loss enforces a fair prior by asserting that this is the case.

To define our Appearance Confusion Loss, we first define a *confusion* function (\mathcal{C}) which operates over the predicted distribution of words $p(\tilde{w}_t)$, a set of woman gender words (\mathcal{G}_w), and a set of man gender words (\mathcal{G}_m):

$$\mathcal{C}(\tilde{w}_t, I') = \left| \sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I') \right|. \quad (2.2)$$

In practice, the \mathcal{G}_w consists only of the word “woman” and, likewise, the \mathcal{G}_m consists only of the word “man”. These are by far the most commonly used gender words in the datasets we consider and we find that using these “sets” results in similar performance as using more complete sets.

We can now define our Appearance Confusion Loss (\mathcal{L}^{AC}) as:

$$\mathcal{L}^{AC} = \frac{1}{N} \sum_{n=0}^N \sum_{t=0}^T \mathbb{1}(w_t \in \mathcal{G}_w \cup \mathcal{G}_m) \mathcal{C}(\tilde{w}_t, I'), \quad (2.3)$$

where $\mathbb{1}$ is an indicator variable that denotes whether or not w_t is a gendered word.

For the remaining non-gendered words that correspond to images I' , we apply the standard cross entropy loss to encourage the model to discuss objects which are still visible in I' . In addition to encouraging sentences to be image relevant even when the gender information has been removed, this also encourages the model to learn representations of words like “dog” and “frisbee” that are not reliant on gender information.

Confident Loss

In addition to being unsure when gender evidence is occluded, we also encourage our model to be confident when gender evidence is present. Thus, we introduce the Confident Loss term, which encourages the model to predict gender words correctly.

Our Confident Loss encourages the probabilities for predicted gender words to be high on images I in which gender information is present. Given functions \mathcal{F}^W and \mathcal{F}^M which measure how confidently the model predicts woman and man words respectively, we can write the Confident Loss as:

$$\mathcal{L}^{Con} = \frac{1}{N} \sum_{n=0}^N \sum_{t=0}^T (\mathbb{1}(w_t \in \mathcal{G}_w) \mathcal{F}^W(\tilde{w}_t, I) + \mathbb{1}(w_t \in \mathcal{G}_m) \mathcal{F}^M(\tilde{w}_t, I)). \quad (2.4)$$

To measure the confidence of predicted gender words, we consider the quotient between predicted probabilities for man and gender words (\mathcal{F}^M is of the same form):

$$\mathcal{F}^W(\tilde{w}_t, I) = \frac{\sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I)}{(\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I)) + \epsilon} \quad (2.5)$$

where ϵ is a small epsilon value added for numerical stability.

When the model is confident of a gender prediction (e.g., for the word “woman”), the probability of the word “woman” should be considerably higher than the probability of the word “man”, which will result in a small value for \mathcal{F}^W and thus a small loss. One nice property of considering the quotient between predicted probabilities is that we encourage the model to distinguish between gendered words without forcing the model to predict a gendered word. For example, if the model predicts a probability of 0.2 for “man”, 0.5 for “woman”, and 0.3 for “person” on a “woman” image, our confidence loss will be low. However, the model is still able to predict gender neutral words, like “person” with relatively high probability. This is distinct from other possible losses, like placing a larger weight on gender words in the cross entropy loss, which forces the model to predict “man”/“woman” words and penalizes the gender neutral words.

The Equalizer Model

Our final model is a linear combination of all aforementioned losses:

$$\mathcal{L} = \alpha \mathcal{L}^{CE} + \beta \mathcal{L}^{AC} + \mu \mathcal{L}^{Con}, \quad (2.6)$$

where α , β , and μ are hyperparameters chosen on a validation set ($\alpha, \mu = 1$, $\beta = 10$ in our experiments).

Our Equalizer method is general and our base captioning framework can be substituted with any other deep captioning framework. By combining all of these terms, the Equalizer model can not only generate image relevant sentences, but also make confident gender predictions under sufficient evidence. We find that both the Appearance Confusion Loss and

the Confident Loss are important in creating a confident yet cautious model. Interestingly, the Equalizer model achieves the lowest misclassification rate only when these two losses are combined, highlighting the complementary nature of these two loss terms.

2.4 Experiments

Datasets

MSCOCO-Bias.

To evaluate our method, we consider the dataset used by [68] for evaluating bias amplification in structured prediction problems. This dataset consists of images from MSCOCO [32] which are labeled as “man” or “woman”. Though “person” is an MSCOCO class, “man” and “woman” are not, so [68] employ ground truth captions to determine if images contain a man or a woman. Images are labeled as “man” if at least one description includes the word “man” and no descriptions include the word “woman”. Likewise, images are labeled as “woman” if at least one description includes the word “woman” and no descriptions include the word “man”. Images are discarded if both “man” and “woman” are mentioned. We refer to this dataset as MSCOCO-Bias.

MSCOCO-Balanced.

We also evaluate on a set where we purposely change the gender ratio. We believe this is representative of real world scenarios in which different distributions of men and women might be present at test time. The MSCOCO-Bias set has a roughly 1:3 woman to man ratio where as this set, called MSCOCO-Balanced, has a 1:1 woman to man ratio. We randomly select 500 images from MSCOCO-Bias set which include the word “woman” and 500 which include “man”.

Person Masks.

To train Equalizer, we need ground truth human rationales for why a person should be predicted as a man or a woman. We use the person segmentation masks from the MSCOCO dataset. Once the masked image is created, we fill the segmentation mask with the average pixel value in the image. We use the masks both at training time to compute Appearance Confusion Loss and during evaluation to ensure that models are predicting gender words by looking at the person. While for MSCOCO the person annotations are readily available, for other datasets e.g. a person detector could be used.

Metrics

To evaluate our methods, we rely on the following metrics.

Error. Due to the sensitive nature of prediction for protected classes (gender words in our scenario), we emphasize the importance of a low error. The error rate is the number of man/woman misclassifications, while gender neutral terms are not considered errors. We expect that the best model would rather predict gender neutral words in cases where gender is not obvious.

Gender Ratio. Second, we consider the ratio of sentences which belong to a “woman” set to sentences which belong to a “man” set. We consider a sentence to fall in a “woman” set if it predicts any word from a precompiled list of female gendered words, and respectively fall in a “man” set if it predicts any word from a precompiled list of male gendered words.

Right for the Right Reasons. Finally, to measure if a model is “right for the right reasons” we consider the pointing game [65] evaluation. We first create visual explanations for “woman”/“man” using the Grad-CAM approach [49] as well as saliency maps created by occluding image regions in a sliding window fashion. To measure if our models are right for the right reason, we verify whether the point with the highest activation in the explanation heat map falls in the person segmentation mask.

Training Details

All models are initialized from the Show and Tell model [58] pre-trained on all of MSCOCO for 1 million iterations (without fine-tuning through the visual representation). Models are trained for additional 500,000 iterations on the MSCOCO-Bias set, fine-tuning through the visual representation (Inception v3 [52]) for 500,000 iterations.

Baselines and Ablations

Baseline-FT.

The simplest baseline is fine-tuning the Show and Tell model through the LSTM and convolutional networks using the standard cross-entropy loss on our target dataset, the MSCOCO-Bias dataset.

Balanced.

We train a Balanced baseline in which we re-balance the data distribution at training time to account for the larger number of men instances in the training data. Even though we cannot know the correct distribution of our data at test time, we can enforce our belief that predicting a woman or man should be equally likely. At training time, we re-sample the images of women so that the number of training examples of women is the same as the number of training examples of men.

UpWeight.

We also experiment with upweighting the loss value for gender words in the standard cross entropy loss to increase the penalty for a misclassification. For each time step where the ground truth caption says the word “man” or “woman”, we multiply that term in the loss by a constant value (10 in reported experiments). Intuitively, upweighting should encourage the models to accurately predict gender words. However, unlike our Confident Loss, upweighting

Model	MSCOCO-Bias		MSCOCO-Balanced	
	Error	Ratio Δ	Error	Ratio Δ
Baseline-FT	12.83	0.15	19.30	0.51
Balanced	12.85	0.14	18.30	0.47
UpWeight	13.56	0.08	16.30	0.35
Equalizer w/o ACL	7.57	0.04	10.10	0.26
Equalizer w/o Conf	9.62	0.09	13.90	0.40
Equalizer	7.02	-0.03	8.10	0.13

Table 2.1: Evaluation of predicted gender words based on error rate and ratio of generated sentences which include the “woman” words to sentences which include the “man” words. Equalizer achieves the lowest error rate and predicts sentences with a gender ratio most similar to the corresponding ground truth captions (Ratio Δ), even when the test set has a different distribution of gender words than the training set, as is the case for the MSCOCO-Balanced dataset.

drives the model to make either “man” or “woman” predictions without the opportunity to place a high probability on gender neutral words.

Ablations.

To isolate the impact of the two loss terms in Equalizer, we report results with only the Appearance Confusion Loss (Equalizer w/o Conf) and only the Confidence Loss (Equalizer w/o ACL). We then report results of our full Equalizer model.

Results

Error.

Table 2.1 reports the error rates when describing men and women on the MSCOCO-Bias and MSCOCO-Balanced test sets. Comparing to baselines, Equalizer shows consistent improvements. Importantly, our full model consistently improves upon Equalizer w/o ACL and Equalizer w/o Conf. When comparing Equalizer to baselines, we see a larger performance gain on the MSCOCO-Balanced dataset. As discussed later, this is in part because our model does a particularly good job of decreasing error on the minority class (woman). Unlike baseline models, our model has a similar error rate on each set. This indicates that the error rate of our model is not as sensitive to shifts in the gender distribution at test time.

Interestingly, the results of the Baseline-FT model and Balanced model are not substantially different. One possibility is that the co-occurrences across words are not balanced (e.g., if there is gender imbalance specifically for images with “umbrella” just balancing the dataset based on gender word counts is not sufficient to balance the dataset). We empha-

Model	Women			Men			Outcome Divergence between Genders
	Correct	Incorrect	Other	Correct	Incorrect	Other	
Baseline-FT	46.28	34.11	19.61	75.05	4.23	20.72	0.121
Balanced	47.67	33.80	18.54	75.89	4.38	19.72	0.116
UpWeight	60.59	29.82	9.58	87.84	6.98	5.17	0.078
Equalizer w/o ACL	56.18	16.02	27.81	67.58	4.15	28.26	0.031
Equalizer w/o Conf	46.03	24.84	29.13	61.11	3.47	35.42	0.075
Equalizer (Ours)	57.38	12.99	29.63	59.02	4.61	36.37	0.018

Table 2.2: Accuracy per class for MSCOCO-Bias dataset. Though UpWeight achieves the highest recall for both men and women images, it also has a high error, especially for women. One criterion of a “fair” system is that it has similar outcomes across classes. We measure outcome similarity by computing the Jensen-Shannon divergence between Correct/Incorrect/Other sentences for men and women images (lower is better) and observe that Equalizer performs best on this metric.

size that balancing across all co-occurring words is difficult in large-scale settings with large vocabularies.

Gender Ratio

We also consider the ratio of captions which include only female words to captions which include only male words. In Table 2.1 we report the *difference* between the ground truth ratio and the ratio produced by each captioning model. Again, the ACL and Confident losses are complementary and Equalizer has the best overall performance.

Performance for Each Gender.

Images with females comprise a much smaller portion of MSCOCO than images with males. Therefore the overall performance across classes (i.e. man, woman) can be misleading because it downplays the errors in the minority class. Additionally, unlike [68] who consider a classification scenario in which the model is forced to predict a gender, our description models can also discuss gender neutral terms such as “person” or “player”. In Table 2.2 for each gender, we report the percentage of sentences in which gender is predicted correctly or incorrectly and when no gender specific word is generated on the MSCOCO-Bias set.

Across all models, the error for Men is quite low. However, our model significantly improves the error for the minority class, Women. Interestingly, we observe that Equalizer has a similar recall (Correct), error (Incorrect), and Other rate across both genders. A caption model could be considered more “fair” if, for each gender, the possible outcomes (correct gender mentioned, incorrect gender mentioned, gender neutral) are similar. This resembles the notion of equalized odds in fairness literature [20], which requires a system to have similar false positive and false negative rates across groups. To formalize this notion of fairness in our captioning systems, we report the outcome type divergence between genders by

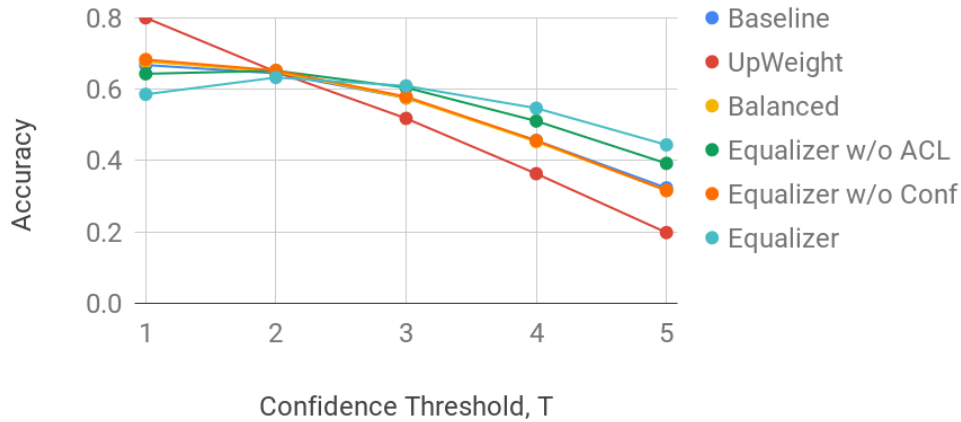


Figure 2.3: Accuracy across man, woman, and gender neutral terms for different models as a function of annotator confidence. When only one annotator describes an image with a gendered word, Equalizer has a low accuracy as it more likely predicts gender neutral words but when more annotations mention gendered words, Equalizer has higher accuracy than other models.

measuring the Jensen-Shannon [31] divergence between Correct/Incorrect/Other outcomes for Men and Women. Lower divergence indicates that Women and Men classes result in a similar distribution of outcomes, and thus the model can be considered more “fair”. Equalizer has the lowest divergence (0.018).

Annotator Confidence.

As described above, gender labels are mined from captions provided in the MSCOCO dataset. Each image corresponds to five captions, but not all captions for a single image include a gendered word. Counting the number of sentences which include a gendered word provides a rough estimate of how apparent gender is in an image and how important it is to mention when describing the scene.

To understand how well our model captures the way annotators describe people, instead of labeling images as either “man” or “woman”, we label images as “man”, “woman”, or “gender neutral” based on how many annotators mentioned gender in their description. For a specific threshold value T , we consider an image to belong to the “man” or “woman” class if T or more annotators mention the gender in their description, and “gender neutral” otherwise. We can then measure accuracy over these three classes. Whereas a naive solution which restricts vocabulary to include no gender words would have low error as defined in Table 2.1, it would not capture the way humans use gender words when describing images. Indeed, the MSCOCO training set includes over 200,000 instances of words which describe people. Over half of all words used to describe people are gendered. By considering accuracy

across three classes, we can better measure how well models capture the way humans describe gender.

Figure 2.3 plots the accuracy of each model with respect to the confidence threshold T . At low threshold values, Equalizer performs worse as it tends to more frequently output gender neutral terms, and the UpWeight model, which almost always predicts gendered words, performs best. However, as the threshold value increases, Equalizer performs better than other models, including at a threshold value of 3 which corresponds to classifying images based off the majority vote. This indicates that Equalizer naturally captures when humans describe images with gendered or gender neutral words.

Object Gender Co-Occurrence.

We analyze how gender prediction influences prediction of other words on the MSCOCO-Bias test set. Specifically, we consider the 80 MSCOCO categories, excluding the category “person”. We adopt the bias amplification metric proposed in [68], and compute the following ratios: $\frac{\text{count}(\text{man}\&\text{object})}{\text{count}(\text{person}\&\text{object})}$ and $\frac{\text{count}(\text{woman}\&\text{object})}{\text{count}(\text{person}\&\text{object})}$, where *man* refers to all male words, *woman* refers to all female words, and *person* refers to all male, female, or gender neutral words. Ideally, these ratios should be similar for generated captions and ground truth captions. However, e.g. for *man* and *motorcycle*, the ground truth ratio is 0.40 and for the Baseline-FT and Equalizer, the ratio is 0.81 and 0.65, respectively. Though Equalizer over-predicts this pair, the ratio is closer to the ground truth than when comparing Baseline-FT to the ground truth. Likewise, for *woman* and *umbrella*, the ground truth ratio is 0.40, Baseline-FT ratio is 0.64, and Equalizer ratio is 0.56. As a more holistic metric, we average the *difference* of ratios between ground truth and generated captions across objects (lower is better). For male words, Equalizer is substantially better than the Baseline-FT (0.147 vs. 0.193) and similar for female words (0.096 vs. 0.99).

Caption Quality.

Qualitatively, the sentences from all of our models are linguistically fluent (indeed, comparing sentences in Figure 2.4 we note that usually only the word referring to the person changes). However, we do notice a small drop in performance on standard description metrics (25.2 to 24.3 on METEOR [29] when comparing Baseline-FT to our full Equalizer) on MSCOCO-Bias. One possibility is that our model is overly cautious and is penalized for producing gender neutral terms for sentences that humans describe with gendered terms.

Right for the Right Reasons.

We hypothesize that many misclassification errors occur due to the model looking at the wrong visual evidence, e.g. conditioning gender prediction on context rather than on the person’s appearance. We quantitatively confirm this hypothesis and show that our proposed model improves this behavior by looking at the appropriate evidence, i.e. is being “right for

Accuracy	Woman	Man	All	Accuracy	Woman	Man	All
Random	22.6	19.5	21.0	Random	25.1	17.5	21.3
Baseline-FT	39.8	34.3	37.0	Baseline-FT	45.3	40.4	42.8
Balanced	37.6	34.1	35.8	Balanced	48.5	42.2	45.3
UpWeight	43.3	36.4	39.9	UpWeight	54.1	45.5	49.8
Equalizer w/o ACL	48.1	39.6	43.8	Equalizer w/o ACL	54.7	47.5	51.1
Equalizer w/o Conf	43.9	36.8	40.4	Equalizer w/o Conf	48.9	46.7	47.8
Equalizer (Ours)	49.9	45.2	47.5	Equalizer (Ours)	56.3	51.1	53.7

(a) Visual explanation is a *Grad-CAM* map.(b) Visual explanation is a *saliency* map.

Table 2.3: *Pointing game* evaluation that measures whether the visual explanations for “man” / “woman” words fall in the person segmentation ground-truth. Evaluation is done for ground-truth captions on the MSCOCO-Balanced.

the right reasons”. To evaluate this we rely on two visual explanation techniques: Grad-CAM [49] and saliency maps generated by occluding image regions in a sliding window fashion.

Unlike [49] who apply Grad-CAM to an entire caption, we visualize the evidence for generating specific words, i.e. “man” and “woman”. Specifically, we apply Grad-CAM to the last convolutional layer of our image processing network, InceptionV3 [52], we obtain 8x8 weight matrices. To obtain saliency maps, we resize an input image to 299×299 and uniformly divide it into 32×32 pixel regions, obtaining a 10×10 grid (the bottom/rightmost cells being smaller). Next, for every cell in the grid, we zero out the respective pixels and feed the obtained “partially blocked out” image through the captioning network (similar to as was done in the occlusion sensitivity experiments in [63]). Then, for the ground-truth caption, we compute the “information loss”, i.e. the decrease in predicting the words “man” and “woman” as $-\log(p(w_t = g_m))$ and $-\log(p(w_t = g_w))$, respectively. This is similar to the top-down saliency approach of [41], who zero-out all the intermediate feature descriptors but one.

To evaluate whether the visual explanation for the predicted word is focused on a person, we rely on person masks, obtained from MSCOCO ground-truth person segmentations. We use the *pointing game* evaluation [65]. We upscale visual explanations to the original image size. We define a “hit” to be when the point with the highest weight is contained in the person mask. The accuracy is computed as $\frac{\#hits}{\#hits+\#misses}$.

Results on the MSCOCO-Balanced set are presented in Table 2.3 (a) and (b), for the Grad-CAM and saliency maps, respectively. For a fair comparison we provide all models with ground-truth captions. For completeness we also report the random baseline, where the point with the highest weight is selected randomly. We see that Equalizer obtains the best accuracy, significantly improving over the Baseline-FT and all model variants. A similar evaluation on the actual generated captions shows the same trends.

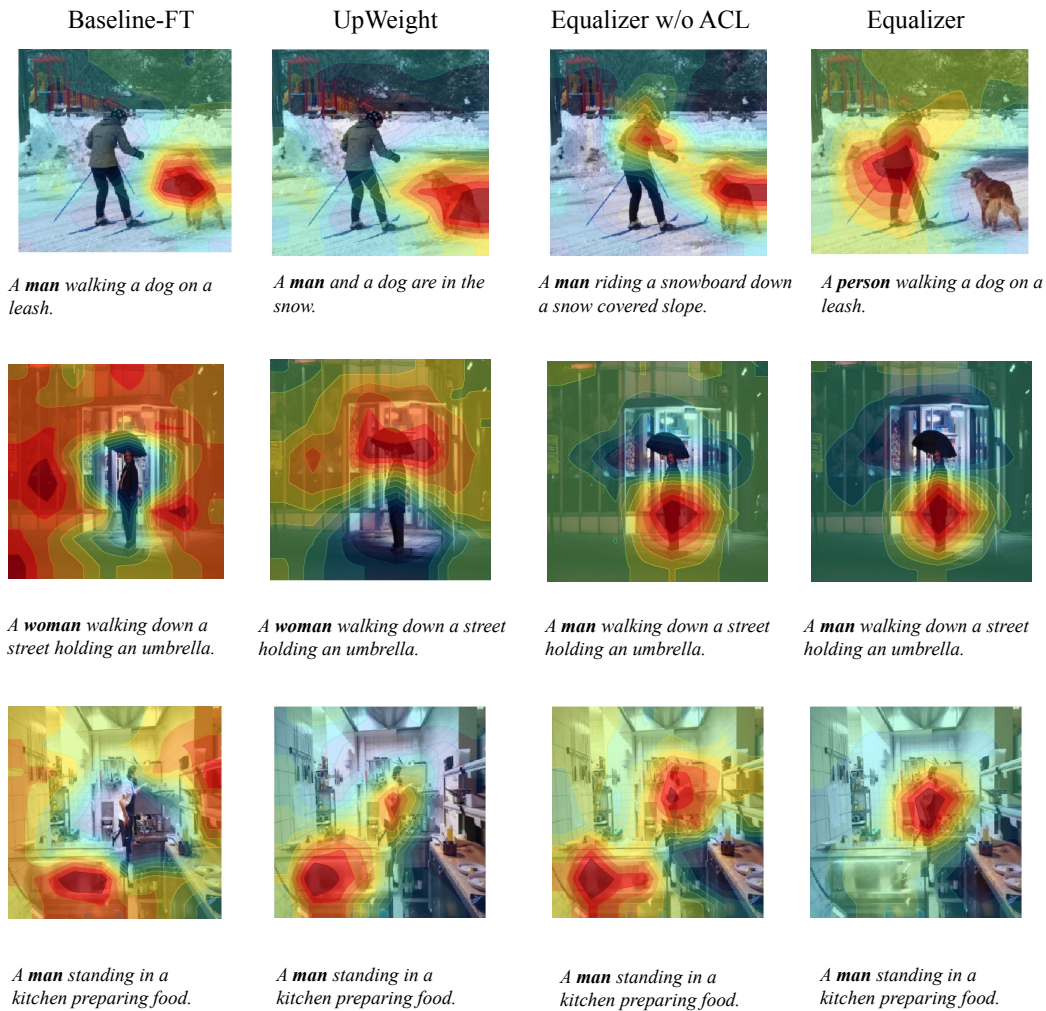


Figure 2.4: Qualitative comparison of multiple baselines and our model. In the top example, being conservative (“person”) is better than being wrong (“man”) as the gender is not obvious. In the bottom example the baselines are looking at the wrong visual evidence.

Looking at objects.

Using our pointing technique, we can also analyze which MSCOCO objects models are “looking” at when they *do not* point at the person while predicting “man”/“woman”. Specifically, we count a hit if the highest activation is on an object in question. We compute the following ratio for each gender: number of images where an object is pointed at to the true number of images with that object. We find that there are differences across genders, e.g. “umbrella”, “bench”, “suitcase” are more often pointed at when discussing women, while

e.g. “truck”, “couch”, “pizza” when discussing men. Our model reduces the overall “delta” between genders for ground truth sentences from an average 0.12 to 0.08, compared to the Baseline-FT. E.g. for “dining table” Equalizer decreases the delta from 0.07 to 0.03.

Performance breakdown for biased words.

We additionally analyze objects which co-occur with one gender more than the other. For a careful analysis, we choose five words that are biased to co-occur with women (umbrella, kitchen, cell phone, table, and food) and five words which frequently co-occur with men (skateboard, baseball, tie, motorcycle, and snowboard). To choose biased words, we compute bias as is done in [68] Section 3 for the most commonly occurring nouns (> 250 times) in the MSCOCO-Bias training set. We compute the error rate and the *difference* between the ground truth ratio of women to men and the ratio produced by each captioning model, for images containing the above objects (Table 2.4). We observe similar trends to our observations in the main paper. Equalizer and Equalizer w/o ACL have the lowest errors, with Equalizer w/o ACL performing slightly better, suggesting the confidence term is important for low error rate. Considering distance to the ground truth gender ratio, the Equalizer model consistently outperforms other models. One particularly interesting case study is the word “kitchen” in which the ground truth woman to man gender ratio is 0.946 (recall that the dataset contains a roughly 1:3 woman to man gender ratio, so a gender ratio close to 1.0 for a specific object suggests that a higher proportion of “woman” images include a “kitchen” than “man” images). The Equalizer model predicts a gender ratio of 1.0 (delta 0.054) whereas the next best model (Equalizer w/o ACL) predicts a gender ratio of 0.806 (delta 0.14). The Baseline-FT model predicts a ratio of 0.586 (delta 0.361).

Masked Images.

We also consider the gender ratio when predicting sentences for masked images (images in the test set are masked in the same way as was done to train the Appearance Confusion Loss term). Ideally, the ratio of predicted gender words should be close to 1.0 on the masked images as gender information is obscured. The man to woman gender ratios for the Equalizer w/o ACL, Equalizer w/o Conf, and Equalizer are 3.45, 2.87, and 1.98 respectively (all other models have larger ratios than Equalizer w/o ACL). This suggests that again both our ACL loss and Conf loss are important for predicting a fair gender ratio when gender information is not present in the image. Again, we achieve the best performance with our full Equalizer model.

Qualitative Results.

Figure 2.4 compares Grad-CAM visualizations for predicted gender words from our model to the Baseline-FT, UpWeight, and Equalizer w/o ACL. We consistently see that our model looks at the person when describing gendered words. In Figure 2.4 (top), all other models look at the dog rather than the person and predict the gender “man” (ground truth label is

	Umbrella	Kitchen	Cell Phone	Table	Food	Skateboard	Baseball	Tie	Motorcycle	Snowboard
Error										
Baseline-FT	0.303	0.277	0.172	0.200	0.154	0.028	0.072	0.017	0.083	0.073
Equalizer w/o ACL	0.210	0.157	0.145	0.100	0.085	0.020	0.085	0.021	0.054	0.031
Equalizer w/o Conf	0.250	0.269	0.158	0.189	0.166	0.028	0.038	0.017	0.107	0.100
Equalizer	0.176	0.181	0.119	0.119	0.128	0.031	0.113	0.011	0.064	0.081
Δ Ratio (Women:Men)										
Baseline-FT	1.074	0.358	0.278	0.351	0.213	0.011	0.004	0.013	0.103	0.094
Equalizer w/o ACL	1.274	0.137	0.028	0.151	0.092	0.009	0.018	0.002	0.086	0.051
Equalizer w/o Conf	1.931	0.291	0.212	0.275	0.127	0.008	0.023	0.015	0.084	0.081
Equalizer	2.335	0.057	0.009	0.131	0.031	0.001	0.046	0.008	0.077	0.033

Table 2.4: Breakdown of error rate and difference to ground-truth woman:man ratio over images with specific biased words. We see that the full Equalizer generally outperforms the Baseline-FT. On error, Equalizer w/o ACL performs best, followed by Equalizer. Equalizer performs best when considering predicted gender ratio.

“woman”). In this particular example, the gender is somewhat ambiguous, and our model conservatively predicts “person” rather than misclassify the gender. In Figure 2.4 (middle), the Baseline-FT and UpWeight example both incorrectly predict the word “woman” and do not look at the person (women occur more frequently with umbrellas). In contrast, both the Equalizer w/o ACL and the Equalizer look at the person and predict the correct gender. Finally, in Figure 2.4 (bottom), all models predict the correct gender (man), but our model is the only model which looks at the person and is thus “right for the right reasons.”

In Figure 2.5 we provide multiple examples of images where our model Equalizer predicts “person” rather than “woman” or “man”. In many cases this occurs when the gender evidence is challenging (e.g. first example where only the person’s hands and arms are visible and second example where the person’s face is occluded by the giraffe) or the person’s pose is unusual (third example). However, we also observe cases like the one at the bottom, where Equalizer predicts “person” despite looking at the clear/correct gender evidence. We attribute this to the Confident Loss term, which allows for neutral words generation when the model is uncertain about gender.

Figure 2.6 presents more qualitative examples for the baselines and our model. At the top we show success cases where our model predicts the right gender for the right reasons. At the bottom we show failure cases with incorrectly predicted gender and the wrong gender evidence.

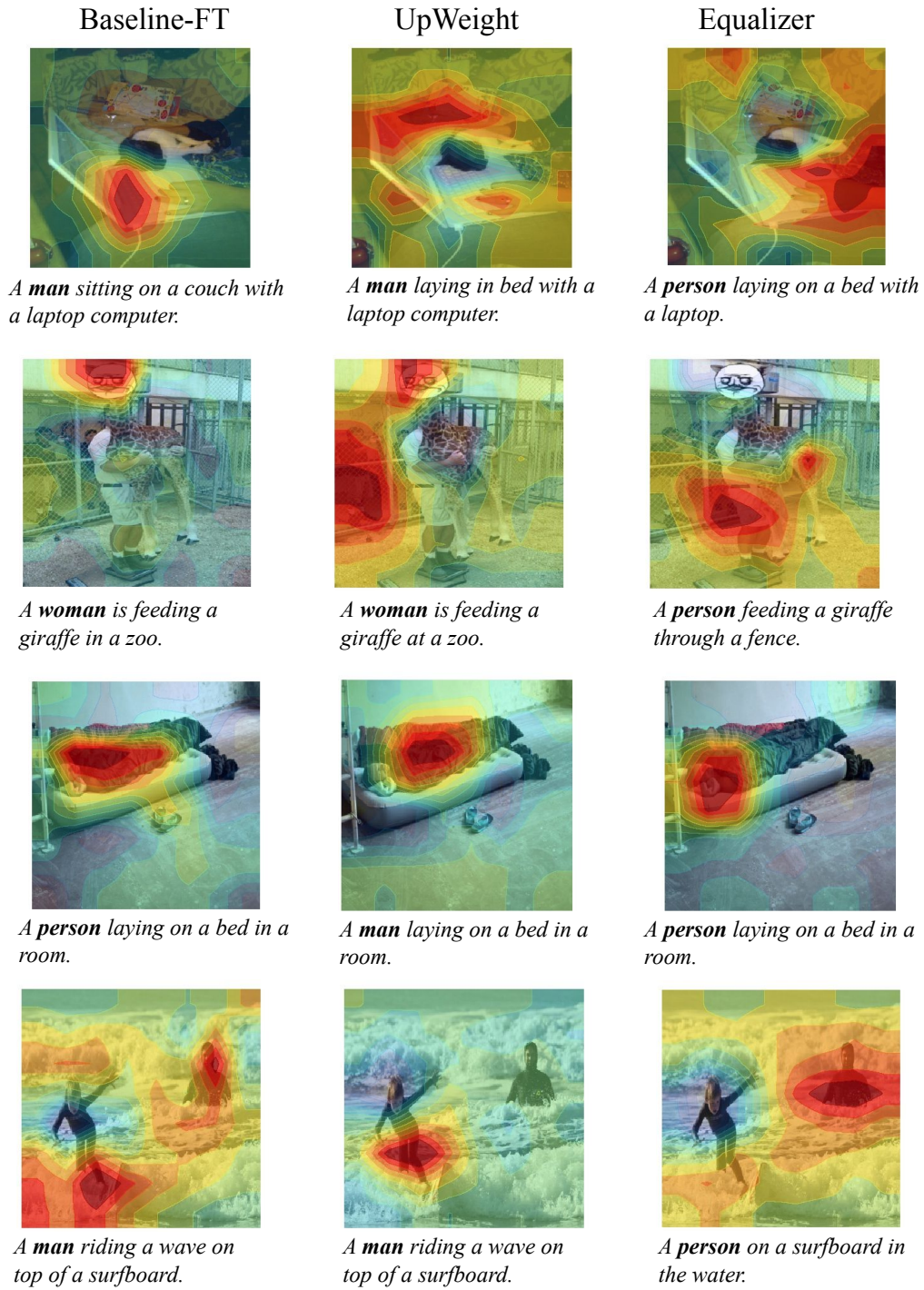


Figure 2.5: Qualitative comparison of baselines and our model when our model predicts “person” rather than “woman” or “man”.

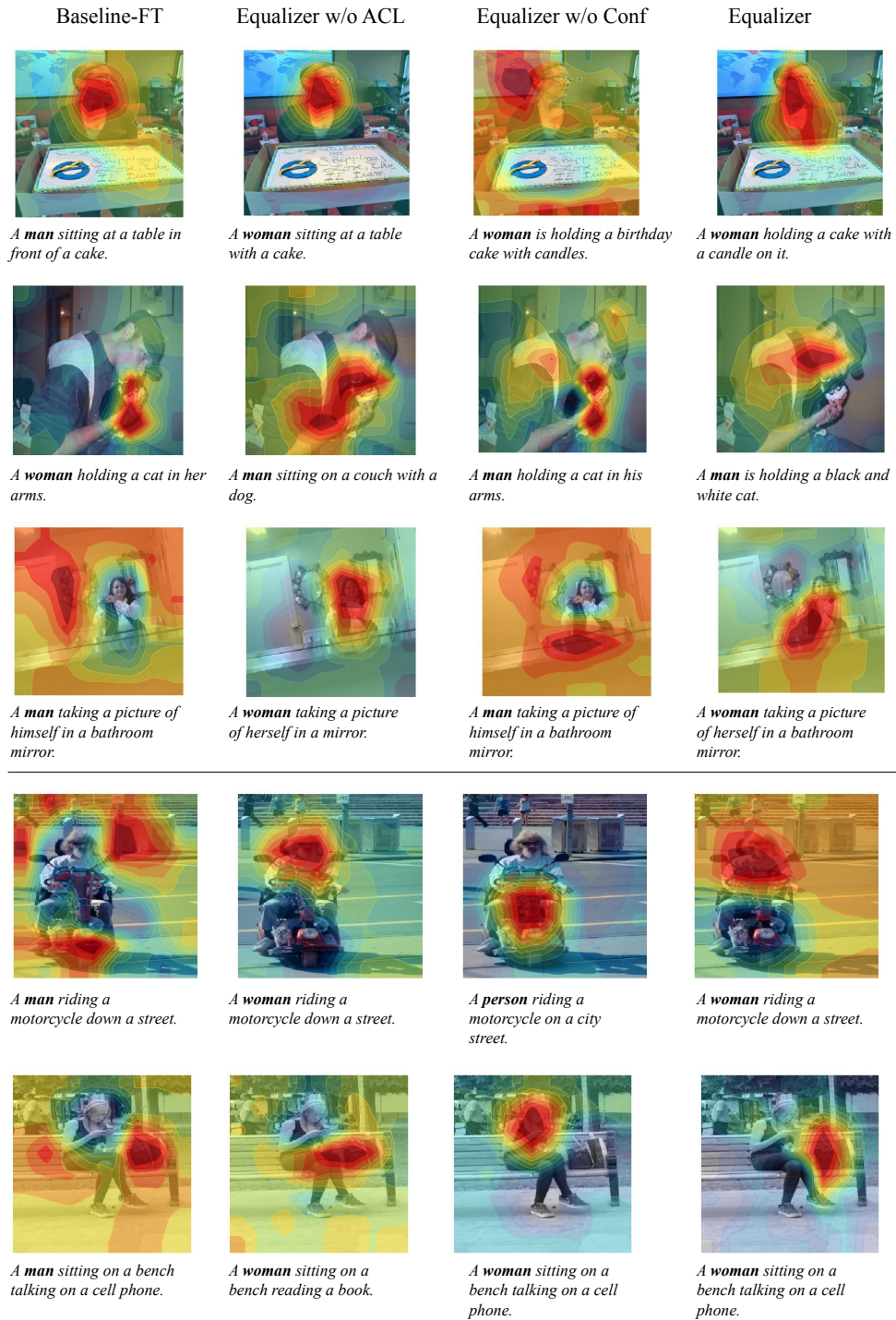


Figure 2.6: Qualitative comparison of baselines and our model. At the top we show success cases where our model predicts the right gender for the right reasons. At the bottom we show failure cases with incorrectly predicted gender and the wrong gender evidence.

2.5 Discussion

We present the Equalizer model which includes an Appearance Confusion Loss to encourage predictions to be confused when predicting gender if evidence is obscured and the Confident Loss which encourages predictions to be confident when gender evidence is present. Our Appearance Confusion Loss, requires human rationales about what is visual evidence is appropriate to consider when predicting gender. We stress the importance of human judgment when designing models which include protected classes. For example, our model can use information about clothing type (e.g., dresses) to predict a gender which may not be appropriate for all applications. Though we concentrate on gender in this work, we believe the generality of our framework could be applied when describing other protected attributes, e.g., race/ethnicity and believe our results suggest Equalizer can be a valuable tool for overcoming bias in captioning models.

2.6 Continuing the Conversation


As research on the impact of gender bias in machine learning progresses, researchers must confront an important ethical question: why do we need gendered words in image captioning?

Aspects of gender appear in captioning datasets in two ways: human annotators embed judgments based on appearance in their descriptions and the individuals in photographs, through their gender expression, provide visual artifacts of gender norms. In this chapter, we discussed how to leverage appearance judgments to balance predictions for female presenting and male presenting individuals and ground predictions about people appropriately in the image. However, classifying gender based on visual appearance can negatively impact vulnerable groups [19]. Indeed, [16] discusses how simply equalizing the probability of predicting “man” and “woman” codifies a cis-gender understanding of human sexuality.

These conversations raise two important questions for future work: does removing judgments about gender from annotations remove bias related to gender expression? if not, how can we develop captioning solutions that are not biased with respect to gender expression without having to make assumptions about gender identity from appearance?

Chapter 3

Object Hallucination in Image Captioning

Despite continuously improving performance, contemporary image captioning models are prone to “hallucinating” objects that are not actually in a scene. One problem is that standard metrics only measure similarity to ground truth captions and may not fully capture image relevance. In this work, we propose a new image relevance metric to evaluate current models with veridical visual labels and assess their rate of object hallucination. We analyze how captioning model architectures and learning objectives contribute to object hallucination, explore when hallucination is likely due to image misclassification or language priors, and assess how well current sentence metrics capture object hallucination. We investigate these questions on the standard image captioning benchmark, MSCOCO, using a diverse set of models. Our analysis yields several interesting findings, including that models which score best on standard sentence metrics do not always have lower hallucination and that models which hallucinate more tend to make errors driven by language priors. 

¹This chapter is based on joint work with Anna Rohrbach, Lisa Anne Hendricks, Trevor Darrell, and Kate Saenko presented at EMNLP 2018 [\[45\]](#). Anna Rohrbach and Lisa Anne Hendricks led the paper. https://people.eecs.berkeley.edu/~lisa_anne/snowboard.html



NBT: A woman talking on a cell phone while sitting on a *bench*.
 CIDEr: **0.87**, METEOR: 0.23, SPICE: **0.22**, CHs: **1.00**, CHi: **0.33**

TopDown: A woman is talking on a cell phone.
 CIDEr: 0.54, METEOR: **0.26**, SPICE: 0.13, CHs: **0.00**, CHi: **0.00**

Figure 3.1: Image captioning models often “hallucinate” objects that may appear in a given context, like e.g. a *bench* here. Moreover, the sentence metrics do not always appropriately penalize such hallucination. Our proposed metrics (CHAIRs and CHAIRi) reflect hallucination. For CHAIR *lower is better*.

3.1 Motivation

Image captioning performance has dramatically improved over the past decade. Despite such impressive results, it is unclear to what extent captioning models actually rely on image content: as we show, existing metrics fall short of fully capturing the captions’ relevance to the image. In Figure 3.1 we show an example where a competitive captioning model, Neural Baby Talk (NBT) [33], incorrectly generates the object “bench.” We refer to this issue as object *hallucination*.

While missing salient objects is also a failure mode, captions are summaries and thus generally not expected to describe all objects in the scene. On the other hand, describing objects that are *not present* in the image has been shown to be less preferable to humans. For example, the LSMDC challenge [46] documents that correctness is more important to human judges than specificity. In another study, [35] analyzed how visually impaired people react to automatic image captions. They found that people vary in their preference of either coverage or correctness. For many visually impaired who value correctness over coverage, hallucination is an obvious concern.

Besides being poorly received by humans, object hallucination reveals an internal issue of a captioning model, such as not learning a very good representation of the visual scene or overfitting to its loss function.

In this paper we assess the phenomenon of object hallucination in contemporary captioning models, and consider several key questions. The first question we aim to answer is: *Which models are more prone to hallucination?* We analyze this question on a diverse set of captioning models, spanning different architectures and learning objectives. To measure object hallucination, we propose a new metric, *CHAIR* (*Caption Hallucination Assessment with*

Image Relevance), which captures image relevance of the generated captions. Specifically, we consider both ground truth object annotations (MSCOCO Object segmentation [32]) and ground truth sentence annotations (MSCOCO Captions [10]). Interestingly, we find that models which score best on standard sentence metrics do not always hallucinate less.

The second question we raise is: *What are the likely causes of hallucination?* While hallucination may occur due to a number of reasons, we believe the top factors include visual misclassification and over-reliance on language priors. The latter may result in memorizing which words “go together” regardless of image content, which may lead to poor generalization, once the test distribution is changed. We propose *image and language model consistency* scores to investigate this issue, and find that models which hallucinate more tend to make mistakes consistent with a language model.

Finally, we ask: *How well do the standard metrics capture hallucination?* It is a common practice to rely on automatic sentence metrics, e.g. CIDEr [57], to evaluate captioning performance during development, and few employ human evaluation to measure the final performance of their models. As we largely rely on these metrics, it is important to understand how well they capture the hallucination phenomenon. In Figure 3.1 we show how two sentences, from NBT with hallucination and from TopDown model [1] – without, are scored by the standard metrics. As we see, hallucination is not always appropriately penalized. We find that by using additional ground truth data about the image in the form of object labels, our metric CHAIR allows us to catch discrepancies that the standard captioning metrics cannot fully capture. We then investigate ways to assess object hallucination risk with the standard metrics. Finally, we show that CHAIR is complementary to the standard metrics in terms of capturing human preference.

3.2 Method: Caption Hallucination Assessment

We first introduce our image relevance metric, *CHAIR*, which assesses captions w.r.t. objects that are actually in an image. It is used as a main tool in our evaluation. Next we discuss the notions of *image and language model consistency*, which we use to reason about the causes of hallucination.

The CHAIR Metric

To measure object hallucination, we propose the *CHAIR (Caption Hallucination Assessment with Image Relevance)* metric, which calculates what proportion of words generated are actually in the image according to the ground truth sentences and object segmentations. This metric has two variants: per-instance, or what fraction of object instances are hallucinated (denoted as CHAIR_i), and per-sentence, or what fraction of sentences include a hallucinated object (denoted as CHAIR_s):

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}$$

For easier analysis, we restrict our study to the 80 MSCOCO objects which appear in the MSCOCO segmentation challenge. To determine whether a generated sentence contains hallucinated objects, we first tokenize each sentence and then singularize each word. We then use a list of synonyms for MSCOCO objects (based on the list from [33]) to map words (e.g., “player”) to MSCOCO objects (e.g., “person”). Additionally, for sentences which include two word compounds (e.g., “hot dog”) we take care that other MSCOCO objects (in this case “dog”) are not incorrectly assigned to the list of MSCOCO objects in the sentence. For each ground truth sentence, we determine a list of MSCOCO objects in the same way. The MSCOCO segmentation annotations are used by simply relying on the provided object labels.

We find that considering both sources of annotation is important. For example, MSCOCO contains an object “dining table” annotated with segmentation maps. However, humans refer to many different kinds of objects as “table” (e.g., “coffee table” or “side table”), though these objects are not annotated as they are not specifically “dining table”. By using sentence annotations to scrape ground truth objects, we account for variation in how human annotators refer to different objects. Inversely, we find that frequently humans will not mention all objects in a scene. Qualitatively, we observe that both annotations are important to capture hallucination. Empirically, we verify that using only segmentation labels or only reference captions leads to higher hallucination (and practically incorrect) rates.

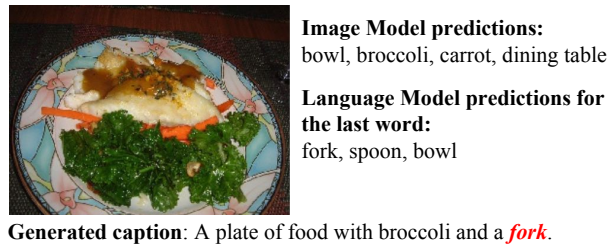


Figure 3.2: Example of image and language consistency. The hallucination error (“fork”) is more consistent with the Language Model.

Image Consistency

We define a notion of *image consistency*, or how consistent errors from the captioning model are with a model which predicts objects based on an image alone. To measure image consistency for a particular generated word, we train an image model and record $P(w|I)$ or the probability of predicting the word given only the image. To score the image consistency of a caption we use the average of $P(w|I)$ for all MSCOCO objects, where higher values mean that errors are *more* consistent with the image model. Our image model is a multi-label classification model with labels corresponding to MSCOCO objects (labels determined the same way as is done for CHAIR) which shares the visual features with the caption models.

Language Consistency

We also introduce a notion of *language consistency*, i.e. how consistent errors from the captioning model are with a model which predicts words based only on previously generated words. We train an LSTM [24] based language model which predicts a word w_t given previous words $w_{0:t-1}$ on MSCOCO data. We report language consistency as $1/R(w_t)$ where $R(w_t)$ is the rank of the predicted word in the language model. Again, for a caption we report average rank across all MSCOCO objects in the sentence and higher language consistency implies that errors are *more* consistent with the language model.

We illustrate image and language consistency in Figure 3.2, i.e. the hallucination error (“fork”) is more consistent with the Language Model predictions than with the Image Model predictions. We use these consistency measures in Section 3.3 to help us investigate the causes of hallucination.

3.3 Exploring Caption Hallucination Results

In this section we present the findings of our study, where we aim to answer the questions posed in Section 3.1: *Which models are more prone to hallucination? What are the likely causes of hallucination? How well do the standard metrics capture hallucination?*

Baseline Captioning Models

We compare object hallucination across a wide range of models. We define two axes for comparison: model architecture and learning objective.

Model architecture. Regarding model architecture, we consider models both with and without attention mechanisms. In this work, we use “attention” to refer to any mechanism which learns to focus on different image regions, whether image regions be determined by a high level feature map, or by object proposals from a trained detector. All models are end-to-end trainable and use a recurrent neural network (LSTM [24] in our case) to output text. For non-attention based methods we consider the **FC model** from [42] which incorporates visual information by initializing the LSTM hidden state with high level image features. We also consider **LRCN** [12] which considers visual information at each time step, as opposed to just initializing the LSTM hidden state with extracted features.

For attention based models, we consider **Att2In** [42], which is similar to the original attention based model proposed by [60], except the image feature is only input into the cell gate as this was shown to lead to better performance. We then consider the attention model proposed by [1] which proposes a specific “top-down attention” LSTM as well as a “language” LSTM. Generally attention mechanisms operate over high level convolutional layers. The attention mechanism from [1] can be used on such feature maps, but Anderson et al. also consider feature maps corresponding to object proposals from a detection model. We consider both models, denoted as **TopDown** (feature map extracted from high level convolutional layer) and **TopDown-BB** (feature map extracted from object proposals from a detection model). Finally, we consider the recently proposed **Neural Baby Talk (NBT)** model [33] which explicitly uses object detections (as opposed to just bounding boxes) for sentence generation.

Learning objective. All of the above models are trained with the standard *cross entropy* (CE) loss as well as the *self-critical* (SC) loss proposed by [42] (with an exception of NBT, where only the CE version is included). The SC loss directly optimizes the CIDEr metric with a reinforcement learning technique. We additionally consider a model trained with a *GAN* loss [50] (denoted **GAN**), which applies adversarial training to obtain more diverse and “human-like” captions, and their respective non-GAN baseline with the CE loss.

TopDown deconstruction. To better evaluate how each component of a model might influence hallucination, we “deconstruct” the TopDown model by gradually removing components until it is equivalent to the FC model. The intermediate networks are *NoAttention*, in which the attention mechanism is replaced by mean pooling, *NoConv* in which spatial feature maps are not input into the network (the model is provided with fully connected

Model	Att.	Cross Entropy					Self Critical				
		S	M	C	CHs	CHi	S	M	C	CHs	CHi
LRCN*		17.0	23.9	90.8	17.7	12.6	16.9	23.5	93.0	17.7	12.9
FC*		17.9	24.9	95.8	15.4	11.0	18.4	25.0	103.9	14.4	10.1
Att2In*	✓	18.9	25.8	102.0	10.8	7.9	19.0	25.7	106.7	12.2	8.4
TopDown*	✓	19.9	26.7	107.6	8.4	6.1	20.4	27.0	117.2	13.6	8.8
TopDown-BB †	✓	20.4	27.1	113.7	8.3	5.9	21.4	27.7	120.6	10.4	6.9
NBT †	✓	19.4	26.2	105.1	7.4	5.4	-	-	-	-	-
GAN ‡		Cross Entropy					GAN				
		18.7	25.7	100.4	10.7	7.7	16.6	22.7	79.3	8.2	6.5

Table 3.1: Hallucination analysis on the Karpathy Test set: Spice (S), CIDEr (C) and METEOR (M) scores across different image captioning models as well as CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). All models are generated with beam search (beam size=5). * are trained/evaluated within the same implementation [34], † are trained/evaluated with implementation publicly released with corresponding papers, and ‡ sentences obtained directly from the author. For discussion see Section 3.3.

feature maps), *SingleLayer* in which only one LSTM is included in the model, and finally, instead of inputting visual features at each time step, visual features are used to initialize the LSTM embedding as is done in the FC model. By deconstructing the TopDown model in this way, we ensure that model design choices and hyperparameters do not confound results.

Implementation details. All the baseline models employ features extracted from the fourth layer of ResNet-101 [21], except for the GAN model which employs ResNet-152. Models without attention traditionally use fully connected layers as opposed to convolutional layers. However, as ResNet-101 does not have intermediate fully connected layers, it is standard to average pool convolutional activations and input these features into non-attention based description models. Note that this means the difference between the *NoAttention* and *NoConv* model is that the *NoAttention* model learns a visual embedding of spatial feature maps as opposed to relying on pre-pooled feature maps. All models except for TopDown-BB, NBT, and GAN are implemented in the same open source framework from [34].²

Training/Test splits. We evaluate the captioning models on two MSCOCO splits. First, we consider the split from Karpathy et al. [25], specifically in that case the models are trained on the respective Karpathy Training set, tuned on Karpathy Validation set and the reported numbers are on the Karpathy Test set. We also consider the *Robust* split, introduced in [33], which provides a compositional split for MSCOCO. Specifically, it is ensured that the object pairs present in the training, validation and test captions do not overlap. In this case the captioning models are trained on the Robust Training set, tuned on the Robust Validation set and the reported numbers are on the Robust Test set.

²<https://github.com/ruotianluo/self-critical.pytorch>

Which Models Are More Prone To Hallucination?

We first present how well competitive models perform on our proposed CHAIR metric (Table 3.1). We report CHAIR at sentence-level and at instance-level (CHs and CHi in the table). In general, we see that models which perform better on standard evaluation metrics, perform better on CHAIR, though this is not always true. In particular, models which optimize for CIDEr frequently hallucinate more. Out of all generated captions on the Karpathy Test set, anywhere between 7.4% and 17.7% include a hallucinated object. When shifting to more difficult training scenarios in which new combinations of objects are seen at test time, hallucination consistently increases (Table 3.2).

Karpathy Test set. Table 3.1 presents object hallucination on the Karpathy Test set. All sentences are generated using beam search and a beam size of 5. We note a few important trends. First, models with attention tend to perform better on the CHAIR metric than models without attention. As we explore later, this is likely because they have a better understanding of the image. In particular, methods that incorporate bounding box attention (as opposed to relying on coarse feature maps), consistently have lower hallucination as measured by our CHAIR metric. Note that the NBT model does not perform as well on standard captioning metrics as the TopDown-BB model but has lower hallucination. This is perhaps because bounding box proposals come from the MSCOCO detection task and are thus “in-domain” as opposed to the TopDown-BB model which relies on proposals learned from the Visual Genome [27] dataset. Second, frequently training models with the self-critical loss actually increases the amount of hallucination. One hypothesis is that CIDEr does not penalize object hallucination sufficiently, leading to both increased CIDEr and increased hallucination. Finally, the LRCN model has a higher hallucination rate than the FC model, indicating that inputting the visual features only at the first step, instead of at every step, leads to more image relevant captions.

We also consider a GAN based model [50] in our analysis. We include a baseline model (trained with CE) as well as a model trained with the GAN loss.³ Unlike other models, the GAN model uses a stronger visual network (ResNet-152) which could explain the lower hallucination rate for both the baseline and the GAN model. Interestingly, when comparing the baseline and the GAN model (both trained with ResNet-152), standard metrics decrease substantially, even though human evaluations from [50] demonstrate that sentences are of comparable quality. On the other hand, hallucination decreases, implying that the GAN loss actually helps decrease hallucination. Unlike the self critical loss, the GAN loss encourages sentences to be human-like as opposed to optimizing a metric. Human-like sentences are not likely to hallucinate objects, and a hallucinated object is likely a strong signal to the discriminator that a sentence is generated, and is not from a human.

We also assess the effect of beam size on CHAIR. We find that generally beam search decreases hallucination. We use beam size of 5, and for all models trained with cross entropy, it outperforms lower beam sizes on CHAIR. However, when training models with the self-critical loss, beam size sometimes leads to worse performance on CHAIR. For example, on

³Sentences were procured directly from the authors.



Figure 3.3: Examples of object hallucination from two state-of-the-art captioning models, TopDown and NBT, see Section 3.3.

	Att	S	M	C	CHs	CHi
FC*		15.5	22.7	76.2	21.3	15.3
Att2In*	✓	16.9	24.0	85.8	14.1	10.1
TopDown*	✓	17.7	24.7	89.8	11.3	7.9
NBT †	✓	18.2	24.9	93.5	6.2	4.2

Table 3.2: Hallucination Analysis on the Robust Test set: Spice (S), CIDEr (C) and METEOR (M) scores across different image captioning models as well as CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). * are trained/evaluated within the same implementation [34], † are trained/evaluated with implementation publicly released with corresponding papers. All models trained with cross-entropy loss. See Section 3.3.

the Att2In model trained with SC loss, a beam size of 5 leads to 12.2 on CHAIRs and 8.4 on CHAIRi, while a beam size of 1 leads to 10.8 on CHAIRs and 8.1 on CHAIRi.

Robust Test set. Next we review the hallucination behavior on the Robust Test set (Table 3.2). For almost all models the hallucination increases on the Robust split (e.g. for TopDown from 8.4% to 11.3% of sentences), indicating that the issue of hallucination is more critical in scenarios where test examples can not be assumed to have the same distribution as train examples. We again note that attention is helpful for decreasing hallucination. We note that the NBT model actually has lower hallucination scores on the robust split. This is in part because when generating sentences we use the detector outputs provided by [33]. Separate detectors on the Karpathy test and robust split are not available and the detector has access to images in the robust split during training. Consequently, the comparison between NBT and other models is not completely fair, but we include the number for completeness.

In addition to the Robust Test set, we also consider a set of MSCOCO in which certain

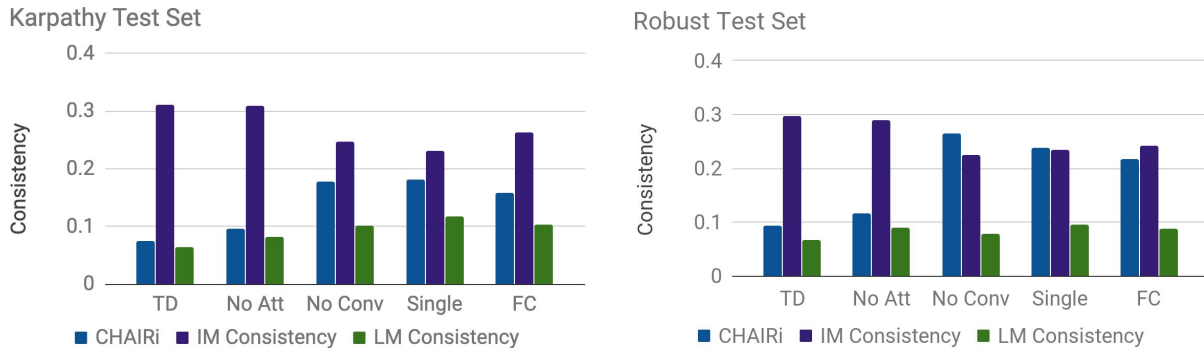


Figure 3.4: Image and Language model consistency (IM, LM) and CHAIRi (instance-level, CHI) on deconstructed TopDown models. Images with less hallucination tend to make errors consistent with the image model, whereas models with more hallucination tend to make errors consistent with the language model, see Section 3.3.

objects are held out, which we call the *Novel Object split* [22]. We train on the training set outlined in [22] and test on the Karpathy test split, which includes objects unseen during training. Similarly to the Robust Test set, we see hallucination increase substantially on this split. For example, for the TopDown model hallucination increases from 8.4% to 12.1% for CHAIRs and 6.0% to 9.1% for CHAIRi.

We find no obvious correlation between the average length of the generated captions and the hallucination rate. Moreover, vocabulary size does not correlate with hallucination either, i.e. models with *more diverse* descriptions may actually *hallucinate less*. We notice that hallucinated objects tend to be mentioned towards *the end of the sentence* (on average at position 6, with average sentence length 9), suggesting that some of the preceding words may have triggered hallucination. We investigate this below.

Which objects are hallucinated and in what context? Here we analyze which MSCOCO objects tend to be hallucinated more often and what are the common preceding words and image context. Across all models the super-category *Furniture* is hallucinated most often, accounting for 20 – 50% of all hallucinated objects. Other common super-categories are *Outdoor objects*, *Sports* and *Kitchenware*. On the Robust Test set, *Animals* are often hallucinated. The *dining table* is the most frequently hallucinated object across all models (with an exception of GAN, where *person* is the most hallucinated object). We find that often words like “sitting” and “top” precede the “dining table” hallucination, implying the two common scenarios: a person “sitting at the table” and an object “sitting on top of the table” (Figure 3.3, row 1, examples 1, 2). Similar observations can be made for other objects, e.g. word “kitchen” often precedes “sink” hallucination (Figure 3.3, row 1, example 3) and “laying” precedes “bed” (Figure 3.3, row 1, example 4). At the same time, if we look at which objects are actually present in the image (based on MSCOCO object annotations), we can similarly identify that presence of a “cat” co-occurs with hallucinating a “laptop”

Karpathy Split	S	M	C	CHs	CHi
TD	19.5	26.1	103.4	10.8	7.5
No Attention	18.8	25.6	99.7	14.2	9.5
No Conv	15.7	22.9	81.3	25.7	17.8
Single Layer	15.5	22.7	80.2	25.7	18.2
FC	16.4	23.3	85.1	23.6	15.8

Table 3.3: Hallucination analysis on deconstructed TopDown models with sentence metrics SPICE (S), METEOR (M), and CIDEr (C), CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). See Section 3.3.

(Figure 3.3, row 2, example 1), a “dog” – with a “chair” (Figure 3.3, row 2, example 2) etc. In most cases we observe that the hallucinated objects appear in the relevant scenes (e.g. “surfboard” on a beach), but there are cases where objects are hallucinated out of context (e.g. “bed” in the bathroom, Figure 3.3, row 1, example 4).

What Are The Likely Causes Of Hallucination?

In this section we investigate the likely causes of object hallucination. We have earlier described how we deconstruct the TopDown model to enable a controlled experimental setup. We rely on the deconstructed TopDown models to analyze the impact of model components on hallucination.

First, we summarize the hallucination analysis on the deconstructed TopDown models (Table 3.3). Interestingly, the *NoAttention* model does not do substantially worse than the full model (w.r.t. sentence metrics and CHAIR). However, removing Conv input (*NoConv* model) and relying only on FC features, decreases the performance dramatically. This suggests that much of the gain in attention based models is primarily due to *access to feature maps with spatial locality*, not the actual attention mechanism. Also, similar to LRCN vs. FC in Table 3.1, initializing the LSTM hidden state with image features, as opposed to inputting image features at each time step, leads to lower hallucination (*Single Layer* vs. *FC*). This is somewhat surprising, as a model which has access to image information at each time step should be less likely to “forget” image content and hallucinate objects. However, it is possible that models which include image inputs at each time step with no access to spatial features overfit to the visual features.

Now we investigate what causes hallucination using the deconstructed TopDown models and the *image consistency* and *language consistency* scores, introduced in Sections 3.2 and 3.2 which capture how consistent the hallucinations errors are with image- / language-only models.

Figure 3.4 shows the CHAIR metric, image consistency and language consistency for the deconstructed TopDown models on the Karpathy Test set (left) and the Robust Test set (right). We note that models with *less* hallucination tend to make errors consistent with



TD: A cat is sitting on a bed in a room.

S: 12.1 M: 23.8 C: 69.7

TD Restrict: A bed with a blanket and a pillow on it.

S: 23.5 M: 25.4 C: 52.5



TD: A cat laying on the ground with a frisbee.

S: 8.0 M: 13.1 C: 37.0

TD Restrict: A black and white animal laying on the ground.

S: 7.7 M: 15.9 C: 17.4

Figure 3.5: Examples of how TopDown (TD) sentences change when we enforce that objects cannot be hallucinated: SPICE (S), Meteor (M), CIDEr (C), see Section 3.3.

the image model, whereas models with *more* hallucination tend to make errors consistent with the language model. This implies that models with less hallucination are better at integrating knowledge from an image into the sentence generation process. When looking at the Robust Test set, Figure 3.4 (right), which is more challenging, as we have shown earlier, we see that image consistency *decreases* when comparing to the same models on the Karpathy split, whereas language consistency is similar across all models trained on the Robust split. This is perhaps because the Robust split contains novel compositions of objects at test time, and all of the models are heavily biased by language.

Finally, we measure image and language consistency during training for the FC model and note that at the beginning of training errors are more consistent with the language model, whereas towards the end of training, errors are more consistent with the image model. This suggests that models first learn to produce fluent language before learning to incorporate visual information.

How Well Do The Standard Metrics Capture Hallucination?

In this section we analyze how well SPICE [2], METEOR [4], and CIDEr [57] capture hallucination. All three metrics do penalize sentences for mentioning incorrect words, either via an F score (METEOR and SPICE) or cosine distance (CIDEr). However, if a caption mentions enough words correctly, it can have a high METEOR, SPICE, or CIDEr score while still hallucinating specific objects.

Our first analysis tool is the TD-Restrict model. This is a modification of the TopDown model, where we enforce that MSCOCO objects which are not present in an image are *not generated* in the caption. We determine which words refer to objects absent in an image following our approach in Section 3.2. We then set the log probability for such words to a very low value. We generate sentences with the TopDown and TD-Restrict model with

	CIDEr	METEOR	SPICE
FC	0.258	0.240	0.318
Att2In	0.228	0.210	0.284
TopDown	0.185	0.168	0.215

Table 3.4: Pearson correlation coefficients between 1-CHs and CIDEr, METEOR, and SPICE scores, see Section 3.3.

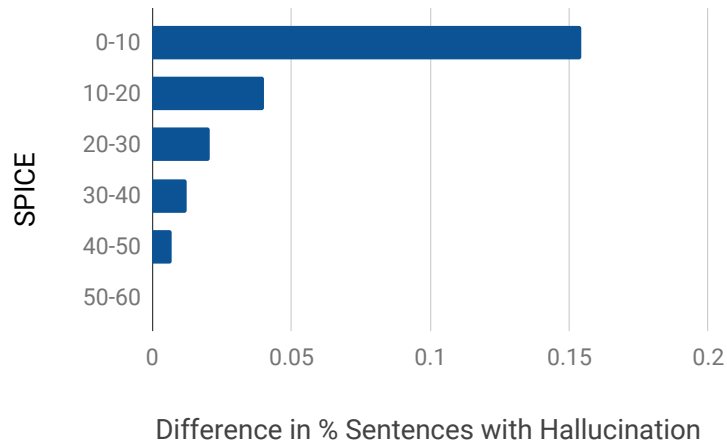


Figure 3.6: Difference in percentage of sentences with *no* hallucination for TopDown and FC models when SPICE scores fall into specific ranges. For sentences with low SPICE scores, the hallucination is generally larger for the FC model, even though the SPICE scores are similar, see Section 3.3.

beam search of size 1, meaning all words produced by both models are the same, until the TopDown model produces a hallucinated word.

We compare which scores are assigned to such captions in Figure 3.5. TD-Restrict generates captions that do not contain hallucinated objects, while TD hallucinates a “cat” in both cases. In Figure 3.5 (left) we see that CIDEr scores the more correct caption much lower. In Figure 3.5 (right), the TopDown model incorrectly calls the animal a “cat.” Interestingly, it then correctly identifies the “frisbee,” which the TD-Restrict model fails to mention, leading to lower SPICE and CIDEr.

In Table 3.4 we compute Pearson correlation coefficient between individual sentence scores and the *absence* of hallucination, i.e. 1-CHAIRs; we find that SPICE consistently correlates higher with 1-CHAIRs. E.g., for the FC model the correlation for SPICE is 0.32, while for METEOR and CIDEr – around 0.25.

We further analyze the metrics in terms of their predictiveness of hallucination risk. Predictiveness means that a certain score should imply a certain percentage of hallucination. Here we show the results for SPICE and the captioning models FC and TopDown. For

	Metric	Metric +(1-CHs)	Metric +(1-CHi)
METEOR	0.269	0.299	0.304
CIDEr	0.282	0.321	0.322
SPICE	0.248	0.277	0.281

Table 3.5: Pearson correlation coefficients between individual/combined metrics and human scores. See Section 3.3.

each model and a score interval (e.g. 10 – 20) we compute the percentage of captions *without* hallucination (1-CHAIRs). We plot the difference between the percentages from both models (TopDown - FC) in Figure 3.6. Comparing the models, we note that even when scores are similar (e.g., all sentences with SPICE score in the range of 10 – 20), the TopDown model has fewer sentences with hallucinated objects. We see similar trends across other metrics. Consequently, object hallucination can *not* be always predicted based on the traditional sentence metrics.

Is CHAIR complementary to standard metrics? In order to measure usefulness of our proposed metrics, we have conducted the following human evaluation (via the Amazon Mechanical Turk). We have randomly selected 500 test images and respective captions from 5 models: non-GAN baseline, GAN, NBT, TopDown and TopDown - Self Critical. The AMT workers were asked to score the presented captions w.r.t. the given image based on their preference. They could score each caption from 5 (very good) to 1 (very bad). We did not use ranking, i.e. different captions could get the same score; each image was scored by three annotators, and the average score is used as the final human score. For each image we consider the 5 captions from all models and their corresponding sentence scores (METEOR, CIDEr, SPICE). We then compute Pearson correlation between the human scores and sentence scores; we also consider a simple combination of sentence metrics and 1-CHAIRs or 1-CHAIRi by summation. The final correlation is computed by averaging across all 500 images. The results are presented in Table 3.5. Our findings indicate that a simple combination of CHAIRs or CHAIRi with the sentence metrics leads to an increased correlation with the human scores, showing the usefulness and complementarity of our proposed metrics.

Does hallucination impact generation of other words? Hallucinating objects impacts sentence quality not only because an object is predicted incorrectly, but also because the hallucinated word impacts generation of other words in the sentence. Comparing the sentences generated by TopDown and TD-Restrict allows us to analyze this phenomenon. We find that after the hallucinated word is generated, the following words in the sentence are different 47.3% of the time. This implies that hallucination impacts sentence quality beyond simply naming an incorrect object. We observe that one hallucination may lead to

another, e.g. hallucinating a “cat” leading to hallucinating a “chair”, hallucinating a “dog” – to a “frisbee”.

3.4 Discussion

In this work we closely analyze hallucination in object captioning models. Our work is similar to other works which attempt to characterize flaws of different evaluation metrics [26], though we focus specifically on hallucination. Likewise, our work is related to other work which aims to build better evaluation tools ([57], [2], [11]). However, we focus on carefully quantifying and characterizing one important type of error: object hallucination.

A significant number of objects are hallucinated in current captioning models (between 5.5% and 13.1% of MSCOCO objects). Furthermore, hallucination does not always agree with the output of standard captioning metrics. For instance, the popular self critical loss increases CIDEr score, but also the amount of hallucination. Additionally, we find that given two sentences with similar CIDEr, SPICE, or METEOR scores from two different models, the number of hallucinated objects might be quite different. This is especially apparent when standard metrics assign a low score to a generated sentence. Thus, for challenging caption tasks on which standard metrics are currently poor (e.g., the LSMDC dataset [44]), the CHAIR metric might be helpful to tease apart the most favorable model. Our results indicate that CHAIR complements the standard sentence metrics in capturing human preference.

Additionally, attention lowers hallucination, but it appears that much of the gain from attention models is due to access to the underlying convolutional features as opposed the attention mechanism itself. Furthermore, we see that models with stronger *image consistency* frequently hallucinate fewer objects, suggesting that strong visual processing is important for avoiding hallucination.

Based on our results, we argue that the design and training of captioning models should be guided not only by cross-entropy loss or standard sentence metrics, but also by image relevance. Our CHAIR metric gives a way to evaluate the phenomenon of hallucination, but other image relevance metrics e.g. those that incorporate missed salient objects, should also be investigated. We believe that incorporating visual information in the form of ground truth objects in a scene (as opposed to only reference captions) helps us better understand the performance of captioning models.

Bibliography

- [1] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and VQA”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [2] Peter Anderson et al. “Spice: Semantic propositional image caption evaluation”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 382–398.
- [3] Shlomo Argamon et al. “Mining the blogosphere: Age, gender and the varieties of self-expression”. In: *First Monday* 12.9 (2007).
- [4] Satanjeev Banerjee and Alon Lavie. “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*. 2005, pp. 65–72.
- [5] Solon Barocas and Andrew D Selbst. “Big data’s disparate impact”. In: *California Law Review* 104 (2016), p. 671.
- [6] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 4349–4357.
- [7] Joy Adowaa Buolamwini. “Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers”. PhD thesis. Massachusetts Institute of Technology, 2017.
- [8] John D Burger et al. “Discriminating gender on Twitter”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 2011, pp. 1301–1309.
- [9] Rich Caruana et al. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1721–1730.
- [10] Xinlei Chen et al. “Microsoft COCO captions: Data collection and evaluation server”. In: *arXiv preprint arXiv:1504.00325* (2015).
- [11] Yin Cui et al. “Learning to Evaluate Image Captioning”. In: *CVPR*. 2018.

- [12] Jeffrey Donahue et al. “Long-term recurrent convolutional networks for visual recognition and description”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 2625–2634.
- [13] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM. 2012, pp. 214–226.
- [14] Eran Eidinger, Roei Enbar, and Tal Hassner. “Age and gender estimation of unfiltered faces”. In: *IEEE Transactions on Information Forensics and Security* 9.12 (2014), pp. 2170–2179.
- [15] Ruth C Fong and Andrea Vedaldi. “Interpretable explanations of black boxes by meaningful perturbation”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.
- [16] Thomas K. Gilbert and Yonatan Mintz. “Epistemic Therapy for Bias in Automated Decision Making”. In: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (2019).
- [17] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. “Contextual action recognition with r* cnn”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1080–1088.
- [18] Jonathan Gordon and Benjamin Van Durme. “Reporting bias and knowledge acquisition”. In: *Proceedings of the 2013 workshop on Automated Knowledge Base Construction*. ACM. 2013, pp. 25–30.
- [19] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. “Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: ACM, 2018, 8:1–8:13. ISBN: 978-1-4503-5620-6. DOI: [10.1145/3173574.3173582](https://doi.org/10.1145/3173574.3173582). URL: <http://doi.acm.org/10.1145/3173574.3173582>.
- [20] Moritz Hardt, Eric Price, Nati Srebro, et al. “Equality of opportunity in supervised learning”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2016, pp. 3315–3323.
- [21] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [22] Lisa Anne Hendricks et al. “Deep compositional captioning: Describing novel object categories without paired training data”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1–10.
- [23] Lisa Anne Hendricks et al. “Women also Snowboard: Overcoming Bias in Captioning Models”. In: *European Conference on Computer Vision (ECCV)* (2018).
- [24] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

- [25] Andrej Karpathy and Li Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3128–3137.
- [26] Mert Kilickaya et al. “Re-evaluating automatic metrics for image captioning”. In: *European Chapter of the Association for Computational Linguistics*. 2016.
- [27] Ranjay Krishna et al. “Visual genome: Connecting language and vision using crowd-sourced dense image annotations”. In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73.
- [28] Brian N Larson. “Gender as a variable in natural-language processing: Ethical considerations”. In: (2017).
- [29] Michael Denkowski Alon Lavie. “Meteor Universal: Language Specific Translation Evaluation for Any Target Language”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 2014, p. 376.
- [30] Gil Levi and Tal Hassner. “Age and gender classification using convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. 2015, pp. 34–42.
- [31] Jianhua Lin. “Divergence measures based on the Shannon entropy”. In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151.
- [32] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2014, pp. 740–755.
- [33] Jiasen Lu et al. “Neural Baby Talk”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [34] Ruotian Luo et al. “Discriminability objective for training descriptive captions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [35] Haley MacLeod et al. “Understanding Blind People’s Experiences with Computer-Generated Captions of Social Media Images”. In: *Proceedings of the 2017 SIGCHI Conference on Human Factors in Computing Systems*. 2017.
- [36] Emiel van Miltenburg. “Stereotyping and bias in the Flickr30k dataset”. In: *Workshop on Multimodal Corpora: Computer vision and language processing*. 2016.
- [37] Ishan Misra et al. “Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2016, pp. 2930–2939.
- [38] United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.
- [39] Novi Quadrianto, James Petterson, and Alex J Smola. “Distribution matching for transduction”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2009, pp. 1500–1508.

- [40] Novi Quadrianto and Viktoriia Sharmanska. “Recycling privileged learning and distribution matching for fairness”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2017, pp. 677–688.
- [41] Vasili Ramanishka et al. “Top-down visual saliency guided by captions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2. 2017, p. 7.
- [42] Steven J Rennie et al. “Self-critical sequence training for image captioning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
- [43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1135–1144.
- [44] Anna Rohrbach et al. “Movie description”. In: *International Journal of Computer Vision* 123.1 (2017), pp. 94–120.
- [45] Anna Rohrbach et al. “Object Hallucination in Image Captioning”. In: *European Conference on Computer Vision (ECCV)* (2018).
- [46] Anna Rohrbach et al. *The Joint Video and Language Understanding Workshop: MovieQA and The Large Scale Movie Description Challenge (LSMDC)*. <https://sites.google.com/site/describingmovies/lsmdc-2017>. 2017.
- [47] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. “Right for the right reasons: Training differentiable models by constraining their explanations”. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2017.
- [48] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. “InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity”. In: *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*. 2018.
- [49] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.
- [50] Rakshith Shetty et al. “Speaking the same language: Matching machine to human captions by adversarial training”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.
- [51] Pierre Stock and Moustapha Cisse. “ConvNets and ImageNet Beyond Accuracy: Explanations, Bias Detection, Adversarial Examples and Model Criticism”. In: *arXiv preprint arXiv:1711.11443* (2017).
- [52] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2818–2826.

- [53] Sarah Tan et al. “Detecting Bias in Black-Box Models Using Transparent Model Distillation”. In: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. 2018.
- [54] Antonio Torralba. “Contextual modulation of target saliency”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2002, pp. 1303–1310.
- [55] Antonio Torralba and Pawan Sinha. “Statistical context priming for object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 1. IEEE. 2001, pp. 763–770.
- [56] Eric Tzeng et al. “Simultaneous deep transfer across domains and tasks”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2015, pp. 4068–4076.
- [57] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4566–4575.
- [58] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE. 2015, pp. 3156–3164.
- [59] Jun Xu et al. “Msr-vtt: A large video description dataset for bridging video and language”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2016, pp. 5288–5296.
- [60] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *Proceedings of the International Conference on Machine Learning (ICML)*.
- [61] Xiang Yan and Ling Yan. “Gender Classification of Weblog Authors.” In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. Palo Alto, CA. 2006, pp. 228–230.
- [62] Peter Young et al. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics (ACL) 2* (2014), pp. 67–78.
- [63] Matthew D. Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2014, pp. 818–833.
- [64] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating Unwanted Biases with Adversarial Learning”. In: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. 2018.
- [65] Jianming Zhang et al. “Top-down neural attention by excitation backprop”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016, pp. 543–559.

- [66] Kaipeng Zhang et al. “Gender and smile classification using deep convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. 2016, pp. 34–38.
- [67] Xu Zhang et al. “Deep transfer network: Unsupervised domain adaptation”. In: *arXiv preprint arXiv:1503.00591* (2015).
- [68] Jieyu Zhao et al. “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2017.
- [69] Luisa M. Zintgraf et al. “Visualizing deep neural network decisions: Prediction difference analysis”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2017.