Visually-Grounded Bayesian Word Learning

Yangqing Jia
Joshua Abbott
Joseph Austerweil
Thomas Griffiths
Trevor Darrell

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-202
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-202.html

October 17, 2012

Visually-Grounded Bayesian Word Learning

Yangqing Jia[1], **Joshua Abbott**[2], **Joseph Austerweil**[2], **Thomas Griffiths**[2], **Trevor Darrell**[1]
[1]Department of EECS [2]Department of Psychology
University of California, Berkeley, CA 94720, USA
{jiayq, joshua.abbott, tom_griffiths, trevor}@berkeley.edu
joseph.austerweil@gmail.com

Abstract

Learning the meaning of a novel noun from a few labeled objects is one of the simplest aspects of learning a language, but approximating human performance on this task is still a significant challenge for current machine learning systems. Current methods typically fail to find the appropriate level of generalization in a concept hierarchy for a given visual stimulus. Recent work in cognitive science on Bayesian models of word learning partially addresses this challenge, but it assumes that the labels of objects are given (hence no object recognition) and it has only been evaluated in small domains. We present a system for learning nouns directly from images, using probabilistic predictions generated by visual classifiers as the input to Bayesian word learning, and compare this system to human performance in an automated, large-scale experiment. The system captures a significant proportion of the variance in human responses. Combining the uncertain outputs of the visual classifiers with the ability to identify an appropriate level of abstraction that comes from Bayesian word learning allows the system to outperform alternatives that either cannot deal with visual stimuli or use a more conventional computer vision approach.

1 Introduction

Learning a language is one of the classic problems that is solved better by the human mind than by any computer. A four-year old child knows the meanings of thousands of words and can learn new words accurately from just a handful of labeled examples [4]. Developing algorithms that approximate human performance on even one aspect of this problem – learning the meaning of a novel noun – is thus a significant challenge. Bayesian word learning models [25] are a step towards answering this challenge, using Bayesian inference to identify the intended level of abstraction referred to by a novel noun (e.g., does the word refer to Dalmatians, dogs, or all mammals?) in a similar manner to human word learning. However, these models do not have a perceptual component and, instead, assume a fixed set of perfectly-recognized stimuli (e.g., it knows a given image is consistent with Dalmatians, dogs, and mammals). We address this limitation by grounding Bayesian word learning with computer vision to produce the first system capable of approximating how people learn nouns directly from images.

Developing a system for determining the referent of a word from labeled images also has the potential to extend the state of the art in computer vision. Over the last decade, computer vision researchers have developed algorithms that can classify images and their contents into a large number of categories [9, 6]. Despite such success, existing image classification algorithms still work in a binary fashion: given an image and a category (e.g. "dog"), the classifier predicts if the image belongs to the category or not. The categories could be mutually exclusive (as in early problems such as digit classification), or nested (as in the context of e.g. ImageNet [6]), in which case the classifier would predict multiple categories that the image belongs to. However, given a set of categories that are all true for an image or a set of images, existing algorithms are not able to further infer which level of

the hierarchy is the true underlying concept (e.g., Dalmatian, dog, or all mammals). Although recent work has proposed using hierarchical structures in object categories [24] or shared attributes [18], learning which objects in an object hierarchy can be referred to by a word (e.g., just Dalmatians or all dog species?) remains an open problem. Unlike other recent computer vision work inspired by human word learning [21], we compare the proposed algorithms to human behavior using a novel large-scale word learning experimental paradigm.

To formulate a visually grounded Bayesian word learning model, we use ImageNet to obtain a large number of images that are used to train a perceptual vision component. In addition to describing a model that replicates human performance in a previous word learning experiment, we present a technique for performing large-scale experimental comparisons of machine and human word learning over a nested hierarchy that contains a broad and deep set of automatically generated domains. We demonstrate that the grounded model and people learn words from small image sets in a similar fashion. Although there still remains room for improving performance, the human results form a new ambitious, yet not impossible, benchmark that should inspire novel algorithms that behave more like human word learners.

2 Background

To explore how children and adults learn words at different levels of abstraction, Xu and Tenenbaum [25] showed participants one or more positive examples of a novel word (e.g., "These three objects are Feps"), while manipulating their taxonomic relationship and asked participants to find the other "Feps" among a variety of both taxonomically related and unrelated objects. For example, a toy Dalmatian could be labeled as a Fep, and people would be asked whether other Dalmatians, dogs, and animals are Feps, along with other objects such as vegetables and vehicles. Xu and Tenenbaum proposed a model of people's responses in which Bayes' rule is used to evaluate a set of hypotheses about which objects are Feps (e.g., Dalmatians are Feps). The taxonomic relationships among the objects define a rooted tree where each object is a leaf node. The space of hypotheses considered by the Bayesian model corresponds to the internal nodes of the tree – for each hypothesis, the objects that are descendants of that node would be identified as Feps. The tree itself was constructed by applying hierarchical clustering [8] to people's judgments of the similarity of all pairs of objects (requiring $O(N^2)$ judgments per participant even though only $N = 45$ objects were used in the experiment).

Relying on human similarity judgments to obtain a set of hypotheses limits the size of the domain in which Bayesian word learning can be applied. To address these issues, recent work [1] proposed automatically constructing a hypothesis space from WordNet, a large lexical database of English represented as a graph of words linked by directed edges denoting semantic relatedness [11, 16]. The resulting model successfully predicted people's generalization judgments in both a replication with the taxonomically-organized domains (animals, vehicles, and vegetables) used by Xu and Tenenbaum [25] and with a set of novel domains that were arranged hierarchically but had a less intuitive taxonomy (clothing, containers, and seats). This WordNet-derived hypothesis space serves as our starting point for developing perceptually grounded bayesian word learning, where inputs are taken directly from the pixels of an image.

All existing Bayesian word learning models assume that people are able to perfectly identify which leaf node of a taxonomy an object should be assigned to based on its appearance (e.g., it knows a given image is consistent with Dalmatians, dogs, and mammals). This is a major limitation of these models from the perspective of providing a solution to the problem of word learning that can be implemented on a computer, as it requires hand-classification of images of objects at all levels of abstraction. In addition, this assumption may not accurately reflect human behavior, as people also perform word learning in a perceptual space that may generate uncertainty. Thus, the challenge for making Bayesian word learning useful in real-world applications is to find a reliable way to connect images (perceptual signals) to the hypotheses evaluated via Bayesian inference.

On the other hand, recent developments in category-level visual appearance modeling make it possible to classify objects into a large number of categories. Specifically, the state-of-the-art classification pipeline usually extracts dense feature from local image patches, using either hand-tuned features such as SIFT [15] or features learned from certain dictionary learning approaches (e.g., 17, 14) and spatially pools the local features [12, 26, 3] to get a vector representation of the input

images. After feature extraction, standard classification methods such as support vector machines (SVMs) or decision trees are learned on a set of training images. Algorithms following such a pipeline have exhibited promising performance on a large number of object categories, such as the Pascal VOC challenge [9] and the ImageNet challenge [6] which involves classifying images that follow an ontology based on WordNet.

Recent work on object recognition has proposed sharing information across categories. Mid-level representations based on attributes [10, 18] focus on extracting common attributes such as "fluffy" and "aquatic" that could be used to semantically describe object categories better than low-level features. Transfer learning approaches have been proposed that jointly learn classifiers sharing a structured regularization term [19]. Of all these previous efforts, our paper is most closely related to work that uses object hierarchies. Salakhutdinov et al. [21] proposed learning a set of object classifiers with a hierarchical regularization term, which improves the classification of leaf node objects by assuming that objects are hierarchically structured. However, they did not address the problem of determining the level of abstraction within the hierarchy at which to make generalizations, which is a key aspect of learning novel words. Deng et al. [7] proposed predicting object labels only to a granularity that the classifier is confident enough with, but their goal was minimizing structured loss instead of mimicking human word learning.

3 Visually-Grounded Bayesian Word Learning

In this section, we formally define the mathematical formulation of the Bayesian word learning framework with vision input, and show its ability to perform word learning on a large number of images.

3.1 The Bayesian Word Learning Framework

The Bayesian word learning framework [25] assumes that we are given a set of N objects $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ that are examples of the concept defined by a novel word, and extend the label to a new object x_{new} with probability given by Bayesian inference in the following manner: Assuming that $\mathcal{C}$ is the set of objects a word can refer to (from which $\mathcal{X}$ is a random sample), the probability that x_{new} is also in $\mathcal{C}$ is given by

$$P(x_{\text{new}} \in \mathcal{C}|\mathcal{X}) = \sum_{k=1}^{K} P(x_{\text{new}} \in \mathcal{C}|h_k)P(h_k|\mathcal{X}), \tag{1}$$

where $\mathcal{H}$ is a set of K hypotheses of the objects referred to by the word, $P(x_{\text{new}} \in \mathcal{C}|h_k)$ is 1 if x_{new} is in the set denoted by h_k and 0 otherwise, and $P(h_k|\mathcal{X})$ is the posterior probability of h_k given the examples $\mathcal{X}$. We assume the objects form a taxonomic hierarchy, where each object x_n belongs to one leaf node in a rooted tree and hypotheses h_k are internal nodes such that $x_n \in h_k$ if x_n is a descendant of h_k.

The posterior distribution over hypotheses is computed using Bayes rule. It is proportional to the product of the *likelihood*, $P(\mathcal{X}|h_k)$, the probability of drawing these examples from h_k uniformly at random and the *prior* probability of h_k, $P(h_k)$. As x_n is drawn uniformly at random from the set of objects picked out by h_k the likelihood is $P(\mathcal{X}|h_k) = (1/|h_k|^N)I(\mathcal{X} \subseteq h_k)$, giving weight to a hypothesis inversely proportional to the number of objects that could be drawn from the hypothesis (with the effect of size increasing with more examples). This "size principle" [22, 23] compares hypotheses at different levels of abstraction that are all consistent with the labeled examples – as the number of examples of a word that are all Dalmatians increases, it becomes increasingly likely that the word applies just to Dalmatians and not to dogs in general, even though both are logically possible. The prior distribution $P(h_k)$ captures biases due to prior knowledge, particular hypotheses with medium granularity or at the basic level (e.g., dogs over Dalmatians or mammals) [20].

3.2 Word Learning with Perceptual Uncertainty

When learning from images, the examples are presented as a set of images $\mathcal{I} = [I_1, I_2, \cdots, I_N]$. If recognition is perfect, each image I_n maps directly to a leaf node x_n. We thus seek to build a set of classifiers that can identify the leaf node x_n for each image I_n. Specifically, given an image I_n, the

classifier predicts a score $f_x(I_n)$ for each leaf node x. A single leaf node can be selected as $\hat{x}_n = \arg\max_x f_x(I_n)$. However, since the classifier prediction usually contains errors, it is beneficial to obtain a probabilistic output $P(x_n = x|I_n)$ from the classifiers. We obtained this probabilistic output by computing the confusion matrix $\mathbf{A}$ by performing cross-validation on the training data, where $A_{i,j}$ is the probability that the true leaf node is i given the classifier output being j.[1] The use of the confusion matrix incorporates the visual ambiguity into the word learning framework: given an image I_n, the probability that the true leaf node is $x_n = x$ is $P(x_n = x|I_n) = A_{x,\hat{x}_n}$ where $\hat{x}_n = \arg\max_x f_x(I_n)$.

Given the probabilistic output from the vision component, the probability of observing I_n as an example of hypothesis h_k now becomes the expectation of $P(x_n|h_k)$ over $P(x_n|I_n)$,

$$P(I_n|h_k) = E_{P(x_n|I_n)}\left[P(x_n|h_k)\right] = \sum_{x=1}^{L} A_{x,\hat{x}_n} P(x_n|h_k). \tag{2}$$

The posterior probability of hypothesis h_k given a set of N images $\mathcal{I}$ is then

$$P(h_k|\mathcal{I}) \propto P(h_k) \prod_{n=1}^{N} P(I_n|h_k), \tag{3}$$

from which we can adopt the standard Bayesian word learning framework (Equation 1) to infer generalization probabilities for new objects.

To obtain the classifiers and the visual ambiguity, we learned a convolutional neural network (following 5) to extract visual features from images obtained from the ImageNet [6]. Specifically, we scale each image to size 32×32 for computational efficiency, densely extract overlapping image patches of size 6×6, and encode ZCA-whitened patches using threshold coding, with a dictionary of size 400 trained by orthogonal matching pursuit. After coding, the local features are average pooled over a 2×2 regular spatial grid to form an 800-dimensional global feature representation for each image. After feature extraction, a standard one-vs-all L_2-SVM is adopted to predict labels for the images (where the weights resulting from the SVM are treated as a vector of noisy labels for an image). We refer the reader to [5] for details about the feature extraction and classification pipeline.

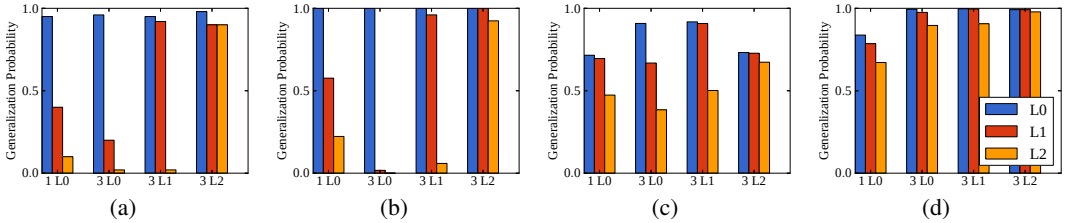
As the images in ImageNet are relatively noisy (e.g., "Dalmatian" may contain the image of a small Dalmatian in a large pumpkin field), we used the Bayesian representativeness approach [see 2] with discretized image features to prune noisy images from each leaf node class, and retain the 400 most representative images. To fairly evaluate our approach, we use 80% of the images as training data to train the classifiers, and use the remaining 20% as testing data to evaluate the performance of concept learning with unseen image input. Additionally, our algorithm is capable of predicting generalization on new, unseen images outside of ImageNet because of the visual model. While the image classifiers are not tuned for specific object classes, we observe an average accuracy rate of around 60% in most domains we tested below.

We believe that the above pipeline is a fair representation of the state-of-the-art computer vision approach. Algorithms using similar approaches have reported competitive performance in image classification on a large number of classes (on the scale of tens of thousands) [13], which provides reassurance about the possibility of learning words using computer vision.

4 Experiments

We now demonstrate the ability of our model to learn the level of generalization of a novel concept in a given hierarchy, in agreement with human performance. We first use a small-scale domain from a previous word learning study [1], but with a much larger number of images (13,200 images as opposed to the 45 images used in the previous experiment). We then propose an automated, large-scale word learning experiment directly from pixel inputs, and show that our model agrees with ground truth performance as given by human behavior.

[1] We also tried converting the classifier output to probabilities via an additional logistic regressor, but found the confusion matrix approach to provide more robust results.



(a) (b) (c) (d)

Figure 1: Generalization results for previously studied domains. (a) Human responses from [1]; (b) Bayesian model assuming perfect labels ($R^2 = 0.981$); (c) Bayesian model with vision input (probabilistic label output) ($R^2 = 0.833$); (d) the conventional vision baseline ($R^2 = 0.570$). Results are grouped in trials denoted by xLy, where x is the number of examples and y is the level from which the examples are sampled. For each trial, the three bars represent the probability that the word can refer to Levels 0, 1, and 2, respectively.

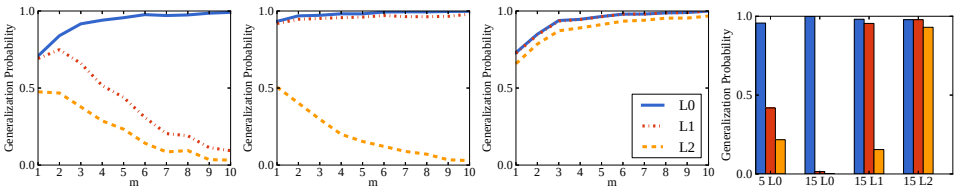
4.1 Initial Experiment

The experiment in [1] involves three object taxonomies (animals, vehicles, and vegetables), with a total of 33 object classes (leaf nodes in the ground-truth concept tree). The set of example images shown to the human participants correspond to four types of trials, expressed at three different levels of abstraction which we will refer to as Levels 0-2 (higher numbers being more abstract). The trial types were: a single Level 0 example (e.g. a Dalmatian); three Level 0 examples (e.g. three Dalmatians); three Level 1 examples, being the single Level 0 example and two examples that share a parent with the Level 0 example (e.g. a Dalmatian, a beagle and a Shih-Tzu); and three Level 2 examples, being the single Level 0 example and two examples sharing a grandparent with the Level 0 example (e.g. a Dalmatian, a toucan and a bear).[2] For the word learning models incorporating image classification, the examples are always selected from test images that are not used to train the classifier. The word learning models then produce generalization probabilities for the examples at Levels 0, 1, and 2 for each test case. We then compare the results against aggregated human subject responses for these domains as reported in [1]. As a quantitative evaluation, we fit a linear model between the per-trial-type per-level probabilities given by each model and the human data, and report the R^2 score.

To the best of our knowledge, there is no vision algorithm that identifies the level of abstraction at which to generalize based on a set of examples. To show the difference between the conventional object classification pipeline and our word learning model, we propose a naïve extension of the conventional 1-vs-all image classifier to predict the level of generalization: the posterior probability of a hypothesis h_k given an example image I is simply defined as the sum of probabilities (given by the confusion matrix) of each leaf nodes in the subtree given I. When there are multiple examples, we take the product of the posterior probabilities computed with each image (i.e., a probabilistic version of the logical AND), and do re-normalization to give the final posterior probability of the hypotheses. Note that no size constraint is utilized in this naive model, so one would expect such a model to prefer high-level concepts (as every example belongs to the root hypothesis - "an object"), and would then generalize to all leaf nodes without much discretion.

Figure 1 shows the generalization probabilities given by people and different models. The word learning model with vision (Figure 1(c)) generalizes in a similar manner to people (Figure 1(a)), but is more conservative than the model assuming perfect labels (Figure 1(b)). For example, when given one Level 0 example, the model with vision is uncertain whether the unknown concept is Level 0 or Level 1, while providing more examples (e.g. three examples) assures the generalization of the Level 0 hypothesis. When provided with three Level 2 examples, the model with vision favors lower-level hypotheses because of the extreme low prior probability of large hypotheses and there is typically at least one lower-level hypothesis that contains classes for each example with non-negligible probability from the vision input.

[2] In this experiment Levels 0, 1, and 2 corresponded to what psychologists identify as subordinate, basic, and superordinate levels [20].



(a) m Level 0 (b) $3m$ Level 1 (c) $3m$ Level 2 (d) $m=5$ ($R^2=0.986$)

Figure 2: Generalization results as a function of the number of input images, using the Bayesian model with vision input. (a)-(c) list three different trial types, and (d) shows the barplot when $m = 5$.

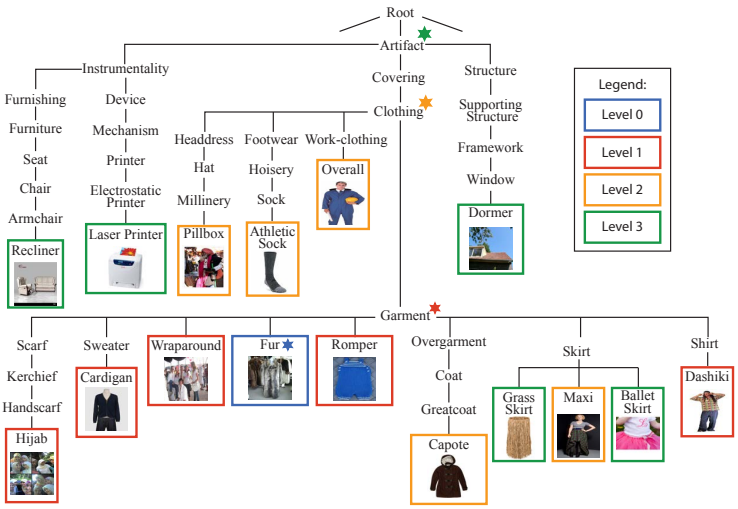


Figure 3: One of the 50 automatically constructed domain used in the experiment that was created by the method detailed in the text. For each level, the root of the corresponding subtree is designated by a star of the appropriate color. Testing leaf nodes are then sampled from the subtree to generate trials.

Without perfect labeling, the additional visual ambiguity (e.g. "is this image a red pepper or a fire truck?") plays an important role. The fact that the visual model does not perfectly classify images causes the generalization probability to "flatten out", and generally, to be lower than the perfect label case. To empirically show this trend, we reduce the visual ambiguity by providing more images in the stimuli to the model with vision. For example, when we provided three Level 1 examples (one example per leaf node), we now provide $3m$ examples (m examples per leaf node, with the same leaf nodes as in the previous experiment). The generalization of our model with respect to m are plotted in Figure 2(a)-(c). It can be observed that using more images as exemplars help to assure the model about the underlying generalization level. For example when $m = 5$ the behavior of our model is very similar to the human subject responses and the Bayesian model assuming perfect labels (Figure 2(d)). As a final note, the conventional vision baseline generalizes poorly, focusing on high-level generalizations as anticipated.

Although the Bayesian word learning model with vision captures human word learning behavior in these three domains, it is unclear if this is due to handpicking the three domains that happen to work with our method. It is also unclear whether our method scales well when there are more realistic numbers of hypotheses involved. In addition, it is worth pointing out that the behavioral experiment reported in [1] was carried out only on a small, manually selected set of images that people recognize easily. This partially explains the observation that the existing Bayesian model assuming perfect labels closely matches human behavior. It is thus beneficial to examine human behavior on a larger scale and determine the effects of visual ambiguity on word learning, as shown next.

4.2 An Automated Large-scale Word Learning Experiment

To perform a large-scale test of word learning from real images, we need to be able to automatically generate a large number of hierarchically nested concepts (domains). Previous methods for testing human word learning use handcrafted domains and images [25] or handpicked domains and images from WordNet and ImageNet [1]. To perform a large-scale evaluation of a computational model, generating the necessary components for testing each domain should be automated. In what follows, we outline a method for automatically constructing four-level hierarchical domains and the images from each object category that can be used in human and machine learning experiments.

To construct a domain, we choose a leaf node in ImageNet uniformly at random from a reduced subset of leaves with at least 800 images in ImageNet. This ensures that each leaf has enough images for training the visual classifier and this initial randomly selected leaf node serves as the most specific category that is to be learned (Level 0). We use an image from the Level 0 category as one of the examples for each of the five trial types. At each increasingly broad trial type, we sample five categories: two categories that are used to create example images shown to participants and three categories that are used to create example images for testing. The test set for a given domain is the same for every trial type and is a randomly ordered set of twenty-four images: twelve in-domain images (three images from the most specific category, and one image from each of the nine broader test categories) and twelve out-of-domain images (four randomly chosen test images from three other automatically generated domains). This results in 21 images in each domain.

For each domain, there are five trial types we show to the human participants:

1. One example from the most specific (Level 0) category.
2. Three examples from the most specific (Level 0) category.
3. Three examples from categories that are 10% of the path to the root of ImageNet (Level 1).
4. Three examples from categories that are 25% of the path to the root (Level 2).
5. Three examples from categories that are 50% of the path to the root (Level 3).

Figure 3 shows the training and test categories for one automatically generated domain.

Using this method, we created fifty domains, each with five trial types (resulting in 1050 images shown to participants). Each trial type was completed by 10 unique participants (resulting in 2500 trials, 250 of which were unique), who were recruited online through Amazon Mechanical Turk (`http://www.mturk.com`) and were compensated $0.05 USD for each completed trial (and were allowed to complete as many unique trials as they wished). The images used as examples in a domain were the same for every replication (the same Level 0 image appeared in every trial type), and the test images of a domain were the same for every trial type and replication. The same random ordering of the example and test images was used for replications.

To learn the perceptual classifiers, we collected images from the corresponding ImageNet categories, and performed feature extraction and per-domain classifier training as previously described. Note that every image used in the human behavior experiment (as examples or test images) was randomly sampled from the set of held-out images that were not used to train any portion of the perceptual model.

4.3 Human and Computational Results on the Large Scale Data

We show the generalization probabilities for people and the different models in Figure 4. The first thing to notice is that people exhibit different behavior from that observed in previous smaller-scale experiments. Specifically, the generalization probabilities are lower than the "cleaner", smaller-scale case discussed in the previous subsection, where identifying objects was easier. Values tend to be farther from the two extremes (zero or one), possibly because the randomly sampled images make the visual input more ambiguous than manually picked ones.

The Bayesian model assuming perfect labeling does not take visual ambiguity into consideration, and still tends to give extreme values to the generalization probabilities. In contrast, human word learning behavior is more similar to our model, which combines image classification and Bayesian word learning components. For example, in both Figure 4 (a) and Figure 4 (c), the response for each level peaks when the given examples are drawn from the same level. The conventional vision baseline still fails to perform generalization, giving high generalization scores to most of the trial

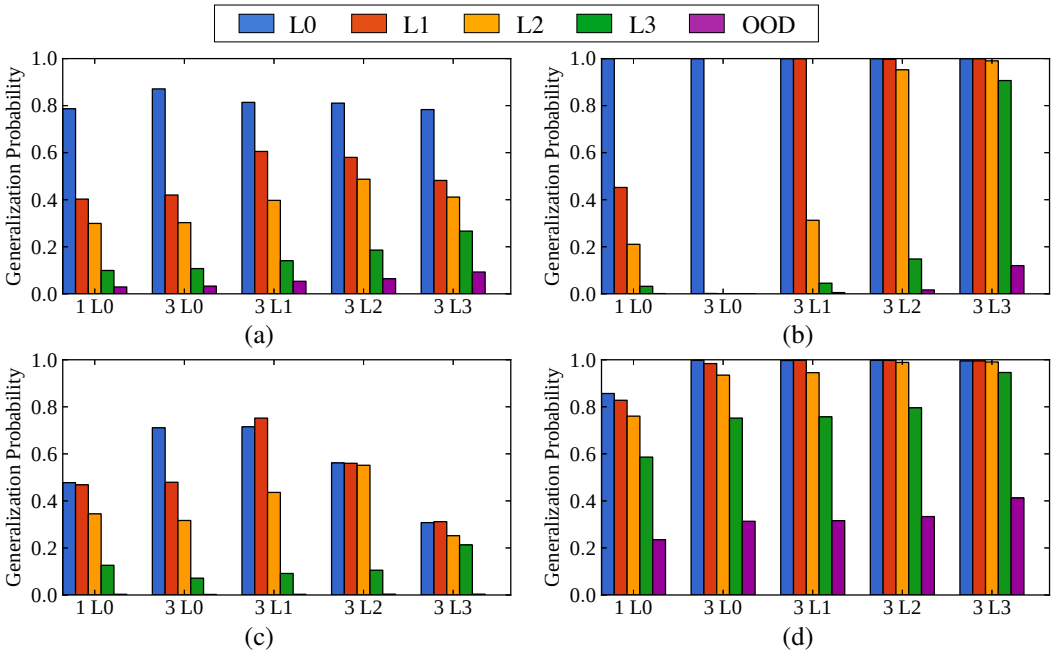


Figure 4: Generalization proabilities given by various models: (a) Human word learning, (b) Bayesian model with perfect label, (c) Bayesian model with vision input, and (d) Conventional vision baseline. Results are first grouped by trial types, and within each group, probabilities from the most specific (Level 0) to the most general (out of domain, OOD) responses are listed.

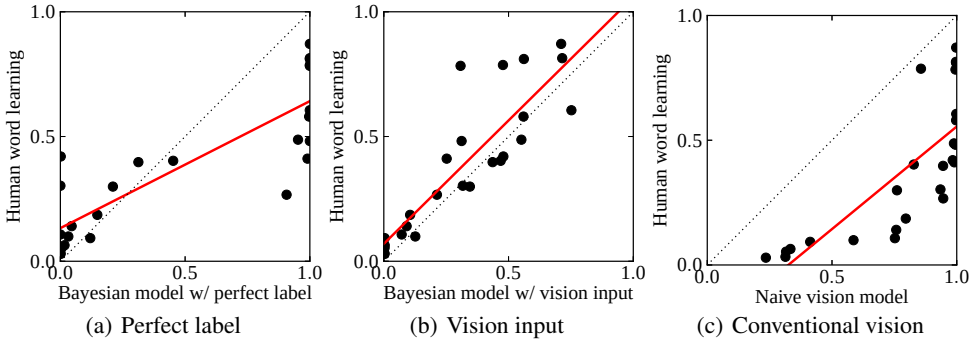


Figure 5: Correlation between the human word learning and the word learning models. The red line shows the fitted linear regression model, and the corresponding R^2 scores are 0.71, 0.77 and 0.59 respectively.

and response types. In summary, our result shows the importance of modeling visual ambiguity in explaining human generalization behavior in a larger scale, more realistic context, which is missing in previous work.

To quantitatively evaluate the performance difference, we show the correlation between human word learning and the different models by comparing the per trial type per response type generalization probabilities. The scatter plots are shown in Figure 5. The model with vision input yields a better R^2 score when we perform a linear fit, with a slope very close to 1 and a positive bias. In contrast, the Bayesian model with perfect label produces more extreme probability values either close to 0 or to 1, and the conventional vision baseline has a large bias towards higher probabilities.

5 Conclusions

In this paper, we proposed a visually grounded Bayesian word learning framework that models human word learning behavior given a set of visual stimuli. By developing a novel visually grounded Bayesian word learning model that accounts for perceptual uncertainty, our approach addresses limitations in the existing word learning and image classification algorithms: the former only being able to use a limited number of stimuli, and the latter not being able to infer the level of generalization. We believe that our work is the first to present a model that learns a visual concept from sets of raw image input, and to empirically show the resemblance of human generalization behavior on a large scale. We hope that this work provides common ground for computer vision, machine learning, and cognitive science researchers, and provides a step towards developing novel object recognition algorithms that better mimic with human behavior.

References

[1] J.T. Abbott, J.L. Austerweil, and T.L. Griffiths. Constructing a hypothesis space from the Web for large-scale Bayesian word learning. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012.

[2] J.T. Abbott, K.A. Heller, Z. Ghahramani, and T.L. Griffiths. Testing a Bayesian measure of representativeness using a large image database. In *NIPS*, 2011.

[3] Y.L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[4] S. Carey. The child as word learner. *Linguistic Theory and Psychological Reality*, 1978.

[5] A. Coates, H. Lee, and A.Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2010.

[6] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] J. Deng, J. Krause, A. Berg, and L. Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *CVPR*, 2012.

[8] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[11] C. Fellbaum. WordNet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[13] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[14] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009.

[15] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[16] G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[17] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[18] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[19] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *CVPR*, 2008.

[20] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

[21] R. Salakhutdinov, A. Torralba, and J.B. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011.

[22] J. B. Tenenbaum. Bayesian modeling of human concept learning. In *NIPS*. 1999.

[23] J. B. Tenenbaum and T. L. Griffiths. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640, 2001.

[24] A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE TPAMI*, 29(5):854–869, 2007.

[25] F. Xu and J.B. Tenenbaum. Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272, 2007.

[26] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.