Learning Hierarchical Visual Representations in Deep Neural Networks Using Hierarchical Linguistic Labels

Joshua Peterson, Paul Soulos, Aida Nematzadeh, & Thomas L. Griffiths
Department of Psychology, University of California, Berkeley
{jpeterson,psoulos,nematzadeh,tom_griffiths}@berkeley.edu

Abstract

Modern convolutional neural networks (CNNs) are able to achieve human-level object classification accuracy on specific tasks, and currently outperform competing models in explaining complex human visual representations. However, the categorization problem is posed differently for these networks than for humans: the accuracy of these networks is evaluated by their ability to identify single labels assigned to each image. These labels often cut arbitrarily across natural psychological taxonomies (e.g., dogs are separated into breeds, but never jointly categorized as "dogs"), and bias the resulting representations. By contrast, it is common for children to hear both *dog* and *Dalmatian* to describe the same stimulus, helping to group perceptually disparate objects (e.g., breeds) into a common mental class. In this work, we train CNN classifiers with multiple labels for each image that correspond to different levels of abstraction, and use this framework to reproduce classic patterns that appear in human generalization behavior.

Category Learning in Children and AI

One of the challenges for a child learning a language is identifying what level in a hierarchical taxonomy a word – category label – refers to. After hearing the word "dog" upon observing a Dalmatian called Sebastian, a child needs to learn whether the category label "dog" refers to only Sebastian, all the Dalmatians, different breeds of dogs, etc. Psychologists have extensively studied this problem. A classic line of psychological research examined what level of abstraction (*e.g.*, dogs vs. Dalmatians) conveys the most information; this level is called the *basic level* (Rosch, 1973; Rosch, Mervis, Gray, M, & Boyes-Braem, 1976). Another line of research has investigated whether people have an innate or a learned basic-level bias, *i.e.*, a tendency to generalize a new category label to the members of the basic level of the taxonomy (*e.g.*, Golinkoff, Mervis, & Hirsh-Pasek, 1994; Markman, 1991).

Artificial intelligence (AI) systems need to address a similar problem. Given an image of a dog, an AI system needs to find the best label to describe that image. In computer vision, this problem is often formulated as an image classification task where the training data consists of images paired with single labels from a limited set of categories; a model needs to learn to predict the correct label for new images. Deep convolutional neural networks have been very successful at both achieving state-of-art accuracy in the image classification task (LeCun, Bengio, & Hinton, 2015) as well as learning representations that best explain human psychological and neural representations for natural images (*e.g.*, Agrawal, Stansbury, Malik, & Gallant, 2014; Mur et al., 2013; Peterson, Abbott, & Griffiths, 2016).

However, this problem formulation is different from what children experience. Multiple labels (*e.g.*, Dalmatian and dog) are commonly used to refer to the same entity in the world children encounter. This explicit use of multiple category labels can help children learn a better representation of the taxonomic relations between categories – the implicit hierarchical structure underlying the world. Different breeds of dogs form a category not only because of their perceptual similarity but also because they are referred to by the same label "dog", implicitly defining a higher level in the taxonomy. Moreover, forming a hierarchical representation can in turn help children better generalize to new items; for example, a new breed of dog, such as a poodle, will be categorized as a dog because of its similarity on crucial dimensions (e.g., long snout) to the members of that category.

In this work, we investigate whether using such human-like supervision (labels from different levels of a hierarchy for a given entity) can help deep neural networks learn better visual representations. We explore the consequences of this multi-level classification by training a near state-of-the-art image classifier (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2015) on a dataset that provides multiple labels for images, each corresponding to a different level of the hierarchical taxonomy. In particular, we focus on two sets of labels: basic and subordinate (in the hierarchical taxonomy, subordinate labels, such as "Dalmatian", are categories that are below the basic level, such as "dog").

To examine the learned representations, we perform three sets of experiments. First, we explore whether the representations learned by each model capture the similarity relations observed in a hierarchical taxonomy. For example, we expect that different breeds of dogs will be better clustered together. We observe that training with basic-level labels results in both better grouping of similar examples and learning of hierarchical structure. Dendrograms reveal stark differences in hierarchical representations learned from basic as opposed to subordinate labels, resulting in interesting high-level groups not present in the original representations. Second, we show that the representations learned by training with basic-level labels match the performance of more commonly used subordinate-level label training schemes in explaining human similarity judgments, indicating that simpler classification tasks are sufficient to capture rich structure in human mental representations. Finally, we show that training with both label sets results in a better match to a classic effect of basic-level generalization bias observed in people using a simple model that restricts generalization with the number of consistent examples, an effect not previously replicated with a fully learned representation.

Background

Exploring Levels of Generalization

Cognitive scientists have extensively studied the problem of finding the correct level of generalization for a given label. The seminal work of Rosch (1973) emphasizes the importance of basic-level objects because they are categorized first during perception and named early during word learning. Research on child word learning suggests that children have a bias in generalizing a new label to the the basic-level category members (*e.g.*, Golinkoff et al., 1994; Markman, 1991).

Xu and Tenenbaum (2007) further examined the basic-level bias by studying how children and adults generalize a new word after observing a few examples. They found that after observing examples from a subordinate level (*e.g.*, three Dalmatians) labeled with a new word such as "dax", the participants often include dogs other than Dalmatians (*e.g.*, poodles) in the category "dax" but exclude animals other than dogs (*e.g.*, cats). Previous models have successfully predicted the generalization patterns observed in this experiment; however, they require the use of pre-specified rather than learned hierarchical taxonomies based on adult similarity judgments (Xu & Tenenbaum, 2007), or limited feature representations (Nematzadeh, Grant, & Stevenson, 2015). The main difference between our work and the existing models is that we explore whether the hierarchical structure of a taxonomy can be learned simply by jointly classifying natural images with multiple labels (or by choosing the right labels), mirroring the multiple references heard by children.

Explaining Human Behavior with Deep Networks

Another line of related work explores how representations from deep neural networks can be used to explain complex human behavior. For example, deep convolutional neural networks (CNNs; LeCun et al., 2015) are specialized for image processing tasks, and can explain human typicality (Lake, Zaremba, Fergus, & Gureckis, 2015) and similarity ratings (Peterson et al., 2016), object memorability (Dubey, Peterson, Khosla, Yang, & Ghanem, 2015), and shape sensitivity (Kubilius, Bracci, & de Beeck, 2016). Peterson et al. (2016) found that deep representations could be tuned post-hoc to improve fit to human similarity judgments by almost 50%. Although the resulting representations better capture the human mental representations, the size of the image dataset used was relatively small, defining a limited context. Further, the reason for the initial discrepancies (and different stimulus groupings) is unknown, and the nature of the learned tuning is opaque. For this reason, it may be useful to instead consider training schemes that more closely match the pressures human categorizors face to more elegantly derive the relevant representations.

Recent work in explainable AI has emphasized the importance of understanding the biases that the training data or a model's architecture can introduce (*e.g.*, Ritter, Barrett, Santoro, & Botvinick, 2017; Zhao, Wang, Yatskar, Ordonez, & Chang, 2017). Our analysis sheds light on another one of these important biases – the basic-level bias – that the image classification model captures.

Multi-level and Multi-label Classification

Computer vision research has recognized the importance of training multi-label classifiers to take advantage of the available rich linguistic information in tasks such as image understanding (*e.g.*, J. Wang et al., 2016). Some of this work makes use of labels from different levels of a hierarchy, but has mostly focused on improving classification performance; for example, Lei, Guo, and Wang (2017) show that using coarse-grained labels can improve the classification accuracy of finer-grained categories. P. Wang and Cottrell (2015) are the first to take inspiration from the work of Rosch (1973), but they also focus on improving subordinate-level benchmarks. They trained CNN classifiers on a set of 308 basic level labels that encompassed an existing set of 1000 ILSVRC12 competition classes (Russakovsky et al., 2015) that clearly fit the traditional conception of subordinate categories (Rosch et al., 1976). They used this network as an initialization that was then fine-tuned on the original 1000 classes, obtaining higher accuracy than their baseline model trained without basic level pre-training. Our work differs in that we assess the structure of the learned representations and implications on later generalization behavior, both for each label set as the sole source of supervision, and as simultaneous and potentially competing objectives.

Training Paradigms

We are interested in examining the extent to which linguistic labels from different levels of the taxonomy (*i.e.*, basic and subordinate categories) affect the structure and hierarchical taxonomy of the learned representations. To examine this hypothesis, we train classifiers using each level separately, and also using multiple labels (first taking a pretrained model on one level and tuning it using labels from both levels). Our goal in the latter case is to more closely mimic a child's experience in language learning – children are known to make use of simple labels first (*e.g.*, "dog"; Rosch et al., 1976), later followed by increasingly sharper distinctions (*e.g.*, "poodle").[1]

More specifically, we pose the multi-level labeling problem simply as learning a set of independent softmax classifiers that are unconnected to each other and fully connected to the final representation layer of a deep CNN. We define the loss function for the multi-level classifiers as $0.5\mathcal{L}_{basic} + 0.5\mathcal{L}_{subordinate}$. Because these models have already been pretrained on one level, this joint loss can be thought of as a method to prevent the fine-tuning procedure from overwritting previous knowledge about the original training domain (*i.e.*, we would like the model to remember what it learned about basic labels while approaching the new problem of more fine-grained subordinate labeling, and vice

[1] Another possibility is to train multi-label classifier simultaneously on the two levels which we will explore in future.

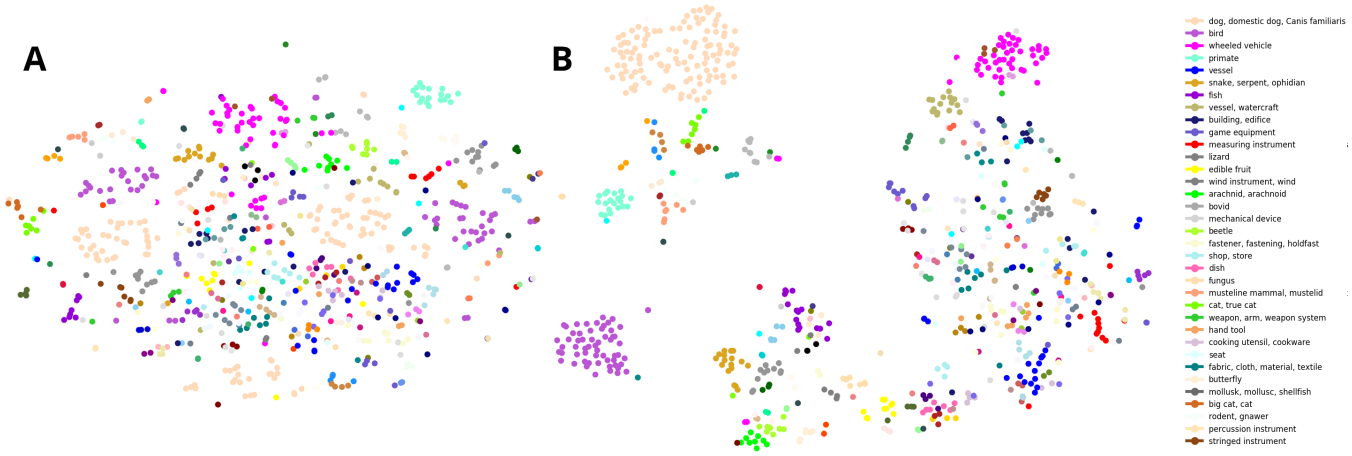

Figure 1: t-SNE visualizations of representations learned for (**A**) Sub training and (**B**) Basic training, colored by each of the 35 basic labels with the most subordinate classes for clarity. The basic model has tighter clusters and more distinct boundaries between categories.

versa). While other alternative approaches exist for defining the network architecture and loss function, this approach provides a single embedding space for all images, which allows us to inspect the representations with classic psychological methods such as hierarchical clustering.

We compare the representations from our multi-level classifiers with ones that are trained using only one level (basic or subordinate). To train on single levels, we use a k-way softmax classifier as the final layer, where k is the number of classes for that level. We report the results for four different models:

- **Sub**: Trained only on subordinate.
- **Basic**: Trained only on basic.
- **Sub**$_{basic}$: Pretrained on basic; tuned with subordinate.
- **Basic**$_{sub}$: Pretrained on subordinate; tuned with basic.

To train our multi-level classifiers, we use the model architecture defined in InceptionV3 (Szegedy et al., 2015) since it obtains high accuracy while still being relatively quick to train. It contains 159 layers, 23 million parameters, and a top-1 validation accuracy of nearly 79% on ILSVRC12 using the default subordinate label set. We use the off-the-shelf pretrained version of this network for the Sub model. For the Basic model, we reinitialize the network parameters randomly and retrain from scratch. For fine-tuning models, we freeze all but the weights in the last block of InceptionV3 to speed training.

Following P. Wang and Cottrell (2015), we use the 1000 labels from ILSVRC12 as our subordinate classes, and basic level labels provided by the same authors, described in the previous section.

Exploring Representations

We perform a set of exploratory qualitative visual analyses to examine whether the representations learned by each model reflect the grouping of similar objects observed in a hierarchical taxonomy. If representations capture such higher-level abstractions, not only would examples of Dalmatians cluster close to each other, but also different breeds. To analyze the learned representations, we extracted a 2048-dimensional feature vector from the last hidden layer before the output, and used for all subsequent analyses in this paper.

We first explored the representations using t-SNE (Maaten & Hinton, 2008), a common visualization method for CNN representations, shown in Figure 1. The t-SNE method reduces the 2048-dimensional vectors to 2 dimensions so that we can visually compare learned representations. We find that the Sub model does not appear to cluster basic level categories very well (*e.g.*, there are multiple spatially separated clusters of dogs), whereas the Basic model does a much more effective job of clustering these like subordinate classes together (*e.g.*, dog breeds are now unified under a single tight cluster). Models trained on both label sets simultaneously, Sub$_{basic}$ and Basic$_{sub}$) exhibited similar clustering to the Basic model, and so they are omitted for this reason.

We also plot dendrograms of the the representations in Figure 2. The Basic model has a clearly defined hierarchy, whereas the Sub model does not. The branch we call "artificial" contains images showing artificial objects such as cars, buildings, household objects, sports, and technology. The other main branch, the "natural" distinction show animals, fish, mushrooms, and other natural stimuli. Interestingly, this high-level distinction is not present explicitly in the basic label set, and is one of the defining categorical divisions found in human mental representations (Mur et al., 2013). As before, models trained on both label sets (Sub$_{basic}$ and Basic$_{sub}$) behaved similarly to the Basic model, suggesting that significant basic-level supervision at any stage of training allows for the preservation of basic-level clustering.

Predicting Human Similarity Judgments

Another way to evaluate the goodness of the learned representations is examine how well they predict human psychologi-

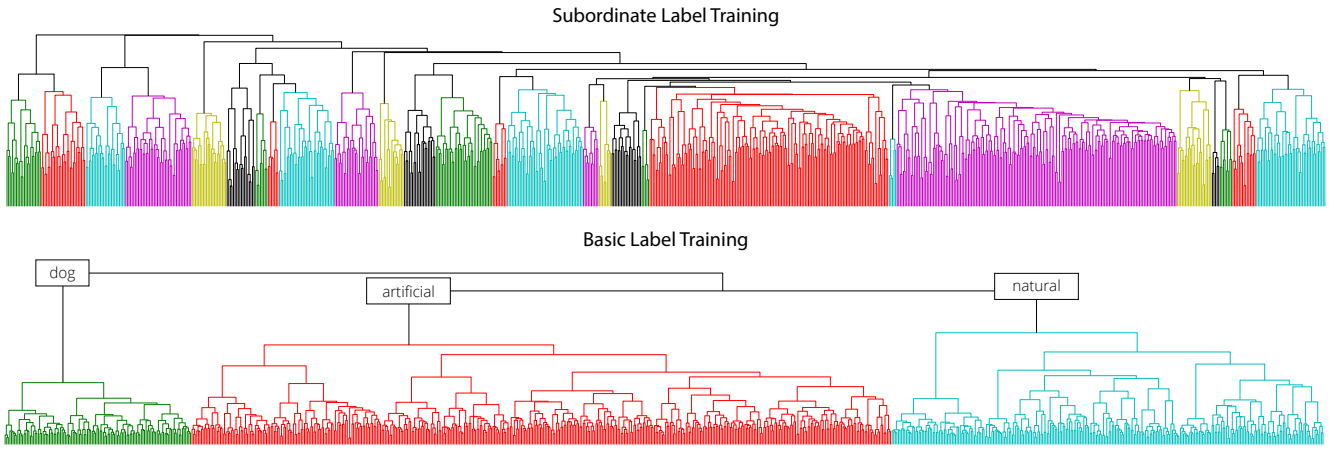

Figure 2: Dendrograms showing the learned representations. The model trained on subordinate labels (top) has no clear hierarchical structure. The model trained on basic labels (bottom) has a clearly defined hierarchical structure which divides the validation data into three top level categories: dogs, images of artificial items, and images of natural items. Over 10% of Imagenet is images of dogs which is why there is a separate branch specifically for these images. We did not label the subordinate model because it did not have a definable structure.

cal representations as captured by pairwise human similarity judgments (Shepard, 1987). As a gold-standard dataset, we use the similarity ratings collected by Peterson et al. (2016). This dataset consists of 7,140 pairwise comparisons of all 120 images presented to MTurk workers, who were asked to rate their similarity between 0 and 10.

For each model, we compute the similarity between image pairs as the inner product of their learned representations. Then, we calculate the Pearson correlation between the representations learned by each model and the gold-standard similarity ratings. We report R^2 to compare directly with previous work (see Table 1). Interestingly, the performance of the models that either train solely on basic labels or are tuned with basic labels greatly surpass those that train solely on subordinate labels or are tuned with subordinate labels. We note that previous work has obtained very similar correlations to our Basic results by using a different (yet comparable) network trained only on subordinate labels (Peterson et al., 2016). This may indicate the current network is generally less apt to predict human similarity judgments, but is more likely to encode information relevant to people given the right supervision from basic labels. In addition, our results show that human psychological representations can be approximated using simpler classifiers with more coarse-grained object distinctions (basic level labels). It is also worth noting that most of the previous work in explaining human visual representations uses subordinate-label trained networks (Agrawal et al., 2014; Mur et al., 2013); our results show a new alternative for predicting these representations.

Generalization Experiments

Xu and Tenenbaum (2007) (henceforth, X&T) examined how people generalize a novel word label after observing a few examples, and whether the number or the taxonomic level from which the examples were drawn changes the generalization behavior. For example, the participants heard a word label such as "dax" while observing one Dalmatian, three Dalmatians, or three different breeds of dogs. Figure 3 shows the set-up of their experiments.

	Sub	Basic	Sub_{basic}	Basic_{sub}
R^2	0.38	0.57	0.41	0.57

Table 1: Variance explained in human similarity judgments by representations formed by each model.

The experiment set-up of X&T's can be thought of as a few-shot learning task. We examine whether our models exhibit the generalization behavior observed in these experiments. We focus on three of their training conditions.

- **1 sub**: the model (or a participant in X&T's experiment) observes 1 (subordinate) example, such as a Dalmatian.
- **3 sub**: the model receives 3 examples from the same subordinate category, such as 3 Dalmatians.
- **3 basic**: 3 examples are drawn from a basic-level, such as a poodle, a Dalmatian, and a Great Pyrenees.

In X&T's experiments, after training, the participants were asked to pick everything that is a "dax" from a fixed set of examples drawn from different levels of taxonomy. We focus on the test objects that are relevant to our experimental conditions, *i.e.*, two basic-level and two subordinate matches.[2] For example, after observing one Dalmatian as a "dax", the participants had to decide whether two other Dalmatians, a poodle, and a golden retriever are also examples of "dax". The results of their experiment is shown in Figure 4. Two interesting observations can be made from their results: First,

[2]X&T also included superordinate examples (such as an animal other than a dog). Here we only focus on the subordinate and basic-level matches because our training data only includes those labels.

people exhibit basic-level bias since they generalize to the basic-level after observing only subordinate examples. Second, their degree of basic-level generalization decreases after observing 3 examples (see "Basic Match" in 1-sub and 3-sub conditions in plot titled "Human" in Figure 4).

We simulate the different conditions of X&T's experiments as a few-shot generalization task, and examine the generalization behavior of models on both subordinate and basic-level matches.

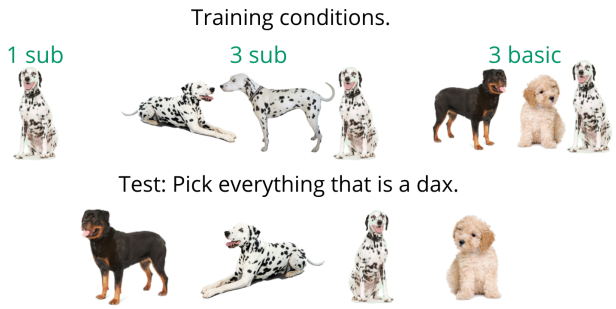

Figure 3: The setup of X&T's experiments Xu and Tenenbaum (2007) used in this paper. The first row shows the three training conditions and the second row is an example of the test trial. Note that during the test trial, the basic matches (dog breeds other than Dalmatians) are different from the ones observed during training (in contrary to what is shown here).

Learning from Positive Examples

To mimic generalization experiments with humans, we use a k-shot generalization formulation meant to learn concepts from positive examples only. Specifically, we use exponentiated euclidean distance (in the spirit of Shepard (1987), given that we are interesting in making qualitative comparisons to human behavior), normalized over distances to all items in the test set to obtain generalization probabilities:

$$g(q_i, c) = \frac{e^{-d(q_i,c)}}{\sum_j e^{-d(q_j,c)}}, \quad (1)$$

where g is the generalization function, c is concept template (a single training example or the mean of several training examples), q_i is the query (test) image, and d is the euclidean distance function described above.

It has been demonstrated that human generalization behavior tightens as the number of examples increases (Tenenbaum, 2000). We can represent this phenomenon in our model through a simple augmentation:

$$g'(q_i, c) = \frac{e^{-nd(q_i,c)}}{\sum_j e^{-nd(q_j,c)}}, \quad (2)$$

where n is the number of positive examples observed.

Since we care only about the relative differences in generalization probabilities across different levels of testing stimuli, we further normalize each set of probabilities for each training set by dividing each probability by the largest $g'(q_i, c)$ in the set (e.g., the largest probability becomes 1).

To evaluate this model, we sample test and train stimuli analogous to X&T from 88/308 of the basic level labels in our set that encompassed at least 3 subordinate classes, a requirement for the final three-example experiment condition.

Generalization Results

The results of the generalization experiments from all four of our models are shown in Figure 4 (see top left and all bottom plots). To some extent, all models exhibit a basic level bias (there is some non-zero probability assigned to generalization at the basic level, even when a single example is given). However, this effect is extremely weak in the Sub model. In all other models that involve basic labels, a stronger basic-level bias is apparent after training on 1 example or 3 subordinate examples, strongly resembling more human-like trends. In addition, this basic-level generalization decreases as the model processes more examples (compare 1 sub to 3 sub conditions). Moreover, most models also exhibit subordinate and basic-level generalization after observing 3 examples drawn from a basic-level category, further resembling humans. Note that the observed basic-level bias for the Sub training, just by capturing the perceptual similarities present in the data, to some extent learns about the higher-level similarities (groupings) of the examples. However, adding nearly any form of higher-level supervision (*i.e.*, using basic-level labels) produces higher basic-level generalization and a better match to human behavior. We also observe a basic-level bias with g (not shown), but without incorporating the number of examples into the formulation, the basic-level generalization is larger than that of people in the 3-sub condition.

Conclusion

Image classification models are often trained on images paired with single labels that correspond to the subordinate level of the hierarchical taxonomy. People, on the other hand, use multiple labels to refer to the same entity in the world (*e.g.*, dog and Dalmatian). This explicit use of multiple labels makes the category learning problem easier as people can use the labels to identify similarities between the concepts (in addition to other features, such as perceptual cues, that signal similarity). Moreover, the use multiple labels helps form hierarchical representations that capture different levels of abstraction. For example, two Dalmatians are more similar to each other than a Dalmatian to a poodle; dalmatians and poodles are still highly similar and should share more relevant features than a Dalmatian and a cow (body size as opposed to black and white spots). This hierarchical nature of the representation is in turn helpful in categorizing and labeling a new entity: a child can use the similarity of an unobserved breed of dog (a Great Pyrenees) to previously-seen dogs to categorize and infer a label for it.

In this paper, we show that training on basic-level and subordinate labels (or just basic-level labels) results in representations that better capture the hierarchical structure of taxonomy of real-world objects. We also show that these representations result in a better match to the basic-level bias observed

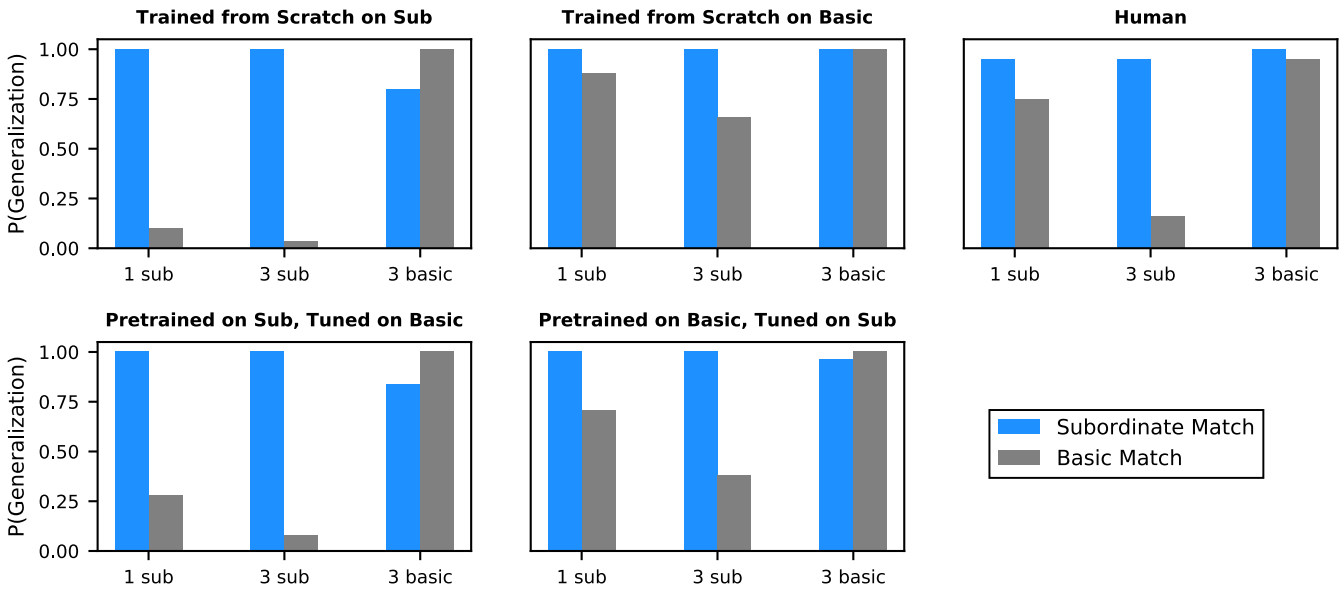

Figure 4: The results of generalization experiments from our trained models and the human data from X&T's experiments.

in human generalization behavior.

Interestingly, we observe that a relatively big portion of ImageNet (10%) are subordinate dog labels. Future work should look at other datasets with more interesting, less skewed basic label stratifications. Additionally, we only explore basic and subordinate levels in this work, which are both known to share many features in common with their members (Rosch, 1978), whereas high-level superordinate classes (e.g., furniture) often have members with few shared features.

Acknowledgements. This work was supported by grant number 1718550 through the National Science Foundation.

References

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to voxels: modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*.

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What makes an object memorable? In *Proceedings of the ieee international conference on computer vision* (pp. 1089–1097).

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of child language*, *21*(01), 125–155.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Cogsci*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lei, J., Guo, Z., & Wang, Y. (2017). Weakly supervised image classification with coarse and fine labels. In *Computer and robot vision (crv), 2017 14th conference on*.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. Mit Press.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in psychology*, *4*.

Nematzadeh, A., Grant, E., & Stevenson, S. (2015). A computational cognitive model of novel word generalization. In *EMNLP Proceedings* (pp. 1795–1804).

Peterson, J., Abbott, J., & Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In D. Grodner, D. Mirman, A. Papafragou, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2363–2368). Austin, TX: Cognitive Science Society.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (p. 111-144). New York, NY: Academic Press.

Rosch, E. (1978). Principles of categorization. In *Cognition and categorization* (p. 27-48).

Rosch, E., Mervis, C. B., Gray, W. D., M, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z.-n. (2015). *Rethinking the inception architecture for computer vision. corr abs/1512.00567 (2015)*.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In *Advances in neural information processing systems* (pp. 59–65).

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. In *Computer vision and pattern recognition (cvpr), 2016 ieee conference on* (pp. 2285–2294).

Wang, P., & Cottrell, G. W. (2015). Basic level categorization facilitates visual object recognition. *CoRR*. Retrieved from http://arxiv.org/abs/1511.04103

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.