Learning deep taxonomic priors for concept learning from few positive examples

Erin Grant (eringrant@berkeley.edu)
Department of Electrical Engineering & Computer Sciences, University of California, Berkeley

Joshua C. Peterson (joshuacp@princeton.edu)
Department of Computer Science, Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University

Abstract

Human concept learning is surprisingly robust, allowing for precise generalizations given only a few positive examples. Bayesian formulations that account for this behavior require elaborate, pre-specified priors, leaving much of the learning process unexplained. More recent models of concept learning bootstrap from deep representations, but the deep neural networks are themselves trained using millions of positive and negative examples. In machine learning, recent progress in meta-learning has provided large-scale learning algorithms that can learn new concepts from a few examples, but these approaches still assume access to implicit negative evidence. In this paper, we formulate a training paradigm that allows a meta-learning algorithm to solve the problem of concept learning from few positive examples. The algorithm discovers a taxonomic prior useful for learning novel concepts even from held-out supercategories and mimics human generalization behavior—the first to do so without hand-specified domain knowledge or negative examples of a novel concept.

Keywords: concept learning; deep neural networks; object taxonomies

Introduction

One of the hallmarks of human intelligence is the ability to rapidly learn new concepts given only limited information (Lake et al., 2016). This task is difficult because we are often presented with only a handful of (positive) examples of a new concept, and no examples outside of the concept (negative examples). Quine (1960) was the first to recognize that this poses a seemingly crippling problem for induction: hearing only the word "gavagai" as a rabbit passes by, we have no way of knowing with certainty whether the new word applies to all animals, all rabbits, one pet rabbit, potential food, or any other of a nearly infinite number of likewise compatible hypotheses.

Nevertheless, humans appear to possess prior knowledge, whether learned, innate, or both, that makes for effective generalizations even under such conditions. In some situations, these constraints are simple and easy to model (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001; Kemp et al., 2007). However, in general, modeling the rich prior knowledge that humans bring to bear on problems in complex domains such as natural images is difficult and reliant on explicit domain knowledge (Xu & Tenenbaum, 2007; Jia et al., 2013). A recent line of follow-up work has made strides by using deep neural networks as a proxy for psychological representations (Campero et al., 2017; Peterson, Soulos, et al., 2018). Although these representations are largely perceptual, they are nevertheless an improvement over hand-specified features given that they are less prone to experimenter bias and have been shown to explain some aspects of human visual representations (Peterson, Abbott, & Griffiths, 2018). However, unlike most cognitive models of concept learning and unlike humans, these networks are trained on millions of both positive and negative examples of mutually exclusive categories. Moreover, they fail to capture the taxonomic biases that humans bring to bear in concept learning (Peterson, Abbott, & Griffiths, 2018).

Challenged by the cognitive science community (Lake et al., 2015), machine learning researchers have developed a number of their own improvements to deep learning algorithms to tackle the problem of learning from few examples (e.g., Vinyals et al., 2016; Ravi & Larochelle, 2017). These approaches constitute impressive new candidate accounts of human concept learning from naturalistic stimuli, but differ from human learning scenarios in that they (1) rely on negative evidence to infer the extent of a novel concept, and (2) ignore the overlapping and hierarchical structure of real-world concepts that humans use to inform their generalization judgments (Rosch et al., 1976; Xu & Tenenbaum, 2007).

In the following paper, we aim to address many of the shortcomings of previous work by demonstrating how a deep meta-learning algorithm combined with a novel stimulus sampling procedure can provide an end-to-end framework for modeling human concept learning, for the first time with no hand-specified prior knowledge or negative examples of a novel concept. We introduce a new, taxonomically structured dataset of concepts compiled by sampling from both internal nodes and leaf nodes within the ImageNet hierarchy (Deng et al., 2009). Our method learns concepts at different levels of this hierarchy, but the hierarchical structure itself is never provided to the model explicitly at any point. To evaluate our model against human behavior, we present a new human benchmark inspired by Rosch's classic object taxonomies (Rosch et al., 1976). Our model not only mimics human generalization behavior, reproducing classic generalization gradients (Shepard, 1987; Xu & Tenenbaum, 2007), but also encompasses a general taxonomic prior that allows for human-like generalization even when presented with novel concepts from different image taxonomies (*i.e.,* held-out supercategories).

Background

Computational models of concept learning in cognitive science have historically focused on the problem of density estima-

tion (Ashby & Alfonso-Reese, 1995). Under this paradigm, learning about a category C amounts to the estimation of the density $p(x \mid C)$, where x represents the space of stimuli. This modeling framework assumes that a density can be learned for each of a set of mutually exclusive categories, where positive examples from one category implicitly serve as negative examples for all other categories. However, the conditions under which humans learn concepts are rarely this straightforward.

Learning concepts from few positive examples. More recent work has begun to examine how humans learn concepts in more natural settings where often only a few positive examples of a single concept are provided. Despite this impoverished learning environment, even young children are able to generalize surprisingly well (Carey, 1978; Markman, 1991). Extending Shepard (1987), Tenenbaum (1999) and Tenenbaum and Griffiths (2001) formalize the concept learning problem as follows: Given n positive examples $\mathbf{x} = \{x_1, ..., x_n\}$ of a concept C, the learner estimates the probability $p(x^* \in C \mid \mathbf{x})$ that a new stimulus x^* is also an example of that concept. The challenge the learner faces in making such a generalization is that the extension of C is underspecified (*i.e.,* it could include only the present examples, all possible stimuli, or anything in between). To address this challenge, the authors propose a Bayesian generalization model that averages the predictions made by a number of hypotheses about the extent of C. By making the plausible assumption that learners expect examples to be randomly sampled from concepts, the authors show that smaller hypotheses will be preferred, thus deriving constraints on the expected extent of C.

Armed with this framework, Xu and Tenenbaum (2007) conducted an extensive analysis of human generalization behavior through *word learning* experiments. Participants were given either one or three examples of a new concept such as "dax" and asked to pick out other instances of that concept from a set of test stimuli. The examples of each concept were unique images that could be drawn from either a subordinate-level (e.g., Dalmatian), basic-level (e.g., dog), or superordinate-level (e.g., animal) category, and the test stimuli were sampled from all three levels. An example of this task is shown in Figure 1. Replicating Shepard (1987), the authors found that generalization from a single example of a concept to a test stimulus decreases with psychological similarity. However, their experiments also yielded two new insights into human concept learning:

1. Given multiple examples of a concept, generalization goes only as far at the most specific level that contains those examples. For example, shown three examples from different dog breeds, other dog breeds are included in the concept at test time, but not other animals.
2. There is a bias towards generalizing to test items at the basic level, in particular when only a single subordinate example is shown. For example, given a single example of a Dalmatian, participants predictably generalize the concept to other Dalmatians, but also generalize to other breeds.

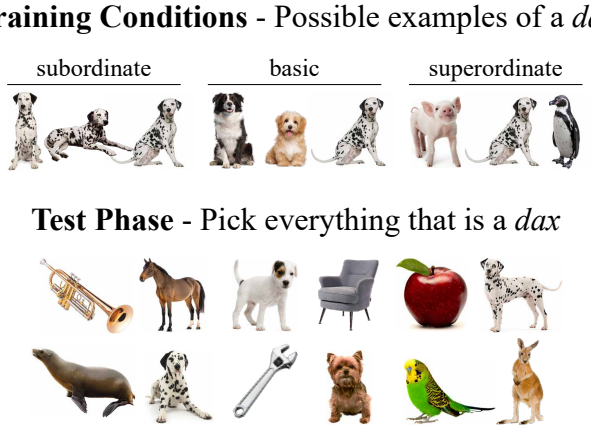

Figure 1: The *word learning* paradigm from Xu and Tenenbaum (2007). In each trial, participants see a few instances exemplifying a novel word such as "dax" and are asked to select other instances that fall under the same word from a test array. The training conditions vary by the levels of the underlying image taxonomy from which the instances are drawn, *e.g.,* Dalmatians (subordinate) vs. dogs (basic) vs. animals (superordinate).

The only modification to the Bayesian concept learning model required to capture these data was a structured, taxonomic prior computed from human similarity judgments over the set of objects used in the experiments. While this work constitutes one of the first successful attempts to explain concept learning in realistic contexts, it arguably leaves much of the structured, taxonomic representation assumed and raises questions about how this knowledge is acquired.

The role of prior knowledge. Given the aforementioned dependence on highly structured priors in explaining people's robust generalization behavior, subsequent work has focused on incorporating this information into the modeling of human concept learning. Jia et al. (2013) provided an automated framework for modeling human generalization behavior by leveraging perceptual stimulus features provided by a computer vision algorithm along with information contained in the WordNet taxonomy (Fellbaum, 1998), but gave no account for how this information is learned by humans. Kemp et al. (2007) provided the first account of how such knowledge could be acquired: The authors start with an unstructured representation and apply a structured hierarchical Bayesian model that learns taxonomic abstractions from data. Despite its elegance, the method does not immediately scale to high-dimensional stimuli such as the images used in Jia et al. (2013).

Deep neural networks (LeCun et al., 2015) have served as both candidate models of object perception and rich image representations that can be used for cognitive modeling. However, these model do not capture even coarse taxonomic information out-of-the-box (Peterson, Abbott, & Griffiths, 2018). Despite this, Peterson and Griffiths (2017) found that the sampling assumptions of Bayesian concept learning could be verified in human generalization judgments when modeling stimuli using deep feature representations. Campero et al. (2017) deployed a hierarchical model similar to Kemp et al. (2007) over a deep

Superordinate	Basic	Subordinates	
Musical Instrument	Guitar	Acoustic guitar	Electric guitar
	Piano	Grand piano	Upright piano
	Drum	Tambourine	Bass drum
Fruit	Apple	Delicious apple	Mackintosh apple
	Currant	Black currant	Red currant
	Grapes	Concord grapes	Thompson seedless grapes
Tool	Hammer	Ball-peen hammer	Carpenter's hammer
	Saw	Hack saw	Cross-cutting saw
	Screwdriver	Phillips screwdriver	Flat tip screwdriver
Clothing	Trousers	Jeans	Sweat pants
	Socks	Athletic socks	Knee-high socks
	Shirt	Dress shirt	Polo shirt
Furniture	Table	Kitchen table	Dining-room table
	Lamp	Floor lamp	Table lamp
	Chair	Armchair	Straight chair
Vehicle	Car	Sports car	Sedan car
	Airplane	Airliner plane	Fighter jet plane
	Truck	Pickup truck	Trailer truck
Fish	Snapper	Grey snapper	Red snapper
	Trout	Rainbow trout	Lake trout
	Salmon	Atlantic salmon	Chinook salmon
Bird	Owl	Barn owl	Great gray owl
	Eagle	Bald eagle	Golden eagle
	Sparrow	Song sparrow	Field sparrow

Table 1: The eight taxonomies adapted from Rosch et al. (1976).

feature space and found both good one-shot learning performance as well as the ability to recover some stimulus clusters representative of human categorization judgments. Noting that most deep networks are trained using subordinate-level labels, Peterson, Soulos, et al. (2018) trained a deep neural network with coarser, basic-level labels to more closely mimic the supervision children receive. A relatively simple generalization model over the resulting representation reproduced both the basic-level bias and the gradient of generalization from Xu and Tenenbaum (2007).

Few-shot learning in machine learning. The problem facing cognitive models of concept learning is closely related to *one-* or *few-shot* classification in machine learning, in which the aim is to learn to discriminate between classes given only a few labeled examples from each class (Fei-Fei et al., 2003; Vinyals et al., 2016). A powerful solution to few-shot learning is *meta-learning*, where learning episodes—themselves consisting of training and testing intervals—are used to train a model to adapt quickly to solve a new task given only a small amount of labeled task data (Schmidhuber, 1987). The learning episodes are leveraged in the form of a data-driven prior that is combined with a small amount of test-time evidence (*i.e.,* a few "shots" of labeled data from a novel task) in order to make a test-time inference.

Modeling Approach

We propose to bridge cognitive science and machine learning by formulating concept learning as a few-shot learning problem. As we will see, the meta-learning problem formulation allows a machine learning model to estimate a decision boundary from only positive samples of a class, similarly to how people learn concepts from only a few positive examples. Moreover, the use of a meta-learning algorithm provides a principled way to present entirely novel concepts at test time as held-out test *tasks*. As such, we can investigate the taxonomic priors encoded in a neural network embedding function, as compared to prior work that examines the representations of images from categories observed during training time (Peterson, Soulos, et al., 2018).

Concept learning as meta-learning. Meta-learning algorithms aim to learn how to learn by extracting task-general knowledge through the experience of solving a number of specific tasks (Thrun & Pratt, 1998; Hochreiter et al., 2001). In the case of concept learning, the jth task corresponds to learning a decision boundary for the jth concept using only positive examples, and meta-learning corresponds to learning how to estimate decision boundaries for arbitrary unseen concepts. We can thus formalize the concept learning problem as the task of predicting a target label y (which indicates whether or not the input belongs to a given category) from an input observation x (i.e., an image). Note that this formulation differs from the standard discriminative classification problem, where the task corresponds to a K-way discriminative classification task in which each of the K class labels are mutually exclusive.

Formally, let $\mathcal{T}_j = (\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}}, \boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$ denote a task drawn from a given task distribution $p(\mathcal{T})$, where $\boldsymbol{X}_j^{\text{trn}}$ and $\boldsymbol{Y}_j^{\text{trn}}$ are a small collection of training inputs and labels, disjoint from validation samples $\boldsymbol{X}_j^{\text{val}}$ and $\boldsymbol{Y}_j^{\text{val}}$ but belonging to the same task $\mathcal{T}_j$. A meta-learning algorithm (*e.g.,* Vinyals et al., 2016; Ravi & Larochelle, 2017) aims to estimate parameters $\boldsymbol{\theta}$ that can be adapted to solve an unseen task $\mathcal{T}_j \sim p(\mathcal{T})$, using only the training samples $(\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}})$, to ensure the updated model achieves good performance on the validation samples $(\boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$ according to some loss function $\mathcal{L}$.

In this work, we use the *model-agnostic meta-learning* (MAML; Finn, Abbeel, & Levine, 2017) algorithm, which formulates meta-learning as estimating the parameters $\boldsymbol{\theta}$ of a model so that when one or a few gradient descent steps are taken from the initialization at $\boldsymbol{\theta}$ on the training data $(\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}})$, the updated model has good generalization performance on that task's validation set, $(\boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$. At test time, a new task from the test set is presented to the model for few-shot adaptation, *i.e.,* gradient descent with $(\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}})$, and computation of test-time performance metrics, *e.g.,* accuracy on $(\boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$. The training examples in the inner gradient computation are strictly positive examples (*i.e.,* $\boldsymbol{Y}_j^{\text{trn}} = 1$) of a particular concept j, whereas validation examples in the outer gradient computation include both positives and negatives (*i.e.,* $\boldsymbol{Y}_j^{\text{val}} \in \{0, 1\}$); thus, at test time, the meta-learning algorithm is able to estimate a decision boundary for a novel concept from only positive examples of that concept.

Behavioral Experiment

In order to compare our method directly to human behavior, we conducted a large-scale human generalization experiment using a test set of naturalistic stimuli used for the simulations in the next section. We assess generalization behavior using a concept learning experiment that follows previous work on Bayesian concept and word learning (Xu & Tenenbaum, 2007; Abbott et al., 2012; Jia et al., 2013).

Stimuli. We mapped a subset of the graph structure embedded in the ImageNet dataset used for the ImageNet Large Scale Visual Recognition Competition (ILSVRC; Russakovsky et al., 2015) to the classic taxonomy used by cognitive scientists and developed by Rosch et al. (1976). ILSVRC is a commonly used object-classification dataset that contains more than 1 million images distributed across 1000 categories. Instead of using the leaf classes as categories, we create concepts by picking a node in the ImageNet hierarchy and sampling images from leaves dominated by the given node. Note that, in this case, concepts are not necessarily mutually exclusive in the sense that a single image may belong to one or more classes (*e.g.,* a Dalmatian may be labeled as both a *dog* and an *animal*). If the exact subordinate node from Rosch et al. (1976) was not available in ImageNet, we found a close semantic match via the WordNet (Fellbaum, 1998) taxonomy. We provide the full taxonomy for this dataset in Table 1.

Task. In each of 8 trials, participants observed 5 images of a single concept, sampled from one of the three levels of taxonomic abstraction. For instance, in a subordinate training condition, the examples could be all Dalmatians; in a basic-level training condition, all dogs; in a superordinate training condition, all animals. To test generalization behavior, participants were then given a test array of 24 images and were asked to pick which images also belonged to the learned concept. The test array comprised 2 subordinate matches (*e.g.,* other Dalmatians), 2 basic-level matches (*e.g.,* other breeds of dog), 4 superordinate matches (*e.g.,* other animals), and 16 out-of-domain items (*e.g.,* inanimate objects), following Xu and Tenenbaum (2007). See Figure 2 for an example set of training and test stimuli. In total, we collected data for 180 unique trials and 1 180 unique images.

Participants. We recruited 900 unique participants from Amazon Mechanical Turk to each complete 8 trials as described above, one randomly sampled for each of the superordinate categories. The test sets were fixed within a superordinate category. Participants were paid $0.40 each.

Results. Figure 3 (a) presents the results of the behavioral experiment for each of the three taxonomic levels. As expected on the basis of previous work, there is an exponentially decreasing generalization gradient as the level of taxonomic abstraction of the test matches (bar color) increases. However, this effect diminishes as the intra-class variation of the few-shot examples (x-axis) increases: Moving from the *subordinate* condition to the *basic*-level condition, we find an increase in the number of basic-level matches selected from the test set. The condition in which there is greatest intra-class variation–the superordinate condition–exhibits only a small generalization gradient.

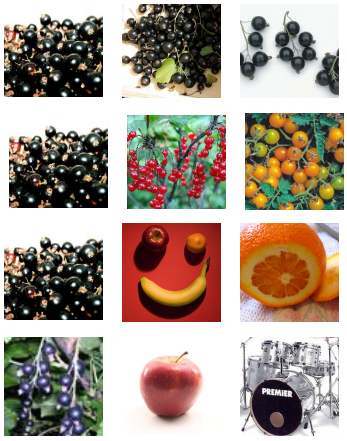

Figure 2: Examples of training stimuli for the **(a)** subordinate, **(b)** basic-level, and **(c)** superordinate level training conditions, as well as **(d)** a subset of the stimuli from the test array for a specific concept learning task (here, learning the concept *black currant* **(a)**, *currant* **(b)** or *fruit* **(c)**). The test array **(d)** displays, from left to right, a subordinate match, a basic-level match and a superordinate match.

Meta-Learning Simulations

Our modeling goal is to investigate whether we can use meta-learning to learn new concepts from only few positive examples, even though these concepts are potentially overlapping and therefore not mutually exclusive. Furthermore, we aim to investigate whether a meta-learning algorithm is able to use information about the underlying concept taxonomy that generates observations of the extension of a concept in order to generalize to novel concepts in a human-like manner.

Meta-learning formalism. Our model observes K positive examples $\boldsymbol{x} = \{x_1, \ldots, x_K\}$ of a concept C, and must learn the generalization function $p(x^* \in C)$ to correctly identify whether a novel example x^* is also a member of the concept. Training proceeds as follows: A concept index j is sampled from the meta-training set. Then, for K-shot learning, $2K$ positive examples of the concept and K negatives are sampled. The parameters $\boldsymbol{\theta}$ are adapted using K of the positives, and then the model is optimized with a loss computed using the remaining positive and negative examples of the concept. At test time, the model with trained parameters θ is presented with K positive examples from a new concept in the test set; the model adapts $\boldsymbol{\theta}$ and is evaluated on its ability to distinguish new positive examples of that concept from negatives.

Taxonomic dataset construction. For training and validation, we created a large-scale taxonomy of classes by using the graph structure embedded in the subset of the ImageNet dataset used for the ImageNet Large Scale Visual Recognition Competition (ILSVRC; Russakovsky et al., 2015), similar to the behavioral experiment described earlier, but using the entirety of the ImageNet hierarchy. We then created few-shot concept learning tasks for training by sampling positive and negative examples for each concept, where negative examples of a concept are generated by sampling from the complement set of leaf nodes. Superordinate-level nodes are not shared between training, validation, and test to ensure that test-time generalization is measured on novel concepts. We use 494, 193, and 223 leaf nodes in the training, validation, and test sets, respectively (*c.f.,* 80, 20, and 20 in the few-shot classifi-

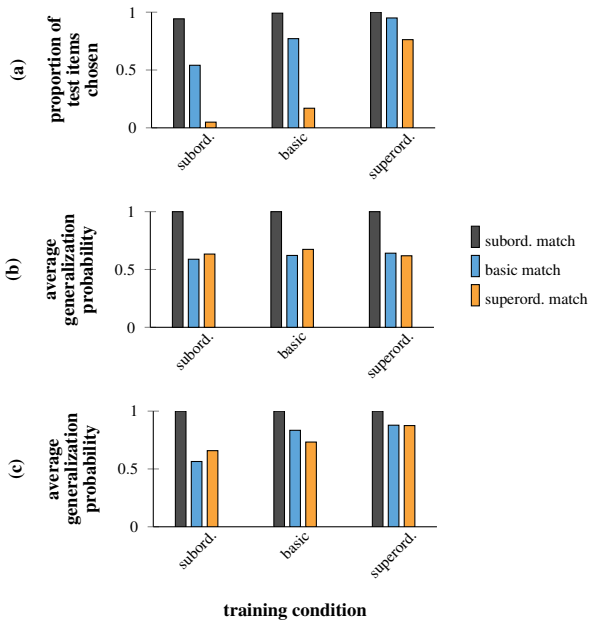

Figure 3: Human behavioral data **(a)** and `flat` **(b)** and `hier` **(c)** modeling results on the concept generalization task. The horizontal axis identifies the training condition (*i.e.,* the level of taxonomic abstraction from which the few-shot examples are drawn). The vertical axis identifies, for each type of match in {subordinate, basic-level, superordinate}, the proportion of selections from the test array **(a)**, or the average probability of generalization **(b, c)**.

cation dataset *mini*ImageNet (Vinyals et al., 2016)). The training, validation, and test node sets do not comprise all of the nodes in the ImageNet hierarchy, as some nodes are redundant (*i.e.,* have a single parent) or are too abstract to appropriately define a visual concept (*e.g., physical entity*, *substance*, *equipment*). We make use of the training and validation dataset for training and hyperparameter selection, respectively; the test set is not used in this work but reserved for future works that may wish to perform large-scale evaluation of concept learning. Instead, the evaluations reported in this work are performed on the Rosch-inspired human benchmark described above. We also wish to emphasize that while we make use of the ImageNet hierarchy, we do so only to generate a natural distribution of concepts to learn from, and never present the explicit hierarchical relations to the model at any time.

We consider two dataset conditions in our simulations: In the `hier` dataset condition, the meta-learning algorithm observes concepts sampled from the internal and leaf nodes of the ImageNet hierarchy, and thus can learn a taxonomic prior; in the `flat` dataset condition, the algorithm observes only leaf-node concepts, and thus has no access to such information.

Hyperparameters. The base model that is optimized by model-agnostic meta-learning (MAML) is a binary classifier consisting of a convolutional neural network with a sigmoid output.[1] In our experiments, we downsample the images to each have a width and height of 84 pixels, as is common in the use of *mini*ImageNet (Vinyals et al., 2016) as a few-shot learning dataset. We select hyperparameters on the same hierarchically structured validation set for both the `hier` and `flat` dataset conditions and evaluate algorithms after a fixed number of training iterations (40K with a batch size of 4). We take the value of the scalar output of the network evaluated on a test example as the *generalization probability* and average this quantity across all test examples from a specific level of taxonomic match to produce the *average generalization probability*. When reporting the average generalization probability metric, we standardize each set of probabilities for each training condition by treating the distractor (out-of-domain) generalization probability as a baseline of zero and further dividing by the largest probability in the set. In line with prior work (Peterson, Abbott, & Griffiths, 2018), this highlights the quantity of interest: the relative differences in average generalization probabilities across the subordinate, basic-level, and superordinate levels of the taxonomy.

Results. The generalization gradient observed in humans is also exhibited by the `hier` dataset condition in Figure 3 (c): When the few-shot examples are taken from a basic-level category (the *basic* condition; *e.g.,* different breeds of dog) as opposed to a subordinate category (the *subord.* condition; *e.g.,* Dalmatians), the model generalizes to more basic-level matches (*e.g.,* different dog breeds) from the test array. In the plot, this can be seen by comparing the ratio of subordinate generalization (black column) to basic-level generalization (blue column) within each training condition (*i.e.,* the gap between the black and blue bars is diminished in the *basic* condition *vs.* the *subord.* condition). Furthermore, when the few-shot examples are taken from a superordinate category (*superord.* condition), both the model in the `hier` dataset condition and humans are equally likely to pick subordinate, basic-level, or superordinate matches from the test array. In Figure 3 (a, c), this can be seen as the generalization to all levels of the taxonomy (black, blue, and yellow bars) being close to equal.

One notable departure of Figure 3 (c), from the human generalization behavior in Figure 3 (a), is overgeneralization to the superordinate category in the subordinate training condition, and to a lesser extent, in the basic-level training condition, suggesting that it is difficult for the algorithm to discriminate between basic-level and superordinate matches given only subordinate examples of a concept. Nevertheless, in comparison to the `flat` dataset condition in Figure 3 (b), which does not change generalization behavior on the basis of the training condition, the behavior of the algorithm exposed to the hierarchically structured `hier` dataset suggests a learned sensitivity to the underlying taxonomic organization of new concepts.

[1] The architecture of the model is similar to prior work in meta-learning (e.g., Ravi & Larochelle, 2017) with 4 convolutional layers each with 32 3×3 filters, leaky ReLU activation functions with a slope of 0.2, and 2×2 max-pooling, all followed by a linear layer with sigmoid activation. We do not employ batch normalization because of strong batch interdependence, as all of the training examples for a concept are of the same (positive) class.

Discussion

When humans are presented with an example from a new concept, they can quickly infer which other instances belong to that same concept even without the strong constraints provided by negative examples. In order to achieve this feat, humans bring to bear information about the taxonomic structure of natural categories. Targeting the robustness of human generalization even in highly novel domains (Schmidt, 2009), we investigated the extent to which taxonomically structured biases for complex, naturalistic stimuli taken from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) could be acquired and leveraged to learn the extent of novel concepts from only a few positive examples. In contrast to previous work (Peterson, Abbott, & Griffiths, 2018), we validate the generalization behavior of our model using *unseen* supercategories drawn from the superordinate levels of Rosch's classic taxonomy (Rosch et al., 1976).

While our method is successful in both learning a general taxonomic prior and exhibiting human-like generalization behavior, there is room for improvement as the quantitative gradients are not a perfect match to humans. However, it should be noted that our model faces the atypically challenging task of both learning a highly structured representation for complex stimuli and making use of it to generalize to entirely novel concepts. As such, this framework draws on many of the strengths of both cognitive models and deep neural networks in machine learning, and constitutes the most comprehensive account of human visual concept learning to date. Lastly, we note that we do not build in any explicit preference for simple concepts or attention to the number of examples (Tenenbaum, 1999; Peterson, Soulos, et al., 2018), although this may be an interesting avenue for improvement in future work.

Acknowledgments. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Lifelong Learning Machines (L2M) program via grant number HR001117S0016 and by the National Science Foundation (NSF) under grant number 1718550. The views and opinions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Government.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Constructing a hypothesis space from the web for large-scale Bayesian word learning. In *Proceedings of the 34th annual meeting of the cognitive science society (cogsci)*.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Campero, A., Francl, A., & Tenenbaum, J. B. (2017). Learning to learn visual object categories by integrating deep learning with hierarchical bayes. In *Cogsci*.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). MIT Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 248–255).

Fei-Fei, L., et al. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the 9th conference on computer vision and pattern recognition (cvpr)* (pp. 1134–1141).

Fellbaum, C. (1998). *Wordnet*. Wiley Online Library.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th international conference on machine learning (icml)*.

Hochreiter, S., Younger, A., & Conwell, P. (2001). Learning to learn using gradient descent. *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 87–94.

Jia, Y., Abbott, J. T., Austerweil, J. L., Griffiths, T., & Darrell, T. (2013). Visual concept learning: Combining machine vision and Bayesian generalization on concept hierarchies. In *Advances in neural information processing systems (nips) 26* (pp. 1842–1850).

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307–321.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 1–101.

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

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

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Peterson, J. C., & Griffiths, T. L. (2017). Evidence for the size principle in semantic and perceptual domains. *arXiv preprint arXiv:1705.03260*.

Peterson, J. C., Soulos, P., Nematzadeh, A., & Griffiths, T. L. (2018). Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *arXiv preprint arXiv:1805.07647*.

Quine, W. V. O. (1960). Word and object, 1960. *Le mot et la chose*, 1977–2000.

Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In *Proceedings of the 5th international conference on learning representations (iclr)*.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning* (Unpublished doctoral dissertation). Institut für Informatik, Technische Universität München.

Schmidt, L. A. (2009). *Meaning and compositionality as statistical induction of categories and constraints* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

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

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Thrun, S., & Pratt, L. (1998). *Learning to learn*. Kluwer Academic Publishers.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems (nips) 29* (pp. 3630–3638).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.