Learning to Learn Visual Object Categories by Integrating Deep Learning with Hierarchical Bayes

Andres Campero (campero@mit.edu)
Andrew Francl (francl@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, MA 02139 USA

Abstract

Humans are capable of generalizing and learning new concepts after very little experience. They have the ability to create semantic structures from concepts they acquire, they can learn appropriate inductive biases that are later used as priors for different tasks, and they can learn novel categories from very few examples. While recent advances in neural networks and other machine learning methods are beginning to approach human-level capabilities in several tasks, building computational models that replicate these abilities has proven difficult. We propose a model that combines powerful features extracted from a deep neural network with a semantic structure inferred using probabilistic Hierarchical Bayes. We test and demonstrate the capabilities of our model in three different tasks: learning a new concept from a single example of a novel category, learning new categories from few examples of different categories, and learning the semantic tree from an unlabeled set of novel objects.

Keywords: hierarchical bayes; one-shot learning; inductive bias; neural networks; unsupervised learning

Introduction

Recent advances in neural networks and other machine learning methods have led to computer vision object-recognition systems that are beginning to approach human-level performance. Trained on thousands of object categories, with thousands of labeled examples for each, deep convolutional networks can tell if a new image contains a familiar category almost as well as human adults can in a brief glance. Yet, even young children have abilities to learn and generalize that go beyond what current machine vision systems can do. Here we focus on three such abilities:

(1) By age 3, children can learn new object categories from just a single example. Furthermore, children generalize in different ways as appropriate for different kinds of categories: labels for artifacts with functionally relevant shapes are preferentially generalized according to those shapes, while labels for non solid substances or arbitrarily shaped objects are more likely to be generalized according to material properties.

(2) Children can learn to learn appropriate inductive biases, such as the shape and material biases described above, from experience with just a few examples each of a small number of categories that exemplify these biases in a consistent way. The shape-bias training studies of Smith and colleagues are the best known examples (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002).

(3) Children can, in a completely unsupervised way, sort novel objects into categories and supercategories in a meaningful way, and then use these hierarchical category structures as strong constraints to learn and generalize names for objects from just one or a few examples.

Previous attempts to capture these abilities in computational models have had some success, but not with models that are "image-computable" on the same stimuli that people see. These earlier models have used either adult similarity judgments (Xu & Tenenbaum, 2007) or highly simplified, idealized feature representations (Kemp, Perfors, & Tenenbaum, 2007) to build their category hierarchies. Here we show that a computational framework can come close to capturing abilities (1-3) by combining two powerful representation-learning techniques: deep learning for feature construction and Hierarchical Bayes for unsupervised taxonomy construction.

We build on work by Salakhutdinov, Tenenbaum, and Torralba (2012) who build a Hierarchical Bayesian model that "learns to learn" by incorporating information from past experience into a prior when inferring statistical properties of a novel category. In particular, when presented with a few image examples of a new category, the model infers a supercategory and uses the higher-order knowledge abstracted from previous categories to identify the relevant features and allow generalization (Figure 1).

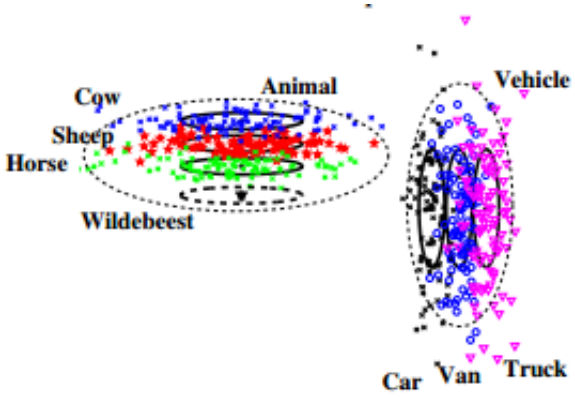

Figure 1: Learning a similarity metric for a new category. The goal is to identify the correct supercategory and estimate an appropriate similarity metric.

That work was extended by the same authors, who harnessed a two layer Deep-Boltzmann Machine to generate low level feature representations of the images while learning a prior using a hierarchical Dirichlet process. (Salakhutdinov,

Tenenbaum, & Torralba, 2013). Their experimental data showed that using this prior in combination with more powerful features gave them a distinct advantage over other methods of classification. This progression of work suggests that building a model that combines complex feature spaces with a hierarchical semantic structure may lead to further increases in performance.

Building on this line of work, we contribute a model that combines the two components: powerful image representations extracted from Deep Neural Networks (DNNs) and a Hierarchical semantic structure that works as a Bayesian prior. We show how the combination of these two components can "learn to learn" in ways that resemble some aspects of child cognition. Additionally, we explore how this model's performance is affected as we vary different aspects of the model architecture and the structure of the training data.

Other approaches to combine probabilistic graphical models and DNNs have recently been proposed that focus on building unsupervised clustering algorithms (Dilokthanakul et al., 2016; Johnson et al. 2016). Instead, the focus of our model is to capture certain aspects of human cognition. This leads to some notable differences. First, representations in our model are a fixed set of visual relevant features instead of being learned for the inference task at hand. In addition, our model's generative component is limited to a hierarchical structure that aims to recover the semantic relations between concepts in a useful and meaningful way while other models are fully generative but tend to have graphs with simpler semantic structures. We therefore propose a relatively simple model that is not intended for general unsupervised learning but that instead focuses on traits of human object and category learning.

More specifically, we test our model's capacity to capture the previously discussed human abilities (1-3) in an image recognition framework. First, we evaluate the ability of our model to learn novel categories from only one or a few examples. To address this we allow the model to construct a semantic structure from labeled examples in a data set and then judge the model's performance on a one-shot learning task. Second, we assess the models capability to construct inductive biases in low data environments. We test this ability by repeating the first task but limiting the training data available to the model when it constructs the semantic tree. Finally, in a third task, we test the model's ability to learn a hierarchical semantic structure of novel objects in a completely unsupervised manner. Results suggest that this approach may be suitable for modeling certain aspects of cognition.

Model and Learning to Learn

Our model combines two Machine Learning approaches that have recently been successful at a range of differing tasks. On one hand, powerful deep networks construct feature spaces that enable rapid and accurate classification. On the other, Hierarchical Bayesian Models have proven successful in creating taxonomies of the different concepts learned from previous experience. These taxonomies can then be used as a prior to identify the relevant features for learning a new category from one or a few examples based on the distribution of other similar categories. We create various versions of our model to compare combinations of feature spaces extracted from different architectures with variants of the Hierarchical Bayesian component.

Learning begins by constructing a 2-level tree of categories and supercategories that best explains the training observations under a Bayesian framework. The model learns structure in the observations by first generating useful general features from a DNN and then developing hierarchical priors that allow previous similar experiences to bias the learning of new concepts and categories. The priors are constructed by inferring the means and variances that define the most relevant dimensions from the DNN feature representations for each category and supercategory (Figure 1).

Deep Network Features

We use features extracted from DNNs pretrained for object classification on ImageNet. We obtain a representation from each image by passing it through a network and extracting the response from the penultimate layer consisting of 4096 real-valued dimensions. In the regular deep network classification scheme, this response is then passed through a linear weighting and a generalized logistic regression layer. This layer maps this representation onto probabilities for each class in the specific classification task for which the network was trained.

We compare the performance of the different versions of our model on features extracted from two different DNN architectures: Alexnet (Krizhevsky, Sutskever, & Hinton, 2012), which was the first implemented Deep Learning Model that significantly improved object classification on images; and VGG-16 (Simonyan & Zisserman, 2014), a more recent architecture with 16 layers that achieves above 90% top 5 classification performance on ImageNet.

Generative Semantic Organization

After obtaining a useful general image representation from the DNN, the Hierarchical Bayesian Model's parameters are inferred by approximating the posterior via Markov Chain Monte Carlo methods in the following way.

Consider a two-level hierarchy where N observed inputs are partitioned into C basic-level categories, these categories are in turn partitioned into K supercategories. In this hierarchy of observations, categories, and supercategories, the higher levels determine a prior over the distribution of the lower levels. In particular, the distribution over observations (feature vector representations of images in our case) of each of the different basic level categories are assumed to be multivariate Gaussian with a category specific mean M_c and with precision terms τ_c^d that are assumed to be independent across the D dimensions of the feature space. These precision terms constitute a similarity metric by determining the relative importance of each of the features. In turn, we place a conjugate

Table 1: Performance results using the area under the ROC curve (AUROC) on the MSR dataset in the one-shot learning task

	# Examples from Withheld Class							
	Alexnet				VGG			
	1ex	2ex	4ex	20ex	1ex	2ex	4ex	20ex
Oracle	.99	1	1	1				
HB-Full	.91	.96	.98	.99	.92	.97	.98	.99
One Supercategory	.87	.94	.97	.99	.88	.95	.98	.99
NearestN	.84	.86	.87	.90	.89	.90	.92	.95
T of T*	.76	.80	.84	.87				

Normal-Gamma prior over $\{M_c, \tau_c\}$, this prior is determined by the supercategory specific level-2 parameters M_k, τ_k, α_k, where M_k and τ_k constitute the expected values of the lower level parameters and α_k controls the variability of τ_c around its mean. Finally, for the conjugate priors over the level-2 parameters, we respectively assume Normal, Exponential and Inverse-Gamma distributions that are further shaped by parameters α_0 and γ_0. The full generative model is given in Figure 2 (Salakhutdinov et al., 2012).

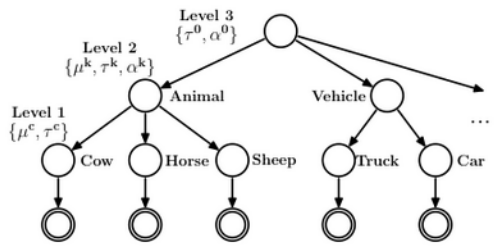

Figure 2: Hierarchical Model

Given a set of observations, the model iteratively performs Bayesian inference by alternating between sampling the parameters and inferring the category assignments. When learning the distributions at each step of the iteration, the supercategory membership is fixed and the parameters are sampled from posteriors that are analytically computed using the conjugate priors[1]. The supercategory membership for each category is learned in a similar way by fixing the currrent parameters and the rest of the hierarchical structure. Every category can be assigned to any of the existing supercategories or to a newly created one. The posterior probability of belonging to a supercategory is computed as a combination the likelihood that the parameters of the category come from the parameters of the supercategory and a Chinese Restaurant Process (CRP) prior (Griffiths & Tenenbaum, 2004). This nonparametric prior is a distribution over a partition on integers in which the n^{th} number is assigned to set k with probability:

$$P(z_n = k | z_1, z_2 ..., z_{n-1}) = \begin{cases} \frac{n^k}{n-1+\gamma} & \text{if } n^k > 0 \\ \frac{\gamma}{n-1+\gamma} & \text{if k is new} \end{cases}$$

Where n^k is the number of previous integers assigned to set k and γ is a concentration parameter sampled from a $Gamma(1,1)$ distribution.

In an unsupervised setting where the categories of the observations are also unknown, the model utilizes a similar strategy to assign observations to categories as is used when assigning categories to supercategories. The model iterates through the observations and assigns each either to an existing or to a newly created category based on the prior and likelihood. By utilizing the CRP prior, the model can create an unbounded number of categories and supercategories. This entire process constitutes a Gibbs sampling procedure where both the tree structure and all of the parameters are simultaneously learned.

Tests and Results

We test the model in scenarios that attempt to capture aspects of human cognition related to learning from limited data. First we measure the model's ability to generalize previous knowledge to learn novel categories from only a few examples. Next, we assess the model on this task when the training data for all of the categories is also limited to only a few examples. Finally, we exploit the model's full hierarchy in a completely unsupervised setting by exploring how the model recovers the underlying semantic structure.

One-Shot Learning on MSR

In the first task, we test the model's ability to learn new categories form one or a few examples. First, we select a category that will be held-out for testing. Labeled observations for all other basic-level categories are provided for training. The model learns the semantic structure of the training set by clustering the basic categories into supercategories and inferring the relevant parameters at all levels of the Bayesian Hierarchy. The challenge is then to generalize the learned structure to the held-out category from only one or a few examples.

To do this, the model first infers the best supercategory from one or a few examples of the withheld category by

[1]For the case of α_k, the conditional posterior cannot be computed analytically and the parameter is sampled with the Metropolis-Hastings rule (Yildirim, 2012).

Table 2: Performance results using the area under the ROC curve (AUROC) on the MSR dataset with limited training data.

	# Examples from Withheld Class							
	Alexnet				VGG			
	1 ex	2 ex	4 ex	20 ex	1 ex	2 ex	4 ex	20 ex
# Training Examples								
1 ex	.87	.87	.88	.89	.90	.90	.90	.92
4 ex	.92	.96	.99	.99	.93	.97	.98	.99
10 ex	.92	.96	.99	.99	.92	.96	.98	.99
18 ex	.92	.95	.98	.99	.91	.96	.98	.99
All examples	.91	.96	.98	.99	.92	.97	.98	.99

marginalizing over the category level parameters. Next, the model uses the supercategory priors and training examples to estimate the category similarity metric and mean for each dimension in the feature space.

We evaluate different versions of our model on the MSR Cambridge dataset (Kohli et al., 2005), which consists of 24 categories with varying numbers of images in each category. In total this dataset contains roughly 800 images. Figure 3 shows a typical partition over all the categories discovered by the full model. To quantify the models accuracy, a testset with unlabeled data from all categories is classified.

We repeatedly trained the model withholding one of the categories at a time and then inferred the withheld category parameters and supercategory membership using one or a few images. Next, we calculated the posterior probability for each testset image belonging to each category and variated a threshold to classify images as belonging to the heldout category or to any of the other categories. This created true and false positive rates for each point along our threshold which traced out a Receiver Operating Characteristic curve (ROC) for classifying objects from the withheld vs. all the other categories. The reported results are calculated by averaging the Area Under the ROC curve (AUROC) for the model trained with each of the 24 categories withheld (Table 1).

Performance is compared for each combination of an Inference Model and a Network Architecture. HB-Full is the full version of the model described above. One Supercategory places all the categories in the same single supercategory. NearestN classifies new points with the label of the nearest neighbor of its feature vector in euclidean distance. Texture of Textures (T of T)* replaces our DNN features with the set of responses from a three layer convolutional neural network that uses precomputed weights that resemble Gabor filters[2]. Finally, the Oracle is the same than our full model, but uses the true empirical mean and variances from the whole population (including testset). Table 1 shows the results for the two different feature spaces used.

[2]Taken from Salakhutdinov et al. (2012)

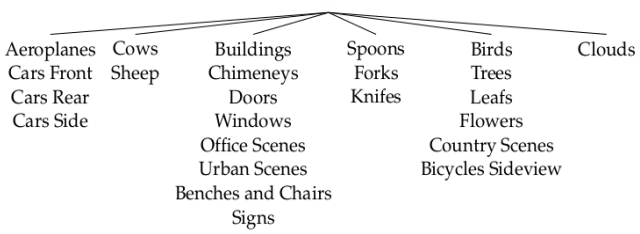

Figure 3: MSR semantic tree discovered by the Full Model

The results show that the model performs best when using the full hierarchy in combination with the feature space extracted from VGG. HB-Full considerably outperforms alternatives under both feature spaces, particularly for trials with one example from the withheld dataset. As more examples become available, the performance difference decreases reflecting the importance of the prior when little data is available. The importance of the learned features is highlighted when comparing with the T of T* feature space where performance is considerably lower. It is interesting to note that the VGG representation improves most over Alexnet when in combination with NearestN, but the effect is mitigated when the hierarchy is used.

Limited training data regimes

In a second task, we test the capability of our model to extract inductive biases from experience with just a few examples. To evaluative this capability, our full model was limited to only 1, 4, 10 or 18 examples of each category used for training. The number of examples from the withheld category was varied separately. Table 2 shows the average AUROC for the same "one vs. all" metric used in the previous task[3]. For comparison, the full model performance from the previous table is included and labeled as "All examples"[4].

We can see that the largest jump in performance happens when moving from 1 to 4 training examples. This likely reflects the fact that a single example provides information about the mean of the category but not about the variance or similarity metric, which has to be inferred completely from

[3]Averages across 10 random repetitions and all categories are reported.

[4]Each category contains a varying number of examples

the prior. However, 4 examples provide adequate information about the variance to allow the model to appropriately infer the parameters for new categories. As the number of training examples continues to increase, there are no further gains in performance. This is consistent with literature showing that children need at least two examples to learn inductive biases in certain contexts (Smith et al., 2002).

Unsupervised Learning on Gazoobian Objects

Humans and children can sort new objects into categories and supercategories in a semantically meaningful way. While our model is also able to of recover meaningful structure from labeled examples (Figure 3), real situations often demand learning where labels are completely absent. Schmidt (2009) explores this human capability with a dataset composed of 45 novel objects that were generated using a modeling software to simulate a specific taxonomic structure. The dataset consists of three supercategories supposed to be alien equivalents of plants, tools and snails from the planet "Gazoob". The objects in each supercategory are further organized into a structure that can be approximated by basic-level categories (gray box in Figure 4).

Our model has the ability to infer both categories and supercategories in an unsupervised manner from observations. Schmidt (2009) shows that a model based on agglomerative clustering that uses adult similarity judgments is able to recover the taxonomic tree (Figure 4). Here our model is tested with the harder task of recovering the taxonomic tree directly from the same images that people saw. The model accomplishes this task in a fully unsupervised manner using a single image of each object.

This "image-computable" model is able, although with some mistakes, to recover the three supercategories and most of the basic-level category structure (Figure 5). Other unsupervised clustering algorithms were also able to capture some of the semantic structure, but the hierarchy between categories and supercategories was not evident.

Discussion

One can think of the task of concept learning as consisting of two elements. The first involves obtaining relevant features to represent the objects and categories commonly observed in the world. The second involves constructing a semantic hierarchical structure with links between categories that humans can use to navigate and perform tasks. While recent results demonstrate the capabilities of DNNs to classify categories provided a large number of training examples, they struggle to perform tasks that require understanding the semantic relationships between classes. The ability of Hierarchical Bayesian Models to build these semantic structures can further help with understanding and classifying new categories.

We demonstrate how these two approaches can complement one another by combining them in a computational model. We tested the model's abilities tasks designed to approximate human capabilities that are currently difficult for computer vision systems such as concept generalization, learning inductive biases, and constructing semantic structures. We show results for three tasks involving limited data availability. The model is able to learn relevant semantic structures from just a few examples of novel objects and effectively transfer appropriate similarity metrics from learned categories in the form of a prior. In all tasks, the computational framework comes close to capturing human abilities that other, more complex, machine vision systems struggle to reproduce.

References

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., & Shanahan, M. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Griffiths, T., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, *16*, 17.

Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., & Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems* (pp. 2946–2954).

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

Kohli, P., Sharp, T., Minka, T., Winn, J., Shotton, J., & Criminisi, A. (2005). *Microsoft research in cambridge image dataset.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2012). One-shot learning with a hierarchical nonparametric bayesian model. In *Icml unsupervised and transfer learning* (pp. 195–206).

Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2013). Learning with hierarchical-deep models. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1958–1971.

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

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

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

Yildirim, I. (2012). Bayesian inference: Metropolis-hastings sampling. *University of Rochester, NY*.

Figure 4: Ground Truth Tree of Gazoobian Objects as Generated from Human Similarity Judgments. Each of the three branches at the top of the tree denotes a supercategoy. The gray box in the lower left hand of the figure denotes a basic-level category.

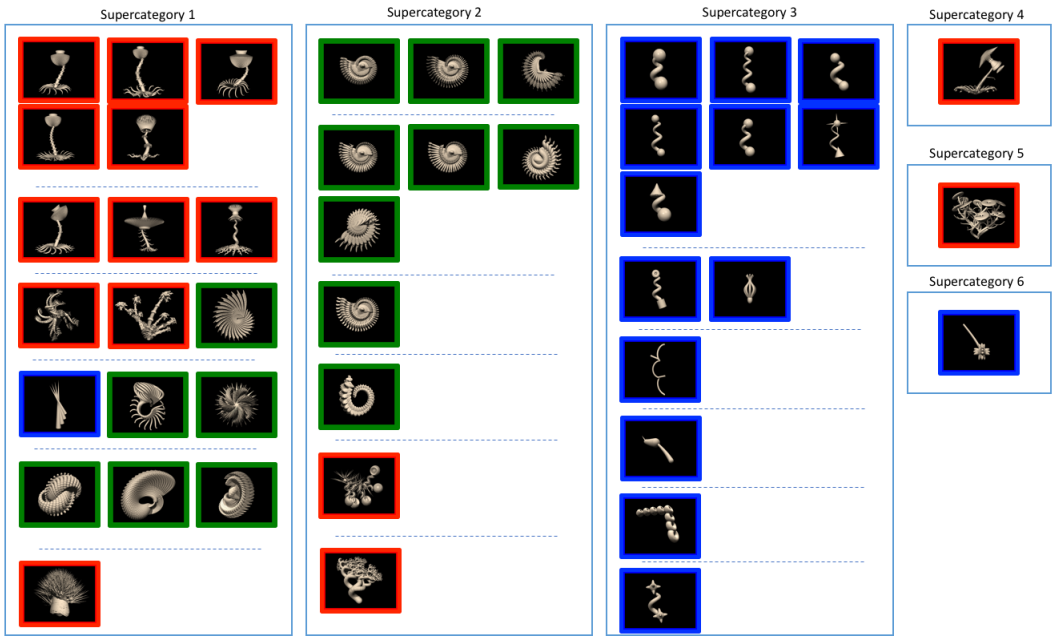

Figure 5: Model's Inferred Semantic Hierarchy of Gazoobian Objects. Outer boxes denote supercategories inferred by the model. Dashed lines separate model generated categories within each supercategory. Colored boxes around each object denote the ground truth supercategories as shown above.