JMLR: Workshop and Conference Proceedings 27:195–207, 2012 Workshop on Unsupervised and Transfer Learning

One-Shot Learning with a Hierarchical Nonparametric Bayesian Model

Ruslan Salakhutdinov RSALAKHU@MIT.EDU
Department of Statistics, University of Toronto
Toronto, Ontario, Canada

Josh Tenenbaum JBT@MIT.EDU
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA, USA

Antonio Torralba TORRALBA@MIT.EDU
CSAIL, MIT
Cambridge, MA, USA

Editor: I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver

Abstract

We develop a hierarchical Bayesian model that learns categories from single training examples. The model transfers acquired knowledge from previously learned categories to a novel category, in the form of a prior over category means and variances. The model discovers how to group categories into meaningful super-categories that express different priors for new classes. Given a single example of a novel category, we can efficiently infer which super-category the novel category belongs to, and thereby estimate not only the new category's mean but also an appropriate similarity metric based on parameters inherited from the super-category. On MNIST and MSR Cambridge image datasets the model learns useful representations of novel categories based on just a single training example, and performs significantly better than simpler hierarchical Bayesian approaches. It can also discover new categories in a completely unsupervised fashion, given just one or a few examples.

1. Introduction

In typical applications of machine classification algorithms, learning curves are measured in tens, hundreds or thousands of training examples. For human learners, however, the most interesting regime occurs when the training data are very sparse. Just a single example is often sufficient for people to grasp a new category and make meaningful generalizations to novel instances, if not to classify perfectly (Pinker, 1999). Human categorization often asymptotes after just three or four examples (Xu and Tenenbaum, 2007; Smith et al., 2002; Kemp et al., 2006; Perfors and Tenenbaum, 2009). To illustrate, consider learning entirely novel "alien" objects, as shown in Fig. 1, left panel. Given just three examples of a novel "tufa" concept (boxed in red), almost all human learners select just the objects boxed in gray (Schmidt, 2009). Clearly this requires very strong but also appropriately tuned inductive biases. A hierarchical Bayesian model we describe here takes a step towards this "one-shot learning" ability by learning abstract knowledge that support transfer of useful inductive biases from previously learned concepts to novel ones.

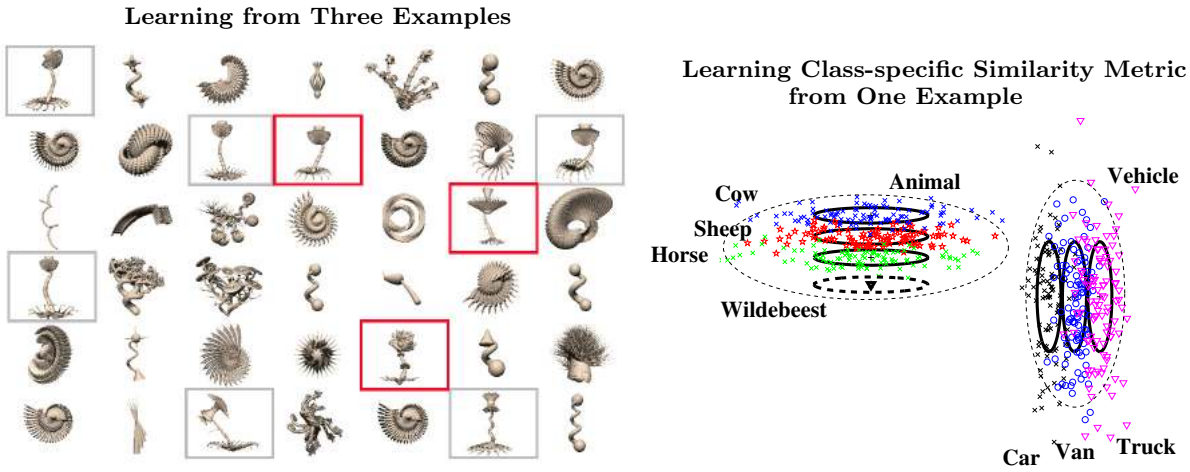


Figure 1: **Left:** Given only three examples (boxed in red) of a novel "tufa" object, which other objects are tufas? Most human learners select just the objects boxed in gray, as shown by Schmidt (2009). **Right:** Learning a similarity metric for a novel "wildebeest" class based on one example. The goal is to identify that the new "wildebeest" belongs to the "animal" super-category, which would allow to transfer an appropriate similarity metric and thereby generalize informatively from a single example.

At a minimum, categorizing an object requires information about the category's mean and variance along each dimension in an appropriate feature space. This is a similarity-based approach, where the mean represents the category prototype, and the inverse variances (or precisions) correspond to the dimensional weights in a category-specific similarity metric. One-shot learning may seem impossible because a single example provides information about the mean or prototype of the category, but not about the variances or the similarity metric. Giving equal weight to every dimension in a large a priori-defined feature space, or using the wrong similarity metric, is likely to be disastrous.

Our model leverages higher-order knowledge abstracted from previously learned categories to estimate the new category's prototype as well as an appropriate similarity metric from just one example. These estimates are also improved as more examples are observed. To illustrate, consider how human learners seeing one example of an unfamiliar animal, such as a wildebeest (or gnu), can draw on experience with many examples of "horse", "cows", "sheep", and more familiar related categories. These similar categories have similar prototypes – horses, cows, and sheep look more like each other than like furniture or vehicles – but they also have similar variability in their feature-space representations, or similar similarity metrics: The ways in which horses vary from the "horse" prototype are similar to the ways in which sheep vary from the "sheep" prototype. We may group these similar basic-level categories into an "animal" super-category, which captures these classes' similar prototypes as well as their similar modes of variation about their respective prototypes, as show in Fig. 1, right panel. If we can identify the new example of "wildebeest" as belonging to this "animal" super-category, we can transfer an appropriate similarity metric and thereby generalize informatively even from a single example.

Learning similarity metric over the high-dimensional input spaces has become an important task in machine learning as well. A number of recent approaches (Weinberger and Saul, 2009; Babenko et al., 2009; Singh-Miller and Collins, 2009; Goldberger et al., 2004;

Salakhutdinov and Hinton, 2007; Chopra et al., 2005) have demonstrated that learning a class-specific similarity metric can provide some insights into how high-dimensional data is organized and it can significantly improve the performance of algorithms like K-nearest neighbours that are based on computing distances. Most this work, however, focused on learning similarity metrics when many labeled examples are available, and did not attempt to address the one-shot learning problem.

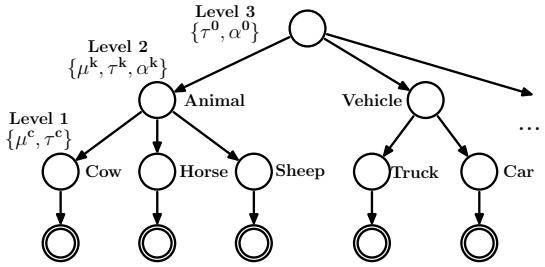
Although inspired by human learning, our approach is intended to be broadly useful for machine classification and AI tasks. To equip a robot with human-like object categorization abilities, we must be able to learn tens of thousands of different categories, building on (and not disrupting) representations of old ones (Bart and Ullman, 2005; Biederman, 1995). In these settings, learning from one or a few labeled examples and performing efficient inference will be crucial. Our method is designed to scale up in precisely these ways: a nonparametric prior allows new categories to be formed at any time in either supervised or unsupervised modes, and conjugate distributions allow most parameters to be integrated out analytically for very fast inference.

2. Related Prior Work

Hierarchical Bayesian models have previously been proposed (Kemp et al. (2006); Heller et al. (2009)) to describe how people learn to learn categories from one or a few examples, or learn similarity metrics, but these approaches were not focused on machine learning settings – large-scale problems with many categories and high-dimensional natural image data. A large class of models based on hierarchical Dirichlet processes (Teh et al. (2006)) have also been used for transfer learning (Sudderth et al. (2008); Canini and Griffiths (2009)). There are two key difference: First, HDPs typically assume a fixed hierarchy of classes for sharing parameters, while we learn the hierarchy in an unsupervised fashion. Second, HDPs are typically given many examples for each category rather than the one-shot learning cases we consider here. Recently introduced nested Dirichlet processes can also be used for transfer learning (Rodriguez and Vuppala (2009); Rodriguez et al. (2008)). However, this work assumes a fixed number of classes (or groups) and did not attempt to address one-shot learning problem. A recent hierarchical model of Adams et al. (2011) could also be used for transfer learning tasks. However, this model does not learn hierarchical priors over covariances, which is crucial for transferring an appropriate similarity metric to new basic-level categories in order to support learning from few examples. These recently introduced models are complementary to our approach, and can be combined productively, although we leave that as a subject for future work.

There are several related approaches in the computer vision community. A hierarchical topic model for image features (Bart et al. (2008); Sivic et al. (2008)) can discover visual taxonomies in an unsupervised fashion from large datasets but was not designed for one-shot learning of new categories. Perhaps closest to our work, Fei-Fei et al. (2006) also gave a hierarchical Bayesian model for visual categories with a prior on the parameters of new categories that was induced from other categories. However, they learned a single prior shared across all categories and the prior was learned only from three categories, chosen by hand.

More generally, our goal contrasts with and complements that of computer vision efforts on one-shot learning. We have attempted to minimize any tuning of our approach to



- For each super-category $k = 1, .., \infty$:
  draw θ^2 using Eq. 4.
- For each basic category $c^k = 1, .., \infty$, placed under each super-category k:
  draw θ^1 using Eq. 2.
- For each observation $n = 1, ..., N$
  draw $\mathbf{z}_n \sim \text{nCRP}(\gamma)$
  draw $\mathbf{x}^n \sim \mathcal{N}(\mathbf{x}^n | \theta^1, \mathbf{z}_n)$ using Eq. 1

Figure 2: **Left:** Hierarchical Bayesian model that assumes a fixed tree hierarchy for sharing parameters. **Right:** Generative process of the corresponding nonparametric model.

specifically visual applications. We seek a general-purpose hierarchical Bayesian model that depends minimally on domain-specific representations but instead learns to perform one-shot learning by finding more intelligent representations tuned to specific sub-domains of a task (our "super-categories").

3. Hierarchical Bayesian Model

Consider observing a set of N *i.i.d* input feature vectors $\{\mathbf{x}^1, ..., \mathbf{x}^N\}$, $\mathbf{x}^n \in R^D$. In general, features will be derived from high-dimensional, highly structured data, such as images of natural scenes, in which case the feature dimensionality D can be quite large (e.g. 50,000). For clarity of presentation, let us first assume that our model is presented with a fixed two-level category hierarchy. In particular, suppose that N objects are partitioned into C basic-level (or level-1) categories. We represent such partition by a vector $\mathbf{z}^b$ of length N, each entry of which is $z_n^b \in \{1, ..., C\}$. We also assume that our C basic-level categories are partitioned into K super-categories (level-2 categories), which we represent by $\mathbf{z}^s$ of length C, with $z_c^s \in \{1, ..., K\}$.

For any basic-level category c, the distribution over the observed feature vectors is assumed to be Gaussian with a category-specific mean μ^c and a category-specific *diagonal* precision matrix, whose entries are $\{\tau_d^c\}_{d=1}^D$. The distribution takes the following product form:

$$P(\mathbf{x}^n | z_n^b = c, \theta^1) = \prod_{d=1}^{D} \mathcal{N}(x_d^n | \mu_d^c, 1/\tau_d^c), \tag{1}$$

where $\mathcal{N}(x|\mu, 1/\tau)$ denotes a Gaussian distribution with mean μ and precision τ and $\theta^1 = \{\mu^c, \tau^c\}_{c=1}^C$ denotes the level-1 category parameters. We next place a conjugate Normal-Gamma prior over $\{\mu^c, \tau^c\}$. Let $k = z_c^s$, i.e. let the level-1 category c belong to level-2 category k, where $\theta^2 = \{\mu^k, \tau^k, \alpha^k\}_{k=1}^K$ denote the level-2 parameters. Then: $P(\mu^c, \tau^c | \theta^2, \mathbf{z}^s) = \prod_{d=1}^D P(\mu_d^c, \tau_d^c | \theta^2, \mathbf{z}^s)$, where for each dimension d we have:

$$P(\mu_d^c, \tau_d^c | \theta^2) = P(\mu_d^c | \tau_d^c, \theta^2) P(\tau_d^c | \theta^2) = \mathcal{N}(\mu_d^c | \mu_d^k, 1/(\nu\tau_d^c)) \Gamma(\tau_d^c | \alpha_d^k, \alpha_d^k/\tau_d^k). \tag{2}$$

Our parameterization of the Gamma density is in terms of its shape α^k and mean τ^k parameters:

$$\Gamma(\tau | \alpha^k, \alpha^k/\tau^k) = \frac{(\alpha^k/\tau^k)^{\alpha^k}}{\Gamma(\alpha^k)} \tau^{\alpha^k - 1} \exp\left(-\tau \frac{\alpha^k}{\tau^k}\right). \tag{3}$$

Such a parameterization is more interpretable, since $\mathrm{E}[\tau] = \tau^k$. In particular, from Eq. 2, we can easily derive that $\mathrm{E}[\mu^c] = \mu^k$ and $\mathrm{E}[\tau^c] = \tau^k$. This gives our model a very intuitive interpretation: the expected values of the basic level-1 parameters θ^1 are given by the corresponding level-2 parameters θ^2. The parameter α^k further controls the variability of τ^c around its mean, i.e. $\mathrm{Var}[\tau^c] = {(\tau^k)^2}/{\alpha^k}$. For the level-2 parameters θ^2, we shall assume the following conjugate priors:

$$P(\mu_d^k) = \mathcal{N}(\mu_d^k|0, 1/\tau^0), \quad P(\alpha_d^k|\alpha^0) = \mathrm{Exp}(\alpha_d^k|\alpha^0), \quad P(\tau_d^k|\theta^0) = \mathrm{IG}(\tau_d^k|a^0, b^0), \tag{4}$$

where $\mathrm{Exp}(x|\alpha)$ denotes an exponential distribution with rate parameter α, and $\mathrm{IG}(x|\alpha, \beta)$ denotes an inverse-gamma distribution with shape parameter α and scale parameter β. We further place a diffuse Gamma prior $\Gamma(1, 1)$ over the level-3 parameters $\theta^3 = \{\alpha^0, \tau^0\}$. Throughout our experimental results, we also set $a^0 = 1$ and $b^0 = 1$.

3.1. Modelling the number of super-categories

So far we have assumed that our model is presented with a two-level partition $\mathbf{z} = \{\mathbf{z}^s, \mathbf{z}^b\}$. If, however, we are not given any level-1 or level-2 category labels, we need to infer the distribution over the possible category structures. We place a nonparametric two-level nested Chinese Restaurant Prior (CRP) (Blei et al. (2003, 2010)) over $\mathbf{z}$, which defines a prior over tree structures and is flexible enough to learn arbitrary hierarchies. The main building block of the nested CRP is the Chinese restaurant process, a distribution on partition of integers. Imagine a process by which customers enter a restaurant with an unbounded number of tables, where the n^{th} customer occupies a table k drawn from:

$$P(z_n = k|z_1, ..., z_{n-1}) = \begin{cases} \frac{n^k}{n-1+\gamma} & n^k > 0 \\ \frac{\gamma}{n-1+\gamma} & k \text{ is new} \end{cases}, \tag{5}$$

where n^k is the number of previous customers at table k and γ is the concentration parameter.

The Nested CRP, nCRP(γ), extends CRP to nested sequence of partitions, one for each level of the tree. In this case each observation n is first assigned to the super-category z_n^s using Eq. 5. Its assignment to the basic-level category z_n^b, that is placed under a super-category z_n^s, is again recursively drawn from Eq. 5 (for details see Blei et al. (2010)). For our model, a two-level nested CRP allows flexibility of having a potentially unbounded number of super-categories as well as an unbounded number of basic-level categories placed under each super-category. Finally, we also place a Gamma prior $\Gamma(1, 1)$ over γ. The full generative model is given in Fig. 2, right panel. Unlike in many conventional hierarchical Bayesian models, here we infer both the model parameters as well as the hierarchy for sharing those parameters.

Our model can be readily used in unsupervised or semi-supervised modes, with varying amounts of label information. Here we focus on two settings. First, we assume basic-level category labels have been given for all examples in a training set, but no super-category labels are available. We must infer how to cluster basic categories into super-categories at the same time as we infer parameter values at all levels of the hierarchy. The training set includes many examples of familiar basic categories but only one (or few) example for a novel class. The challenge is to generalize the new class intelligently from this

one example by inferring which super-category the new class comes from and exploiting that super-category's implied priors to estimate the new class's prototype and similarity metric most accurately. This training regime reflects natural language acquisition, where spontaneous category labeling is frequent, almost all spontaneous labeling is at the basic level (Rosch et al., 1976) yet children's generalizations are sensitive to higher superordinate structure (Mandler, 2004), and where new basic-level categories are typically learned with high accuracy from just one or a few labeled examples. Second, we consider a similar labeled training set but now the test set consists of many unlabeled examples from an unknown number of basic-level classes – including both familiar and novel classes. This reflects the problem of "unsupervised category learning" a child or robot faces in discovering when they have encountered novel categories, and how to break up new instances into categories in an intelligent way that exploits knowledge abstracted from a hierarchy of more familiar categories.

4. Inference

Inferences about model parameters at all levels of hierarchy can be performed by MCMC. When the tree structure $\mathbf{z}$ of the model is not given, the inference process will alternate between fixing $\mathbf{z}$ while sampling the space of model parameters θ and fixing θ while sampling category assignments.

Sampling level-1 and level-2 parameters: Given level-2 parameters θ^2 and $\mathbf{z}$, the conditional distribution $P(\mu^c, \tau^c|\theta^2, \mathbf{z}, \mathbf{x})$ is Normal-Gamma (Eq. 2), which allows us to easily sample level-1 parameters $\{\mu^c, \tau^c\}$. Given $\mathbf{z}$, θ^1, and θ^3, the conditional distributions over the mean μ^k and precision τ^k take Gaussian and Inverse-Gamma forms. The only complicated step involves sampling α^k that control the variation of the precision term τ^c around its mean (Eq. 3). The conditional distribution over α^k cannot be computed in closed form and is proportional to:

$$p(\alpha^k|\mathbf{z}, \theta^1, \theta^3, \tau^k) \propto \frac{(\alpha^k/\tau^k)^{\alpha^k n_k}}{\Gamma(\alpha^k)^{n_k}} \exp\left(-\alpha^k\left(\alpha^0 + S^k/\tau^k - T^k\right)\right), \tag{6}$$

where $S^k = \sum_{c:z(c)=k} \tau^c$ and $T^k = \sum_{c:z(c)=k} \log(\tau^c)$. For large values of α^k the density, specified by Eq. 6, is similar to a Gamma density (Wiper et al. (2001)). We therefore use Metropolis-Hastings with a proposal distribution given by the Gamma density. In particular, we generate a new candidate

$$\alpha^* \sim Q(\alpha^*|\alpha^k) \quad \text{with} \quad Q(\alpha^*|\alpha^k) = \Gamma(\alpha^*|t, t/\alpha^k)$$

and accept it with M-H rule. In all of our experiments we use $t = 3$, which gave an acceptance probability of about 0.6. Sampling level-3 parameters is similar to sampling level-2 parameters.

Sampling assignments z: Given model parameters $\theta = \{\theta^1, \theta^2\}$, combining the likelihood term with the nCRP(γ) prior, the posterior over the assignment $\mathbf{z}_n$ can be calculated as follows:

$$p(\mathbf{z}_n|\theta, \mathbf{z}_{-n}, \mathbf{x}^n) \propto p(\mathbf{x}^n|\theta, \mathbf{z}_n) p(\mathbf{z}_n|\mathbf{z}_{-n}), \tag{7}$$

where $\mathbf{z}_{-n}$ denotes variables $\mathbf{z}$ for all observations other than n. We can further exploit the conjugacy in our hierarchical model when computing the probability of creating a new basic-level category. Using the fact that Normal-Gamma prior $p(\mu^c, \tau^c)$ is the conjugate prior of a normal distribution, we can easily compute the following marginal likelihood:

$$p(\mathbf{x}^n|\theta^2, \mathbf{z}_n) = \int_{\mu^c,\tau^c} p(\mathbf{x}^n, \mu^c, \tau^c|\theta^2, \mathbf{z}_n) = \int_{\mu^c,\tau^c} p(\mathbf{x}^n|\mu^c, \tau^c)p(\mu^c, \tau^c|\theta^2, \mathbf{z}_n).$$

Integrating out basic-level parameters θ^1 lets us more efficiently sample over the tree structures[1]. When computing the probability of placing $\mathbf{x}^n$ under a newly created super-category, its parameters are sampled from the prior.

5. One-shot Learning

One of the key goals of our work is to develop a model that has the ability to generalize from a single example. Consider observing a single new instance $\mathbf{x}^*$ of a *novel category* c^* [2]. Conditioned on the current setting of the level-2 parameters θ^2 and our current tree structure $\mathbf{z}$, we can first infer which super-category the novel category should belong to, i.e. we can compute the posterior distribution over the assignments $\mathbf{z}_c^*$ using Eq. 7. We note that our new category can either be placed under one of the existing super-categories, or create its own super-category, if it is sufficiently different from all of the existing super-categories.

Given an inferred assignment $\mathbf{z}_c^*$ and using Eq. 2, we can infer the posterior mean and precision terms (or similarity metric) $\{\mu^*, \tau^*\}$ for our novel category. We can now test the ability of the HB model to generalize to new instances of a novel category by computing the conditional probability that a new test input $\mathbf{x}^t$ belongs to a novel category c^*:

$$p(c^*|\mathbf{x}^t) = \frac{p(\mathbf{x}^t|\mathbf{z}_c^*)p(\mathbf{z}_c^*)}{\sum_{\mathbf{z}} p(\mathbf{x}^t|\mathbf{z})p(\mathbf{z})}, \tag{8}$$

where the prior is given by the nCRP(γ) and the log-likelihood takes form:

$$\log p(\mathbf{x}^t|c^*) = \frac{1}{2}\sum_d \log(\tau_d^*) - \frac{1}{2}\sum_d \tau_d^*(x_d^t - \mu_d^*)^2 + C,$$

where C is a constant that does not depend on the parameters. Observe that the relative importance of each feature in determining the similarity is proportional to the category-specific precision of that feature. Features that are salient, or have higher precision, within the corresponding category contribute more to the overall similarity of an input.

6. Experimental results

We now present experimental results on the MNIST handwritten digit and MSR Cambridge object recognition image datasets. During the inference step, we run our hierarchical Bayesian (HB) model for 200 full Gibbs sweeps, which was sufficient to reach convergence

1. In the supervised case, inference in simplified by only considering which super-category each basic-level category is assigned to.
2. Observing several examples of a new category is treated similarly.

Table 1: Performance results using the area under the ROC curve (AUROC) on the MNIST dataset. The Average panel shows results averaged over all 10 categories, using leave-one-out test format.

Model	Category: Digit 9				Category: Digit 6				Average			
	1 ex	2 ex	4 ex	20 ex	1 ex	2 ex	4 ex	20 ex	1 ex	2 ex	4 ex	20 ex
HB	0.81	0.85	0.88	0.90	0.85	0.89	0.92	0.97	0.85	0.88	0.90	0.93
HB-Flat	0.71	0.77	0.84	0.90	0.73	0.79	0.88	0.97	0.74	0.79	0.86	0.93
HB-Var	0.72	0.81	0.86	0.90	0.72	0.83	0.90	0.97	0.75	0.82	0.89	0.93
Euclidean	0.70	0.73	0.76	0.80	0.74	0.77	0.82	0.86	0.72	0.76	0.80	0.83
Oracle	0.87	0.89	0.90	0.90	0.95	0.96	0.96	0.97	0.90	0.92	0.92	0.93
MLE	0.69	0.75	0.83	0.90	0.72	0.78	0.87	0.97	0.71	0.77	0.84	0.93

and obtain good performance. We normalize input vectors to zero mean and scale the entire input by a single number to make the average feature variance be one.

In all of our experiments, we compare performance of the HB model to the following four alternative methods for one-shot learning. The first model, "Euclidean", uses a Euclidean metric, i.e. all precision terms are set to one and are never updated. The second model, that we call "HB-Flat", always uses a single super-category. When presented with a single example of a new category, HB-Flat will inherit a similarity metric that is shared by all existing categories, as done in Fei-Fei et al. (2006). Our third model, called "HB-Var", is similar in spirit to the approach of Heller et al. (2009) and is based on clustering only covariance matrices without taking into account the means of the super-categories. Our last model, "MLE", ignores hierarchical Bayes altogether and estimates a category-specific mean and precision from sample averages. If a category contains only one example, the model uses the Euclidean metric. Finally, we also compare to the "Oracle" model that is the same as our HB model, but always uses the correct, instead of inferred, similarity metric.

6.1. MNIST dataset

The MNIST dataset contains 60,000 training and 10,000 test images of ten handwritten digits (zero to nine), with 28×28 pixels. For our experiments, we randomly choose 1000 training and 1000 test images (100 images per class). We work directly in the pixel space because all handwritten digits were already properly aligned. In addition, working in the pixel space allows us to better visualize the kind of transfer of similarity metrics our model is performing. Fig. 3, left panel, shows a typical partition over the basic level categories, along with corresponding mean and similarity metrics, that our model discovers.

We first study the ability of the HB model to generalize from a single training example of handwritten digit "nine". To this end, we trained the HB model on 900 images (100 images of each of zero-to-eight categories), while withholding all images that belong to category "nine". Given a single new instance of a novel "nine" category our model is able to discover that the new category is more like categories that contain images of seven and four, and hence this novel category can inherit the mean and the similarity metric, shared by categories "seven" and "four".

Table 1 further quantifies performance using the area under the ROC curve (AUROC) for classifying 1000 test images as belonging to the "nine" vs. all other categories. (an area of 0.5 corresponds to the classifier that makes random predictions). The HB model

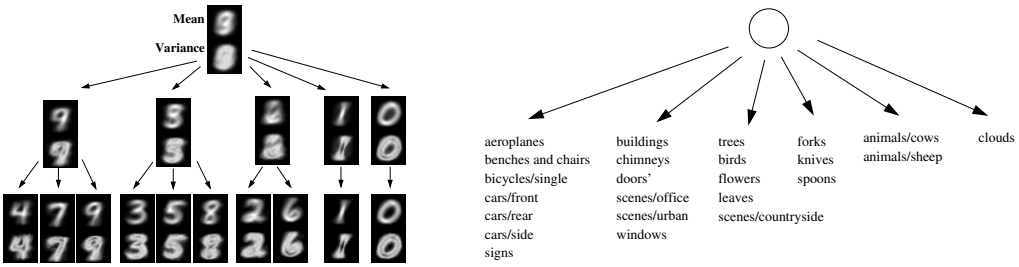


Figure 3: **Left:** MNIST Dataset: A typical partition over the 10 categories discovered by the HB model. Top panels display means and bottom panels display variances (white encodes larger values). **Right:** MSR Dataset: A typical partition over the 24 categories discovered by the HB model.

Table 2: Performance results using the area under the ROC curve (AUROC) on the MSR dataset. The Average panel shows results averaged over all 24 categories, using leave-one-out test format.

Model	Category: Cow				Category: Flower				Average			
	1 ex	2 ex	4 ex	20 ex	1 ex	2 ex	4 ex	20 ex	1 ex	2 ex	4 ex	20 ex
HB	0.77	0.81	0.84	0.89	0.71	0.75	0.78	0.81	0.76	0.80	0.84	0.87
HB-Flat	0.62	0.69	0.80	0.89	0.59	0.64	0.75	0.81	0.65	0.71	0.78	0.87
HB-Var	0.61	0.73	0.83	0.89	0.60	0.68	0.77	0.81	0.64	0.74	0.81	0.87
Euclidean	0.59	0.61	0.63	0.66	0.55	0.59	0.61	0.64	0.63	0.66	0.69	0.71
Oracle	0.83	0.84	0.87	0.89	0.77	0.79	0.80	0.81	0.82	0.84	0.86	0.87
MLE	0.58	0.64	0.78	0.89	0.55	0.62	0.72	0.81	0.62	0.67	0.77	0.87

achieves an AUROC of 0.81, considerably outperforming HB-Flat, HB-Var, Euclidean, and MLE that achieve an AUROC of 0.71, 0.72, 0.70, and 0.69 respectively. Moreover, with just four examples, the HB model is able to achieve performance close to that of the Oracle model. This is in sharp contrast to HB-Flat, MLE and Euclidean models, that even with four examples perform far worse.

6.2. MSR Cambridge Dataset

We now present results on a considerably more difficult MSR Cambridge dataset[3], that contains images of 24 different categories. Fig. 3, right panel, shows 24 basic-level categories along with a typical partition that our model discovers. We use a simple "texture-of-textures" framework for constructing image features (DeBonet and Viola (1997)).

We first tested the ability of our model to generalize from a single image of a cow. Similar to the experiments on the MNIST dataset, we first train the HB model on images corresponding to 23 categories, while withholding all images of cows. In general, our model is able to discover that the new "cow" category is more like the "sheep" category, as opposed to categories that contain images of cars, or forks, or buildings. This allows the new "cow" category inherit sheep's similarity metric.

Table 2 show that the HB model, based on a single example of cow, achieves an AUROC of 0.77. This is compared to an AUROC of only 0.62, 0.61, 0.59, and 0.58 achieved by the HB-Flat, HB-Var, Euclidean, and MLE models. Similar to the results on the MNIST dataset, the HB model with just one example performs comparably the HB-Flat and MLE

3. Available at http://research.microsoft.com/en-us/projects/objectclassrecognition/

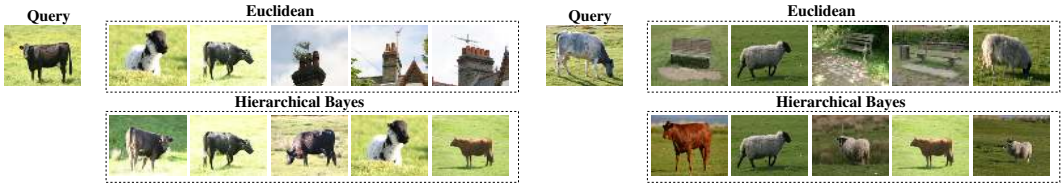


Figure 4: Retrieval results based on observing a single example of cow. Top five most similar images were retrieved from the test set, containing 360 images corresponding to 24 categories.

Figure 5: Unsupervised category discovery. **Left:** Six representative test images, sorted by the posterior probability of forming a novel category. **Right:** When presented with 18 unlabeled test images, the model correctly places nine "familiar" images in nine different basic-level categories, while also correctly forming three novel basic-level categories with three examples each.

models that make use of four examples. Fig. 4 further displays retrieval results based on a single image of a cow. As expected, the HB model performs much better compared to the Euclidean model that does not learn a similarity metric.

6.3. Unsupervised Category Discovery

Another key advantage of the hierarchical nonparametric Bayesian model is its ability to infer category structure in an unsupervised fashion, discovering novel categories at both levels 1 and 2 of the hierarchy. We explored the HB model's category discovery ability by training on labeled examples of 21 basic-level MSR categories, leaving out clouds, trees, and chimneys. We then provided six test images: one in each of the three unseen categories and one in each of three familiar basic-level categories (car, airplane, bench). For each test image, using Eq. 8, we can easily compute the posterior probability of forming a new basic-level category. Figure 5, left panel, shows six representative test images, sorted by the posterior probability of forming a novel category. The model correctly identifies the car, the airplane and the bench as belonging to familiar categories, and places much higher probability on forming novel categories for the other images. With only one unlabeled example of these novel classes, the model still prefers two of them in familiar categories: the "tree" is interpreted as an atypical example of "countryside" while the "chimney" is classified as an atypical "building".

The model, however, can correctly discover novel categories given only a little more unlabeled data. With 18 unlabeled test images (see Fig. 5), after running a Gibbs sampler for 100 steps, the model correctly places nine "familiar" images in nine different basic-level categories, while also correctly forming three novel basic-level categories with three examples each. Most interestingly, these new basic-level categories are placed at the appropriate level

of the category hierarchy: the novel "tree" category is correctly placed under the super-category containing "leaves" and "countrysides"; the novel "chimney" category is placed together with "buildings" and "doors"; while "clouds" category is placed in its own super-category – all consistent with the hierarchy we originally found from a fully labeled training set (see Fig. 3). Other models we tried for this unsupervised task perform much worse; they confuse "chimneys" with "cows" and "trees" with "countrysides".

7. Conclusions

In this paper we developed a hierarchical nonparametric Bayesian model for learning a novel category based on a single training example. Our experimental results further demonstrate that our model is able to effectively transfer appropriate similarity metric from the previously learned categories to a novel category based on observing a single example. There are several key advantages to our model. First, due to efficient Gibbs moves that can exploit conjugacy, the model can be efficiently trained. Many of the Gibbs updates can be run in parallel, which will allow our model to potentially handle a large number of basic-level categories. Second, the model is able to discover meaningful super-categories and be able to form coherent novel categories. Finally, given a single example of a novel category, the model is able to quickly infer which super-category the new basic-level ategory should belong to. This in turns allows us to efficiently infer the appropriate similarity metric for this novel category.

References

R. Adams, Z. Ghahramani, and M. Jordan. Tree-structured stick breaking processes for hierarchical data. In *To appear in NIPS*. MIT Press, 2011.

B. Babenko, S. Branson, and S. J. Belongie. Similarity functions for categorization: from monolithic to category specific. In *ICCV*, 2009.

E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, pages 672–679, 2005.

E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, pages 1–8, 2008.

I. Biederman. Visual object recognition. *An Invitation to Cognitive Science*, 2:152–165, 1995.

D. M. Blei, T. L. Griffiths, M. I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS*. MIT Press, 2003.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), 2010.

K. R. Canini and T. L. Griffiths. Modeling human transfer learning with the hierarchical Dirichlet process. In *NIPS 2009 workshop: Nonparametric Bayes*, 2009.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages I: 539–546, 2005.

J. S. DeBonet and P. A. Viola. Structure driven image database retrieval. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *NIPS*. The MIT Press, 1997.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(4):594–611, April 2006.

J. Goldberger, S. T. Roweis, G. E. Hinton, and R. R. Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, 2004.

K. Heller, A. Sanborn, and N. Chater. Hierarchical learning of dimensional biases in human categorization. In *NIPS*, 2009.

C. Kemp, A. Perfors, and J. Tenenbaum. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321, 2006.

J.M. Mandler. *The foundations of mind: Origins of conceptual thought*. Oxford University Press, USA, 2004.

A. Perfors and J.B. Tenenbaum. Learning to learn categories. In *31st Annual Conference of the Cognitive Science Society*, pages 136–141, 2009.

S. Pinker. *How the Mind Works*. W.W. Norton, 1999.

A. Rodriguez and R. Vuppala. Probabilistic classification using Bayesian nonparametric mixture models. Technical Report, 2009.

A. Rodriguez, D. Dunson, and A. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103:1131–1144, 2008.

E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories* 1. *Cognitive psychology*, 8(3):382–439, 1976.

R. R. Salakhutdinov and G. E. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11, 2007.

L. Schmidt. Meaning and compositionality as statistical induction of categories and constraints. Ph.D Thesis, Massachusetts Institute of Technology, 2009.

N. Singh-Miller and M. Collins. Learning label embeddings for nearest-neighbor multi-class classification with an application to speech recognition. In *Advances in Neural Information Processing Systems*. MIT Press, 2009.

J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, pages 1–8, 2008.

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

E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77(1-3):291–330, 2008.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 3, 2009.

M. Wiper, D. R. Insua, and F. Ruggeri. Mixtures of Gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10(3), September 2001.

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