One-Shot Learning with a Hierarchical Nonparametric Bayesian Model

R. Salakhutdinov, J. Tenenbaum and A. Torralba
*MIT Technical Report, 2010*

Presented by Esther Salazar
Duke University

June 10, 2011
Summary

**Contribution:** The authors develop a hierarchical nonparametric Bayesian model for learning a novel category based on a single training example.

The mode discovers how to group categories into meaningful super-categories that express different priors for new classes.

Given a single example of a novel category, the model is able to quickly infer which super-category the new basic-level category should belong to.
Motivation

- Categorizing an object requires information about the category’s mean and variance in an appropriate feature space.

- Similarity-based approach:
  - the mean represents the category prototype
  - the inverse variances (precisions) correspond to the dimensional weights in a category-specific similarity metric

- In this context, a one-shot learning is not possible because a single example can provide information about the mean but not about the variances (or similarity metric).

Proposal: The proposed model leverages higher-order knowledge abstracted from previously learned categories to estimate the new category’s prototype (mean) as well as an appropriate similarity metric (variance) from just one single example or more.
Hierarchical Bayesian model

- Consider observing a set of $N$ input features
  \[ \{x^1, \ldots, x^n\}, \quad x^n \in \mathbb{R}^D \]

- Features derived from high-dimensional data such images, (e.g. $D = 50,000$).

- Assumption: fixed two-level category hierarchy.

- Suppose that $N$ objects are partitioned into $C$ basic-level (level-1) categories. That partition is represented by a vector $z^b = (z^b_1, \ldots, z^b_N)$ such that $z^b_n \in \{1, \ldots, C\}$

- The $C$ basic-level categories are partitioned into $K$ super-categories (level-2) represented by $z^s = (z^s_1, \ldots, z^s_C)$ such that $z^s_c \in \{1, \ldots, K\}$
For any basic-level category $c$, the distribution over the observed feature vectors is

$$P(\mathbf{x}^n|z^b_n = c, \theta^1) = \prod_{d=1}^{D} \mathcal{N}(x^n_{d}|\mu^c_{d}, 1/\tau^c_{d})$$

(1)

where $\theta^1 = \{\mu^c, \tau^c\}_{c=1}^{C}$ denotes the level-1 category parameters.

Let $k = z^s_c$, the level-1 category $c$ belong to level-2 category $k$ and $\theta^2 = \{\mu^k, \tau^k, \alpha^k\}_{k=1}^{K}$ the level-2 parameters.

$$P(\mu^c, \tau^c|\theta^2, z^s) = \prod_{d=1}^{D} P(\mu^c_{d}, \tau^c_{d}|\theta^2, z^s)$$

$$P(\mu^c_{d}, \tau^c_{d}|\theta^2) = P(\mu^c_{d}|\tau^c_{d}, \theta^2)P(\tau^c_{d}|\theta^2) = \mathcal{N}(\mu^c_{d}|\mu^k_{d}, 1/(\nu\tau^k_{d}))\Gamma(\tau^c_{d}|\alpha^k_{d}, \alpha^k_{d}/\tau^k_{d})$$

(2)
\[ P(\mu^c_d, \tau^c_d | \theta^2) = P(\mu^c_d | \tau^c_d, \theta^2)P(\tau^c_d | \theta^2) \]
\[ = \mathcal{N}(\mu^c_d | \mu^k_d, 1/(\nu \tau^c_d)) \Gamma(\tau^c_d | \alpha^k_d, \alpha^k_d / \tau^k_d) \]

- It can be easily derived that

\[ E[\mu^c] = \mu^k, \quad E[\tau^c] = \tau^k \]

- That means: the expected values of the basic level-1 parameters \( \theta^1 \) are given by the corresponding level-2 parameters \( \theta^2 \)

- The parameters \( \alpha^k \) controls the variability of \( \tau^c \) around its mean

- For the level-2 parameters \( \theta^2 \) they assume:

\[ P(\mu^k_d) = \mathcal{N}(\mu^k_d | 0, 1/\tau^0), \quad P(\alpha^k_d | \alpha^0) = \text{Exp}(\alpha^k_d | \alpha^0), \quad P(\tau^k_d | \theta^0) = \text{IG}(\tau^k_d | a^0, b^0), \]
Modeling the number of super-categories

- The model is presented with a two-level partition \( z = \{ z^s, z^b \} \) that assumes a fixed hierarchy for sharing parameters.

- **Aim:** To infer the distribution over the possible categories \( z \).

- **Proposal:** The authors place a nonparametric two-level nested Chinese Restaurant Prior (nCRP) (Blei et al. 2003, 2010) over \( z \).

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

The nCRP(\( \gamma \)) extends CRP to nested sequence of partitions, one for each level of the tree (prior: \( \gamma \sim \Gamma(1, 1) \)).

- In this case, each observation \( n \) is first assigned to the super category \( z^s_n \) using Eq. 4.

- Its assignment to the basic-level category \( z^b_n \) (under a super category \( z^s_n \)) is again recursively drawn from Eq. 4.

- Full generative model in Fig. 1 right panel.
Inference

- **Sampling level-1 parameters**: Given $\theta^2$ and $z$ the conditional distribution $P(\mu^c, \tau^c|\theta^2, z, x)$ is Normal-Gamma also

  $$P \left( \{\mu^c, \tau^c\}_{c=1}^{C} | \theta^2, z \right) = \prod_{c=1}^{C} \prod_{d=1}^{D} P(\mu^c_d, \tau^c_d | \theta^2, z).$$

- **Sampling level-2 parameters**: Given $z$ and $\theta^1$ the conditional distributions over the mean $\mu^k$ and precision $\tau^k$ take Gaussian and Inverse-Gamma forms. Also, for $\alpha^k$

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

  $$p(z_n | \theta, z_{-n}, x^n) \propto p(x^n | \theta, z_n)p(z_n | z_{-n}), \quad (5)$$
One-shot learning

- Consider observing a single new instance $x^*$ of a novel category $c^*$
- Conditioned on the level-2 parameters $\theta^2$ and the tree structure $z$.
  - Compute the posterior distribution over the assignments $z_c^*$ using Eq. 5 (which super-category the novel category belong to)
  - The new category can either be placed under one of the existing super-categories or create its own super-category
- Given an inferred assignment $z_c^*$ and using Eq. 2, we can infer the posterior mean and precisions terms $\{\mu^*, \tau^*\}$ for the novel category
- Also, the conditional probability that a new test input $x^t$ belongs to a novel category $c^*$ is:
  $$p(c^*|x^t) = \frac{p(x^t|z_c^*)p(z_c^*)}{\sum_z p(x^t|z)p(z)},$$
The authors present experimental results on the **MNIST handwritten digit** and **MSR Cambridge object recognition image datasets**.

The proposed model was compared with four base-line models

- **Euclidean**: use Euclidean metric, i.e. all precision terms are set to one and are never updated
- **HB-Flat**: always uses a single super-category
- **HB-Var**: is based on clustering only covariance matrices without taking into account the means of the super-categories
- **MLE**: ignores hierarchical Bayes altogether and estimates a category-specific mean and precision from sample averages
- **Oracle**: model that always uses the correct, instead of inferred, similarity metric
Table 1: Performance results using the area under the ROC curve (AUROC) on the MNIST dataset. The rightmost Average panel shows results averaged over all 10 categories, using leave-one-out test format.

<table>
<thead>
<tr>
<th>Model</th>
<th>Category: Digit 9</th>
<th>Category: Digit 6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 ex</td>
<td>2 ex</td>
<td>4 ex</td>
</tr>
<tr>
<td>HB</td>
<td>0.81</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>HB-Flat</td>
<td>0.71</td>
<td>0.77</td>
<td>0.84</td>
</tr>
<tr>
<td>HB-Var</td>
<td>0.72</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>Euclidean</td>
<td>0.70</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>Oracle</td>
<td>0.87</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>MLE</td>
<td>0.69</td>
<td>0.75</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Figure 2: MNIST dataset. Left: A typical partition over the 10 basic-level categories discovered by the HB model. Top panels display means and bottom panels display variances (white encodes larger values). Middle: Transfer of similarity metric based on a single example of a novel ‘nine’ category. Right: Retrieval results: Top eight most similar images retrieved from the test set of 1000 images corresponding to 10 categories. Note that due to metric transfer, the HB model is able to avoid mistakes made by the Euclidean model.
The Cambridge Dataset contains images of 24 different categories. The figure shows 24 basic-level categories along with a typical partition that the model discovers, where many super-categories contain semantically similar basic-level categories.
Figure 5: 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 6: Unsupervised category discovery. **Left:** Six representative test images, sorted by the posterior probability of forming a novel category. **Right:** The model is presented with 18 unlabeled test images. 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.