Modeling Transfer Learning in Human Categorization with the Hierarchical Dirichlet Process

Kevin R. Canini KEVIN@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720 USA

Mikhail M. Shashkov MSHASHKOV@BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720 USA

Thomas L. Griffiths TOM_GRIFFITHS@BERKELEY.EDU
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Transfer learning can be described as the distillation of abstract knowledge from one learning domain or task and the reuse of that knowledge in a related domain or task. In categorization settings, transfer learning is the modification by past experience of prior expectations about what types of categories are likely to exist in the world. While transfer learning is an important and active research topic in machine learning, there have been few studies of transfer learning in human categorization. We propose an explanation for transfer learning effects in human categorization, implementing a model from the statistical machine learning literature – the hierarchical Dirichlet process (HDP) – to make empirical evaluations of its ability to explain these effects. We present two laboratory experiments which measure the degree to which people engage in transfer learning in a controlled setting, and we compare our model to their performance. We find that the HDP provides a good explanation for transfer learning exhibited by human learners.

1. Introduction

In most learning environments, people do not start from a *tabula rasa* when forming hypotheses and drawing inferences. Rather, knowledge gained from one domain is abstracted and reused in other domains. This behavior, called *transfer learning*, has been demonstrated in children as young as 3 years (Brown & Kane, 1988) and is central to people's ability to quickly adapt to new situations given a limited amount of data. It is also a major area of machine learning research, since automated systems that are capable of learning rapidly in a variety of domains will need to be able to transfer the knowledge that they acquire in one domain to another (Thrun & Pratt, 1998). Despite the importance of transfer learning as part of an explanation for how people learn new concepts, most studies of human category learning have focused on settings where people learn independent categories, with little opportunity for the learning of one category to influence the learning of another. In this paper, we explore how people who learn systems of inter-related categories exhibit transfer effects, and how these effects can be modeled using tools from hierarchical Bayesian statistics.

Transfer learning in human categorization can occur in a number of ways. One example is dimensional attention: previously learned categories can reveal which dimensions are more diagnostic of category membership. This type of transfer can be modeled by explicitly representing a saliency factor for each dimension, and is a standard part of existing categorization models, e.g., (Nosofsky, 1986; Kruschke, 1992; Perfors & Tenenbaum, 2009). However, recent work on rational models of human category learning has taken an approach that can support a novel kind of transfer learning. In these models – the rational model of categorization (Anderson, 1991), the mixture model of categorization (Rosseel, 2002), and the hierarchical Dirichlet process model (Griffiths et al., 2007a) – categories are represented as a set of clusters of objects. Learning a system of categories involves learning which objects cluster together and which clusters

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

belong to each category. Importantly, these models can allow for the possibility that categories share clusters of objects, providing a way for transfer learning to take place: the clusters induced by learning one category can be used to inform learning about other categories. For example, if learning the category of *cats* leads to the formation of clusters corresponding to different kinds of cats – such as Siamese, Persians, and tabbies – one might more quickly learn the category of *striped objects*, having the expectation that if one of the members of a cluster (such as a tabby cat) belongs to the category, the other members will too.

Using category labels to identify meaningful clusters of objects is a particularly interesting form of transfer learning, as it means that learners are dynamically developing a vocabulary with which to describe novel concepts. Whether people perform this kind of transfer learning has not previously been explored. Consequently, we present an empirical analysis of how simultaneously learning clusters and categories can affect categorization performance. We describe two laboratory experiments in which people are taught systems of related categories and we measured the degree to which their training affects future decisions.

The plan of the paper is as follows. Section 2 introduces the idea of representing categories with learned clusters of objects, laying out the probabilistic model – the hierarchical Dirichlet process – we use to describe this form of learning. Section 3 outlines the motivation for our human learning experiments, and Sections 4 and 5 present experiments examining whether these effects are observed with discrete and continuous stimuli, respectively. Section 6 discusses the implications of our results and concludes the paper.

2. Representing categories with learned clusters of objects

Rational models of cognition explain human behavior as an attempt to find optimal solutions to the computational challenges posed by the environment (Anderson, 1990). Rational models of categorization represent categories as probability distributions over stimuli and decide what category a new stimulus belongs to by performing Bayesian inference (Anderson, 1991; Griffiths et al., 2007b). More formally, if each category c is associated with a probability distribution $p(x|c)$ over stimuli x, Bayes' rule indicates the posterior probability that a stimulus x belongs to category c is

$$p(c|x) = \frac{p(x|c)p(c)}{\sum_{c'} p(x|c')p(c')}, \tag{1}$$

where the sum in the denominator ranges over all the categories, and $p(c)$ reflects the prior probability that a stimulus belongs to category c.

This probabilistic perspective provides a way to unify existing psychological models of categorization (Ashby & Alfonso-Reese, 1995; Griffiths et al., 2007b). Much empirical work in psychology has focused on differentiating between prototype models, e.g., (Reed, 1972), in which a category is represented by a prototypical member, and exemplar models, e.g., (Medin & Schaffer, 1978; Nosofsky, 1986), in which a category is represented by all of the observed members. These two models can be shown to correspond to different statistical strategies for estimating $p(x|c)$, with prototype models corresponding to parametric density estimation and exemplar models corresponding to nonparametric density estimation (Ashby & Alfonso-Reese, 1995). Recently, several psychological models have been proposed that compromise between prototype and exemplar models, with a category being represented by a set of clusters, each of which has its own prototype, e.g., (Love et al., 2004; Vanpaemel et al., 2005). These models also have probabilistic analogues, in which $p(x|c)$ is a mixture distribution, with

$$p(x|c) = \sum_{z} p(x|z)p(z|c), \tag{2}$$

where $p(x|z)$ is a distribution over stimuli associated with mixture component z, and $p(z|c)$ is the weight assigned to that component by category c (Anderson, 1991; Rosseel, 2002; Griffiths et al., 2007a). Intuitively, each mixture component corresponds to a cluster of stimuli, and a category is a combination of the clusters.

Categorization models based on mixture distributions support a novel kind of transfer learning, where clusters are shared between categories. We focus on one such model here: the hierarchical Dirichlet process (HDP; Teh et al., 2006). Focusing on this model does not limit the scope of our conclusions, as the other models we have mentioned can be shown to be special cases of the HDP (Griffiths et al., 2007a). The model samples $p(x|c)$ from a Dirichlet process for each category c, so categories are represented by an unbounded number of components. Sharing between categories is implemented by generating the categories' common base distribution from a global Dirichlet process, allowing components to occur in multiple categories.

An intuitive understanding of the way categories are represented in the HDP can be obtained by considering the process by which the model assumes stimuli are generated. When a new stimulus is generated from category c, the model selects a cluster z from which to generate that stimulus. The cluster z might already exist in c, or it could be new to category c. The probability of each cluster is given by a stochastic process known as the Chinese restaurant process (Aldous, 1985; Pitman & Picard, 2006), with

$$p(z = k|\mathbf{z}_c) = \begin{cases} \frac{n_k}{n_c+\alpha} & k \text{ is an existing cluster} \\ \frac{\alpha}{n_c+\alpha} & k \text{ is a new cluster} \end{cases} \tag{3}$$

where $\mathbf{z}_c$ are the cluster assignments of all other stimuli from category c, n_k is the number of stimuli from cluster k in the category, n_c is the total number of stimuli in the category, and α is the *concentration parameter* of the model. If the stimulus is to be generated from a cluster that is new to category c, the cluster is generated from another Chinese restaurant process (with concentration parameter γ), this time selecting clusters based on their frequency across all categories. If this second process elects to generate a completely new cluster, parameters θ for that cluster are sampled from a global base distribution, and then the stimulus x is generated from a parametric distribution $p(x|\theta)$.

The generative process assumed by the HDP provides a way to generate categories that share clusters, with the tendency to share being regulated by the parameters of the model. As α increases, the number of clusters used to represent each category increases. As γ increases, the degree to which these clusters are shared between categories decreases. The model can be used to estimate a set of densities associated with categories by inferring the values of these parameters and the settings of the latent variables – the assignment of stimuli to clusters and clusters to categories – that make the observed data most probable. This can be done using standard Markov chain Monte Carlo algorithms developed for the HDP (Teh et al., 2006; Griffiths et al., 2007a). As a consequence, this model allows us to explore how the capacity to share clusters between categories influences category learning by comparing the performance of models that allow sharing (with γ free to vary) and models that do not (with $\gamma \to \infty$). We can compare the results of both of these types of models to human performance on categorization tasks to explore whether cluster-sharing models provide an explanation for certain types of human transfer learning.

3. Motivation for experiments

Most empirical studies of human categorization focus on learning category systems in isolation, eliminating the possibility of transfer effects. Instead, we are interested in measuring the extent to which the inductive biases that guide human learning of a new category depend on the previous or simultaneous learning of other categories. To this end, we conducted two experiments in which participants learned multiple categories over the same set of stimuli, creating an opportunity for domain knowledge to be transferred between categories.

When learners engage in transfer learning in a categorization task, they are not simply learning what objects belong to each category; they are also learning how to learn about categories in general. This meta-level knowledge can be described in terms of overhypotheses that govern what types of categories are more likely *a priori* to be observed than others. In the hierarchical Bayesian framework, this is realized by placing a distribution over the categories in which they are probabilistically dependent, so knowledge about one category affects the distribution over the others. The categories in our experiments are constructed so that the HDP can express the relevant overhypotheses, with certain clusters of objects reappearing across categories. We conducted two experiments, one with discrete stimuli and one with continuous stimuli, to evaluate the extent to which this type of transfer learning is evident in human learning.

4. Experiment 1: Discrete stimuli

4.1. Method

Fifty-seven undergraduate students, aged 18 to 27, participated in the experiment for course credit. The experiment is comprised of three sessions in which participants learn to partition a set of stimuli into two categories. The stimuli have four binary-valued features, and the same $2^4 = 16$ objects are used in each session. The category members are chosen as follows. In the first session, the stimuli are split into two categories based on their feature value (0 vs. 1) on a single dimension. In the second session, their category membership is based on their feature value on a second dimension. The two relevant dimensions are randomly chosen for each participant. In the third session, category membership is determined using an exclusive-or (XOR) rule. The two dimensions involved in the XOR rule depended on which of two conditions the participant is assigned to. In the *consistent* condition, the rule is an XOR over the same two dimensions used in the first two sessions. In the *inconsistent* condition, the XOR rule involves the remaining two dimensions.

The stimuli are adopted from a previous categorization experiment (Sakamoto & Love, 2004) and consist of square figures with five binary-valued dimensions: size (small or large), color (blue or purple), border (yellow or white), texture (smooth or dotted), and diagonal cross (present or absent). The dimensions were found to be independent and equally salient by a multidi-

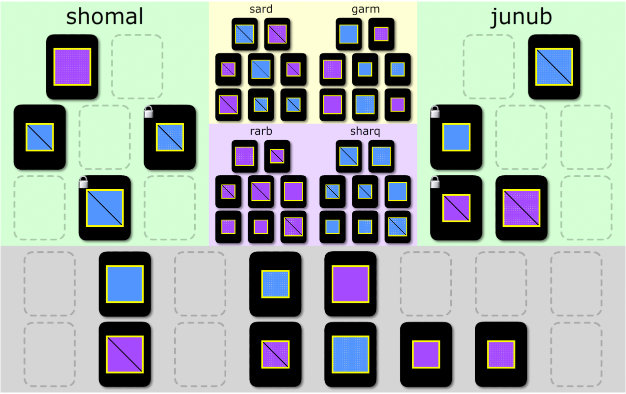

Figure 1. A screenshot from the third session of Experiment 1. Four cards are locked, four unlocked cards have been placed, and eight remain to be placed. The categories of the first two sessions are shown for reference.

mensional scaling of similarity ratings collected from a separate pool of participants (Sakamoto & Love, 2004). In our experiment, each participant is exposed to only 16 of the 32 possible stimuli, selected by choosing a random feature value, for example, a yellow border, and eliminating all stimuli with that feature.

Figure 1 shows a screenshot from the experimental setup. The stimuli are depicted on cards, and participants categorize them by dragging each card into one of two regions labelled with the two category names. When a participant finishes placing all of the cards, one card is randomly chosen, and the participant receives visual feedback telling them whether that card was correctly categorized or not. If the card is correct, it is locked in place, and feedback continues with another randomly-chosen card. If it is incorrect, it is moved to the correct category and locked in place, and the remaining unlocked cards are moved to the bottom of the screen for the participant to try again. Each time a card is locked is considered the end of a block. The number of correctly categorized cards per block is recorded for each participant.

4.2. Results

The results are shown in Figure 2 (a). As expected, the human learners were able to quickly achieve a high level of accuracy on the two training sessions, in which the stimuli were split based on just one dimension. They learned more slowly in the test conditions, in which category membership was based on an XOR rule. Most interestingly, there was a statistically significant difference between the observed accuracies in the consistent and inconsistent test conditions. The experiment was designed so that the same feedback is presented to the participants in both test conditions until block 9, so any differences in performance before then are statistical noise. Correspondingly, we performed a one-sided t-test of the null hypothesis that the the average accuracy over blocks 10 through 15 for the participants in the consistent condition was less than or equal to that of the participants in the inconsistent condition, producing $p = 0.032$. Participants were more accurate when learning an exclusive-or (XOR) rule based on the two dimensions that they used in the training sessions than for an XOR rule based on the other two dimensions.

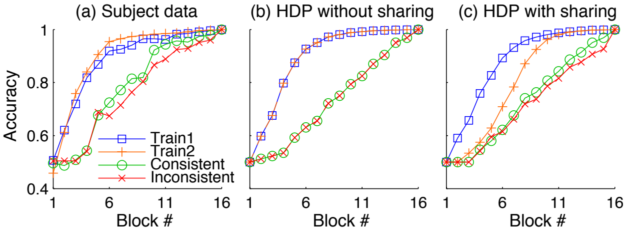

Figure 2. Results from Experiment 1. Each plot shows the accuracy in categorizing the stimuli for each of the 16 blocks of each session.

4.3. Modeling

To evaluate the degree to which the type of transfer learning used by the HDP can explain the effect found in the human data, we ran an HDP model with cluster-sharing (with $0<\gamma<\infty$) and one without cluster-sharing (with $\gamma \to \infty$) on the same tasks. The models were evaluated by separately estimating the probabilities of all 74 assignments of stimuli to categories (modulo rotations and inversions) using Gibbs sampling (burn-in of 10000, followed by 50000 samples, with concentration parameters being estimated from the data). Calculation of the posterior probability of an inferred category structure, conditioned on observed stimuli, then reduced to looking up one of the 74 pre-computed probabilities and dividing by the sum of the probabilities of the assignments consistent with the observed data.

The performance of the two learning algorithms is shown in Figure 2 (b) and (c). The HDP without sharing treats categories independently; therefore, the two training sessions are equivalent under this model, as are the two test conditions. It achieves a good fit to human performance for all the individual sessions, but it fails to account for the difference in performance between the two test conditions. The HDP model with sharing shows an increased learning rate for the consistent test condition compared to the inconsistent test condition, providing a possible explanation for the difference found in human performance.

Interestingly, the HDP model with sharing also predicts a transfer effect between the two training sessions, where the second training session is learned more slowly than the first. One possible explanation for why this effect is not found in the human data is that participants were not sufficiently aware that the category systems in the first two sessions are meant to be related to each other. The fact that the HDP model without sharing accurately predicts human performance for the first two sessions provides evidence for this hypothesis (the difference in human performance between the two training sessions is not statistically significant: running the same type of t-test used to compare the test sessions resulted in $p = 0.12$).

The concentration hyperparameters for the models were chosen to best fit the human data by visual inspection. The parameters for the HDP model without sharing are as follows. The base distribution for each Bernoulli parameter was set at $\text{Beta}(0.1, 0.1)$, and the concentration parameter α was learned from the data using a $\text{Gamma}(10, 14)$ prior distribution. The best-fitting hyperparameters for the HDP model with sharing were as follows. For the two training sessions, the base distribution for each Bernoulli parameter was set at $\text{Beta}(0.7, 0.7)$, the concentration parameter α was learned from the data using a $\text{Gamma}(10, 50)$ prior distribution. For the two test sessions, the base distribution for each Bernoulli parameter was set at $\text{Beta}(0.1, 0.1)$, and the concentration parameter α was learned from the data using a $\text{Gamma}(10, 25)$ prior distribution. For all four sessions, the concentration parameter γ was learned from the data using a $\text{Gamma}(10, 6)$ prior distribution. The posterior distributions were fairly sensitive to these prior distributions; this indicates that to the extent that the HDP model accurately describes the way people represent categories, human performance can provide information about their prior distributions over concentration parameters. For all sessions and both models, the estimated posterior distribution was raised to the power of 1.5 before sampling the model's responses, as in the exponentiated Luce choice rule often used in psychological categorization models (Kruschke, 1992).

Although the increased learning rate for participants in the consistent test condition provides evidence for the HDP model of categorization, this effect could also be explained in terms of dimensional salience parameters. Attention parameters allow a model to learn after the first two training sessions that dimensions 1 and 2 are more salient than dimensions 3 and 4 because they have been more relevant for determining category membership. Models can then use these parameters to increase the significance of differences along dimensions 1 and 2 in subsequent learning sessions, leading to the same observed transfer learning effect. Our second experiment eliminates the use of dimensional salience parameters as a potential explanation for transfer learning, and broadens our investigation to a new experimental paradigm and stimulus set.

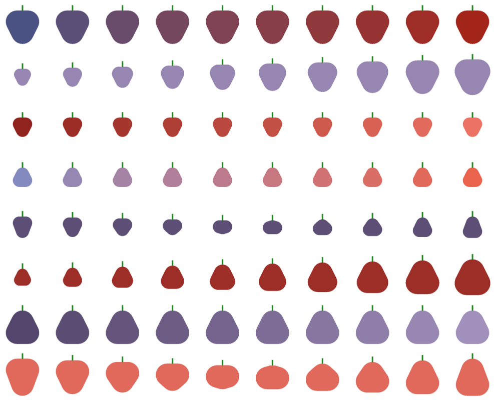

Figure 3. Examples of stimuli from the eight clusters used in Experiment 2. Each row contains ten examples from a cluster. Each cluster varies along one of the four dimensions: size, shape, hue, and lightness. For each of the four dimensions, there are exactly two clusters varying along that dimension, so that after learning the clusters, none of the dimensions is more globally salient than the others.

5. Experiment 2: Continuous stimuli

5.1. Method

The experiment was completed by 73 participants, aged 18 through 42, who were compensated with either course credit or payment. The experiment is comprised of three sessions: a pre-test session, a training session, and a post-test session. The two test sessions are identical, except that their trials are presented in random orders. Thus, the experiment allows us to measure the extent to which the training session alters participants' inductive biases about new categories by looking for any differences in performance between the pre-test and post-test sessions.

Participants are given a cover story about being a farmer's apprentice and having to learn about four different types of crops. The stimuli are adopted from a previous psychological experiment (Sanborn et al., 2009). They are generated using six parameters, but we selected a four-dimensional manifold within the original space to create our stimuli. The four roughly psychologically-independent dimensions we selected correspond to size, shape, hue, and lightness. Within this 4-dimensional space, we chose eight multivariate Gaussian distributions to serve as clusters which comprise the categories. The eight clusters are depicted in Figure 3. Each cluster is axis-aligned, i.e., has zero correlation between the dimensions, and has non-negligible variance in only one of the dimensions. Furthermore, there are exactly two clusters that vary in any given dimension, so given on observations from all clusters, the dimensions are equally salient. Figure 4 depicts the arrangement of the eight clusters in the four-dimensional space in which they vary.

The pre-test is comprised of 48 trials, during each of which, participants are presented with a single stimulus observation and told to imagine that it belongs to a novel category. They are then asked to click on one of two alternative stimuli that appear below it, indicating which one they believe to be more likely to belong to the same category as the first stimulus. In each of these trials, the observed stimulus and one of the two alternatives are generated from one of the eight Gaussian clusters, and the other alternative is generated from a different cluster. Critically, the two alternatives are equidistant from the observed stimulus in the underlying 4-dimensional space. For example, if the observed stimulus in a test trial occupies the front-upper-left corner of the left cube in Figure 4, the two alternatives would occupy the rear-upper-left and front-upper-right corners of the left cube. Since the participants have not yet been exposed to the clusters of stimuli, they have no reason to prefer one test alternative over the other, and we predict they will choose the within-cluster alternative 50% of the time on average. We record the percentage of pre-test trials for which the within-cluster alternative is selected and call this the participant's pre-test score.

Following the pre-test is the training session, in which participants are asked to help the farmer harvest four different types of crops. The eight clusters are randomly grouped into pairs to form the four categories, and participants repeat training blocks until they learn to differentiate between the four categories. First, an 8×8 grid of crops is displayed. Eight stimuli are generated from each cluster and randomly positioned to fill the grid. Participants are then asked to harvest all of the crops belonging to one of the four categories, which are labeled with the nonsense words *relts*, *pasps*, *worbs*, and *broms*. Participants harvest crops by clicking on them, and each time a correct stimulus is chosen, it is moved to the right side of the screen in a region corresponding to the appropriate category. Once all the category members are harvested, the field is randomly repopulated and the participant is asked to harvest the next group of crops, rotating through all four until the performance criterion (no more than four errors in the last four blocks of training) is met.

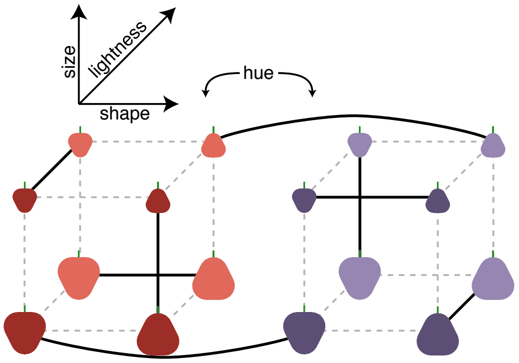

Figure 4. Spatial orientation of the stimuli in Experiment 2. A 4-dimensional hypercube is depicted as two cubes, with the dimensions labeled above. The probability density of the stimuli under each of the eight clusters is an axis-aligned multivariate normal distribution with non-negligible variance in only one dimension. The clusters are shown as the thick black edges of the hypercube; the other edges are dashed gray lines.

The post-test follows the training session. It is identical to the pre-test, with the trials randomly permuted. Again, we record the percentage of post-test trials for which the within-cluster alternative is selected and call this the participant's post-test score.

5.2. Results

The results of the experiment are given in Figure 5 (a). All of the pre-test scores are clustered around 50%, as expected. Although many of the post-test scores are also clustered around 50%, a significant number of human learners exhibited a transfer learning effect where after being exposed to the eight Gaussian clusters during training, they preferred to generalize novel categories to stimuli within the same cluster as observed stimuli. A one-tailed, within-subjects t-test confirmed that the difference between pre-test scores and post-test scores is statistically significant ($p < 10^{-8}$).

Further analysis of the results suggested that the post-test transfer learning was an effect of participants preferring to generalize to stimuli within the same category as the observed stimulus, rather than the more specific preference for generalizing to stimuli within the same cluster. When the analysis is restricted to only those test trials where both alternatives are within the same category as the observed stimulus (but one is in the same cluster and one is not), participants perform at chance (scoring near 50% on average) for the post-test. The results thus indicate a strong transfer learning effect, but one in which participants assume that entire categories, rather than clusters within those categories, can be repeated.

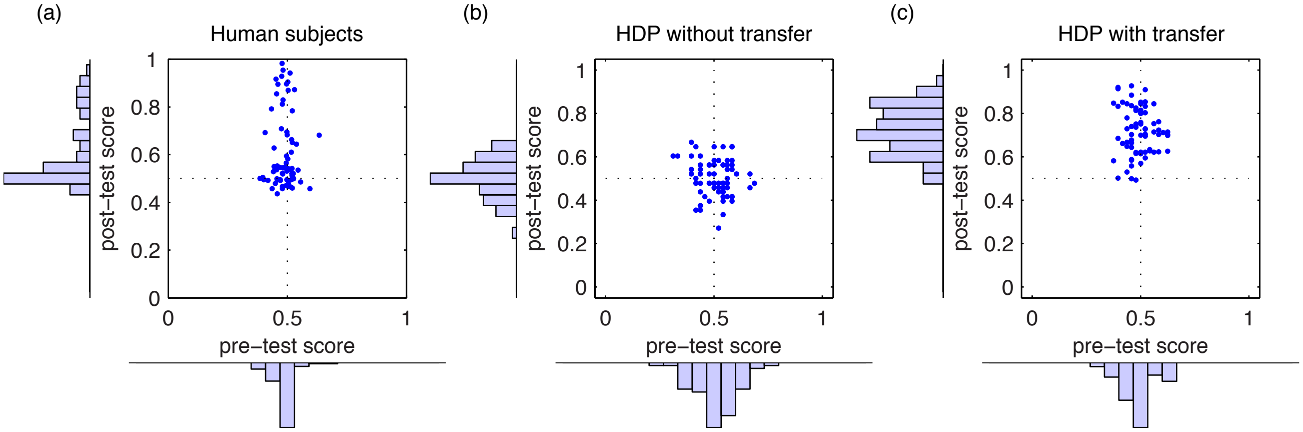

Figure 5. Results from (a) the human learners and (b,c) the two models in Experiment 2. Each marker represents one participant or one sample from the models. The x axis on each plot is the percentage of test trials where the learner generalized a novel category to a stimulus within the same cluster as the observed stimulus. The y axis is the same percentage for the post-test trials. Marginal histograms are shown for each axis. Before training, participants are very tightly clustered around 50% (random chance for trials with two options), while after training, some participants show a strong bias to generalize to within-cluster fruit. This post-test bias is also evident in the cluster-sharing HDP.

5.3. Modeling

As in Experiment 1, we modeled the human learners using both a cluster-sharing and a non-cluster-sharing HDP model. One sample from each of the models was taken for each participant, with the models receiving the same amount of training data as the corresponding participant. The model predictions were made using an independent round of Gibbs sampling for each pre-test and post-test trial. In the cluster-sharing HDP model, the γ parameter was learned from the data using a Gamma($2, \frac{1}{10}$) prior distribution. In both models, the α parameters were learned using a Gamma($2, \frac{2}{3}$) prior distribution. The cluster-specific multivariate Gaussian densities were assumed to be axis-aligned, with parameters μ and σ^2 being drawn from a Gaussian-inverse gamma distribution with parameters $\lambda = \frac{1}{2}$, $\nu = 1$, $\alpha = 3$, and $\beta = 1$.

The model results are shown in Figure 5 (b) and (c). Before training, the models have no observations to condition on, so their pre-test scores are drawn from a Binomial($48, 0.5$) distribution. This aligns with the human performance, although the human learners seem to be more tightly clustered around 50%. This could be due to the use of particular classification strategies throughout the pre-test. For example, a participant who decides to categorize on the basis of color alone will have a pre-test score of exactly 50%.

Because the HDP model without cluster sharing treats categories independently, the training session has no effect on the post-test, so its post-test scores are also clustered around 50%. The HDP model with cluster sharing shows a strong transfer effect, as seen in the human data. However, this version of the model predicts that generalization should take place at the level of clusters rather than categories, which is inconsistent with the human data. Generalization at the level of categories rather than clusters could be produced by constraining the model to have only a single cluster per category, taking $\alpha \to 0$ rather than estimating α from the data. In future work we hope to further investigate this phenomenon, getting a better sense of the level at which people transfer knowledge of category structure.

6. Conclusion

Most theoretical and empirical studies of human categorization have focused on learning a small number of categories independently of each other. However, there is much evidence that humans use knowledge acquired during past category learning episodes when learning new categories of objects. In this paper, we evaluated the degree to which transfer learning in human categorization can be explained by simultaneously learning categories consisting of shared clusters. The categorization model we adopted to achieve this effect, the hierarchical Dirichlet process (HDP), is a generalization of several recently proposed rational models of human categorization. Although we do not propose the HDP as a complete account of transfer learning in humans, we believe it explains some types of transfer. We have experimentally confirmed through two experiments that human learners engage in transfer learning in categorization tasks, and we have shown that the HDP can provide an explanation for these effects.

Beyond the specific model and experiments we considered in this paper, we anticipate that understanding how people transfer knowledge from one domain to another will play a significant role in building automated systems capable of rapidly learning new concepts. The form of transfer learning investigated here establishes an interesting connection with machine learning, where hierarchical Bayesian models have been proposed as a formalism in which to develop systems capable of transfer learning (Baxter, 1997). Hierarchical Bayesian models can be used to express hypotheses at multiple levels of abstraction and can consequently form generalizations that span multiple domains. Such models are already being used to explain human transfer learning in a variety of settings, laying the foundations for a formal framework that brings human and machine learning closer together.

Acknowledgments

This work was supported by grants FA 9550-07-1-0351 from the Air Force Office of Sponsored Research and IIS-0845410 from the National Science Foundation.

References

Aldous, David. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, pp. 1–198. Springer, Berlin, 1985.

Anderson, John R. *The Adaptive Character of Thought.* Erlbaum, Hillsdale, NJ, 1990.

Anderson, John R. The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429, 1991.

Ashby, F. Gregory and Alfonso-Reese, Leola A. Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39:216–233, 1995.

Baxter, Jonathan. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28:7–39, 1997.

Brown, Ann L. and Kane, Mary Jo. Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20(4):493–523, 1988.

Griffiths, Thomas L., Canini, Kevin R., Sanborn, Adam N., and Navarro, Daniel J. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pp. 323–328, 2007a.

Griffiths, Thomas L., Sanborn, Adam N., Canini, Kevin R., and Navarro, Daniel J. Categorization as nonparametric Bayesian density estimation. In Chater, Nick and Oaksford, Mike (eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, pp. 303–328. Oxford University Press, 2007b.

Kruschke, John K. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44, January 1992.

Love, Bradley C., Medin, Douglas L., and Gureckis, Todd M. SUSTAIN: A network model of category learning. *Psychological Review*, 111(2):309–332, 2004.

Medin, Douglas L. and Schaffer, Marguerite M. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

Nosofsky, Robert M. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57, 1986.

Perfors, Amy and Tenenbaum, Joshua B. Learning to learn categories. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.

Pitman, Jim and Picard, Jean. *Combinatorial Stochastic Processes.* Springer, Berlin, 2006.

Reed, Stephen K. Pattern recognition and categorization. *Cognitive Psychology*, 3:393–407, 1972.

Rosseel, Yves. Mixture models of categorization. *Journal of Mathematical Psychology*, 46:178–210, 2002.

Sakamoto, Yasuaki and Love, Bradley C. Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, 133(4):534–553, 2004.

Sanborn, Adam N., Griffiths, Thomas L., and Shiffrin, Richard M. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, 2009.

Teh, Yee Whye, Jordan, Michael I., Beal, Matthew J., and Blei, David M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101 (476):1566–1581, December 2006.

Thrun, Sebastian and Pratt, Lorien (eds.). *Learning to Learn.* Kluwer Academic Publishers, 1998.

Vanpaemel, Wolf, Storms, Gert, and Ons, Bart. A varying abstraction model for categorization. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 2277–2282, 2005.