Bayesian teaching of image categories

Wai Keen Vong* **(waikeenvong@gmail.com),**
Ravi B. Sojitra* (ravisoji@gmail.com),
Anderson Reyes (reyesanderson428@gmail.com),
Scott Cheng-Hsin Yang (scott.cheng.hsin.yang@gmail.com),
Patrick Shafto (patrick.shafto@gmail.com)
Department of Mathematics and Computer Science, 110 Warren Street,
Newark, NJ, 07102

Abstract

Humans learn from other knowledgeable informants who choose data to foster learning. Mathematical models of teaching and learning have formalized this process of learning from helpful others. While these approaches have been successful in capturing teaching and learning in a variety of contexts, they have been limited to relatively simple domains. One of the open questions regarding Bayesian teaching is whether it can scale to teach from naturalistic domains with more interesting datasets. In this work, we show how to apply Bayesian teaching to teach human participants categories learned by a supervised machine learning model. The effectiveness of teaching is measured by how well the participants can predict the behavior of the target machine learning model. Our results demonstrate that Bayesian teaching can be applied to naturalistic domains, show that the best sets of examples according to the model yield better learning, and suggest avenues for improving our ability to automate teaching of image categories.

Keywords: Bayesian teaching; category learning; pedagogy; prototype model

Introduction

Teaching is a common method of knowledge transmission, which occurs in both formal and informal contexts (Csibra & Gergely, 2009). In such pedagogical situations, a knowledgeable and helpful informant—a teacher—provides data or examples that best communicate concepts to a learner, who in turn assumes that the data presented is intended to be helpful, allowing them to learn more efficiently than other methods. This mutual cooperation toward the goal of learning has been formalized in probabilistic models of Cooperative Inference, which involves recursive reasoning by the teacher and learner (Shafto & Goodman, 2008; Shafto, Goodman, & Griffiths, 2014; Yang et al., 2018), and is a generalization of Bayesian learning and Bayesian teaching.[†]

These models of teaching have been successful in capturing teaching behavior, including the implications for learning, in a variety of laboratory studies. For example, Shafto and Goodman (2008) showed that in a simple concept learning game (the "rectangle game"), participants both selected examples in line with Bayesian teaching and rapidly identified the target hypothesis when presented with those pedagogical examples. Similarly, work by Bonawitz, Shafto et al. (2011) showed that when pre-schoolers were shown various functions of a toy, those provided pedagogically caused them to explore less. Further, Eaves, Fledman, Griffiths, and Shafto (2016) showed that infant-directed speech is consistent with the sounds Bayesian teaching would produce to teach phonetic categories of adult speech. Rafferty, Brunskill, Griffiths, and Shafto (2016) used Bayesian teaching in a planning problem to improve human performance in simple concept-learning tasks. Finally, Ho, Littman, MacGlashan, Cushman, and Austerweil (2016) explored Bayesian teaching for reinforcement learners, showing that examples provided by teaching differ from following the policy that maximizes an agent's utility.

One desideratum for computational models of teaching is the automatic selection of examples to teach relevant, real-world concepts. However, due to computational constraints, successes have been limited to small, schematic domains characteristic of concept learning in the lab. Thus, one of the open questions regarding this framework is whether it can scale to teach realistic domains with large, complex, naturalistic data sets. Extracting information from such domains can be made efficient with machine learning models, which can process vast datasets much faster than humans do. In this view, teaching a domain becomes a matter of teaching the machine learning models that are trained on the relevant datasets.

In this paper, we explore this problem by investigating Bayesian teaching with image categories. We adapt a prototype-based machine learning model (Probabilistic Linear Discriminant Analysis; Ioffe, 2006), formalize teaching for this model, and run a classification experiment to test the effectiveness on teaching image categories. The effectiveness of teaching is measured by how well humans can predict the machine learning model's predictions. Our results indicate that Bayesian teaching is helpful for learning what the model learns about natural image categories.

Bayesian Teaching

The goal of Bayesian teaching is to select small subsets of data that induce a target model in the learner (Shafto & Goodman, 2008; Shafto, Goodman, & Griffiths, 2014; cf. Griffiths & Tenenbaum, 2001) . In this paper, given a set of training data $D = \{d_1, d_2, \ldots, d_N\}$ and a teaching set size $n < N$, Bayesian teaching conveys a target model Θ^* to a learner by sampling a teaching set $\mathcal{D} \subseteq D$ from the space of possible candidate teaching sets, $\boldsymbol{\mathcal{D}} = \{\mathcal{D} \big| \mathcal{D} \in \mathcal{P}(D) \wedge |\mathcal{D}| = n\}$, ac-

[*]Both authors contributed equally to this paper.

[†]Cooperative inference does not require that one necessarily be more knowledgeable, only that there be a target hypothesis.

cording to Bayes' rule

$$P_T(\mathcal{D} \mid \Theta^*) = \frac{p_L(\Theta^* \mid \mathcal{D})p_T(\mathcal{D})}{\sum_{\mathcal{D}' \in \boldsymbol{\mathcal{D}}} p_L(\Theta^* \mid \mathcal{D}')p_T(\mathcal{D}')}$$
$$= \frac{p_L(\Theta^* \mid \mathcal{D})}{\sum_{\mathcal{D}' \in \boldsymbol{\mathcal{D}}} p_L(\Theta^* \mid \mathcal{D}')}, \quad (1)$$

where $\mathcal{P}(D)$ is the power set of D, $p_T(\mathcal{D})$ is the prior over teaching sets, $p_L(\Theta^* \mid \mathcal{D})$ is the learner's posterior and $P_T(\mathcal{D} \mid \Theta^*)$ is the teacher's posterior. Since the prior over teaching sets $p_T(\mathcal{D})$ is assumed to be uniform, the effect of this prior cancels out, yielding the second equality. Therefore, the posterior probability of selecting a particular teaching set to teach a target model is proportional to the learner's posterior probability of the target model after observing the same teaching set.

Probabilistic Linear Discriminant Analysis (PLDA)

PLDA is a supervised learning model, taking in training data with class labels and can be used for classification of new, unlabeled data (Ioffe, 2006). We use PLDA as the basis for Bayesian teaching of image categories for two reasons: (1) PLDA has previously been applied to supervised classification of image categories with good performance (Ioffe, 2006) and (2) PLDA is a probabilistic model which makes it amenable to Bayesian teaching. The model assumes that both the means of the categories and examples from each category are samples from multivariate Gaussian distributions, and its objective is to maximize the distance *between* the category means while also minimizing the distance between examples *within* each category.

Formally, the generative model is

$$\boldsymbol{v}^k \sim N(\boldsymbol{0},\ \boldsymbol{\Psi}),$$
$$\boldsymbol{u}_i^k \sim N(\boldsymbol{v}^k,\ \boldsymbol{I}),$$
$$\boldsymbol{d}_i^k = \boldsymbol{m} + \boldsymbol{A}\boldsymbol{u}_i^k.$$

The category means $\boldsymbol{v}^k$ are sampled from a multivariate Gaussian distribution with mean $\boldsymbol{0}$ and diagonal covariance $\boldsymbol{\Psi}$. Then, for each category k, a sample $\boldsymbol{u}_i^k$ is drawn from a multivariate Gaussian with mean $\boldsymbol{v}^k$ and identity covariance. Finally, samples from all categories are linearly transformed from latent space to the data space with shift $\boldsymbol{m}$ and rotation $\boldsymbol{A}$. Under this model, $\boldsymbol{\Psi}$, $\boldsymbol{m}$, and $\boldsymbol{A}$ are free parameters and fitted via maximum likelihood of the data.

Given the fitted parameters and a set of data $\boldsymbol{d}_1^k, \boldsymbol{d}_2^k, ..., \boldsymbol{d}_{N^k}^k$ for category k, we transform the data to $\boldsymbol{u}_1^k, \boldsymbol{u}_2^k, ..., \boldsymbol{u}_{N^k}^k$ in latent space, and the posterior on $\boldsymbol{v}^k$ is

$$p_L(\Theta^* \mid \mathcal{D}) = p_L(\boldsymbol{v}^k \mid \boldsymbol{u}_1^k, \boldsymbol{u}_2^k, ..., \boldsymbol{u}_{N^k}^k) \quad (2)$$
$$= \frac{N(\boldsymbol{v}^k|\boldsymbol{0},\boldsymbol{\Psi})\prod_i^{N^k} N(\boldsymbol{u}_i^k|\boldsymbol{v}^k,\boldsymbol{I})}{\int_{\boldsymbol{v}} N(\boldsymbol{v}|\boldsymbol{0},\boldsymbol{\Psi})\prod_i^{N^k} N(\boldsymbol{u}_i^k|\boldsymbol{v},\boldsymbol{I})d\boldsymbol{v}}$$
$$= N(\boldsymbol{v}^k|N^k\boldsymbol{\Lambda}_k\bar{\boldsymbol{u}}^k, \boldsymbol{\Lambda}_k), \quad (3)$$

where $\boldsymbol{\Lambda}_k = \frac{\boldsymbol{\Psi}}{N^k\boldsymbol{\Psi}+\boldsymbol{I}}$ and $\bar{\boldsymbol{u}}^k = \frac{1}{N^k}\sum_i^{N^k} \boldsymbol{u}_i^k$.

The posterior predictive probability for a datum $\boldsymbol{u}^*$ is given by

$$p_L(\boldsymbol{u}^* \mid \boldsymbol{u}_1^k, \boldsymbol{u}_2^k, \ldots, \boldsymbol{u}_{N^k}^k) = \int_{\boldsymbol{v}} p_L(\boldsymbol{u}^* \mid \boldsymbol{v}) p_L(\boldsymbol{v} \mid \boldsymbol{u}_1^k, \boldsymbol{u}_2^k, \ldots, \boldsymbol{u}_{N^k}^k) d\boldsymbol{v}$$
$$= N(\boldsymbol{u}^*|N^k\boldsymbol{\Lambda}_k\bar{\boldsymbol{u}}^k, \boldsymbol{\Lambda}_k + \boldsymbol{I}). \quad (4)$$

which is used for classification of new, unlabeled data by computing this for each category k and selecting the category with the highest posterior predictive probability.

Generating teaching sets requires three steps: (1) train a PLDA model on labeled data to obtain a target model, (2) use the target model's predicted labels (not the training labels) for teaching because the target model is what we wish to convey, and (3) generate teaching sets by using Equation (5) below.

Training the target model. The training of the PLDA target model is described in the previous section and is done on a preprocessed dataset containing images of faces with emotion labels (see the Dataset and Preprocessing sections). To obtain the target model's predictions for each image, we first compute the posterior predictive probabilities with respect to each category using Equation (4), and then select the category with the highest probability to be the predicted label.

Generating teaching sets. The representation leading to the target model's predictions is defined by the parameters $\boldsymbol{\Psi}$, $\boldsymbol{m}$, and $\boldsymbol{A}$ and the posterior distributions over the mean of each category. Each of these distributions for category k is characterized by its mean $\boldsymbol{v}^* = N^k\boldsymbol{\Lambda}_k\bar{\boldsymbol{u}}^k$, and the teacher's objective is to convey these category means to a learner by generating teaching sets.

To do this, the teacher assumes the learner to have the same $\boldsymbol{\Psi}$, $\boldsymbol{m}$, and $\boldsymbol{A}$ as the target model, but not necessarily the same category means. Explicitly, as given by Equation (1), the teaching equation for teaching a particular category learned using PLDA is:

$$P_T(\mathcal{D} \mid \Theta^*) = \frac{p_L(\Theta^* \mid \mathcal{D})}{\sum_{\mathcal{D}' \in \boldsymbol{\mathcal{D}}} p_L(\Theta^* \mid \mathcal{D}')}$$
$$= \frac{N(\boldsymbol{v}^*|n^k\boldsymbol{\lambda}_k\bar{\boldsymbol{u}}^k, \boldsymbol{\lambda}_k)}{\sum_{\bar{\boldsymbol{u}}^{k'} \in U_{\boldsymbol{\mathcal{D}}}} N(\boldsymbol{v}^*|n^{k'}\boldsymbol{\lambda}_{k'}\bar{\boldsymbol{u}}^{k'}, \boldsymbol{\lambda}_k)}, \quad (5)$$

where $\boldsymbol{\lambda}_k = \frac{\boldsymbol{\Psi}}{n^k\boldsymbol{\Psi}+\boldsymbol{I}}$.[‡] Here, the teacher samples a teaching set $\mathcal{D}$ to teach $\Theta^* = \boldsymbol{v}^*$. Given a dataset size N^k and a teaching set size n^k such that $N^k > n^k$, the number of possible teaching sets in $\boldsymbol{\mathcal{D}}$ is $\binom{N^k}{n^k}$. To compute the individual $p_L(\Theta^* \mid \mathcal{D}')$, each data point in $\mathcal{D}'$ is first transformed into latent space. In latent space, $p_L(\Theta^* \mid \mathcal{D}') = N(\boldsymbol{v}^*|n^k\boldsymbol{\lambda}_k\bar{\boldsymbol{u}}^k, \boldsymbol{\Lambda}_k)$, where $\bar{\boldsymbol{u}}^k = \frac{1}{n^k}\sum_i^{n^k} \boldsymbol{u}_i^k$ and the $\boldsymbol{u}_i^k$'s are the transformed data points. Note, $U_{\boldsymbol{\mathcal{D}}}$ simply denotes the space of $\bar{\boldsymbol{u}}^k$ that emerges from applying both the transformation and computing the average on each teaching set in $\boldsymbol{\mathcal{D}}$.

Simulating the learner. We simulate human behavior in the 2AFC task (see Experiment section) using Equation (4).

[‡]In the experiment, since participants were shown three examples from each teaching set, we set $n^k = 3$ for all k.

We compute the posterior predictive probability of a target image $\boldsymbol{u}^*$ belonging to either category, which are illustrated by teaching sets with three examples each. Specifically, this is done by computing the following:

$$p_L(k^* = T|\boldsymbol{u}^*; \boldsymbol{u}^T_{1:3}, \boldsymbol{u}^O_{1:3}) = \frac{N(\boldsymbol{u}^*|n^T \boldsymbol{\lambda}_k \bar{\boldsymbol{u}}^T, \boldsymbol{\lambda}_k + \boldsymbol{I})}{\sum_{k' \in \{T,O\}} N(\boldsymbol{u}^*|n^{k'} \boldsymbol{\lambda}_k \bar{\boldsymbol{u}}^{k'}, \boldsymbol{\lambda}_k + \boldsymbol{I})}, \tag{6}$$

where T is the **target** category, and O is the **other** category (see below).

Experiment

We ran a study involving human participants recruited from Amazon's Mechanical Turk to determine whether it is possible to teach a trained machine learning model (PLDA) to participants. In our approach, we presented participants with sets of examples selected from the Bayesian teaching PLDA (BT-PLDA) model, manipulating the helpfulness of these teaching sets, and determining whether or not participants' responses matched the predictions of the target model.

Dataset

In order to teach a target model to participants, we required a dataset of images (with category labels) to train the target model with. Our main criteria was to use data sufficiently challenging for the model to learn completely, while also not being too easy for humans either.

Hence, we selected the Child Affective Facial Expressions dataset by LoBue and Thrasher (2015), which consists of images of children expressing a variety of different emotions. While people have ample experience with facial expressions of emotions, categorizing faces according to their emotion is challenging, with performance well under ceiling (LoBue & Thrasher, 2015). Moreover, in our experiment, we do not explicitly tell participants the images are categorized by emotion, which further increases the task difficulty.

The dataset consisted of 1192 images of children 2-8 years old, expressing six basic emotions (angry, disgust, fearful, happy, sad and surprise), in addition to a neutral facial expression. For the purposes of our task, we used a subset of this dataset consisting of mouth open versions of the six basic emotions (excluding neutral faces). This resulted in a dataset for training the model consisting of 484 images from six emotion categories (84 angry, 95 disgust, 61 fearful, 95 happy, 46 sad and 103 surprise).

Preprocessing

For each of the 484 images, we pre-processed the images by grayscaling and resizing them to be 400 $\times$ 400 pixels. We then applied Principal Components Analysis to further reduce the dimensionality of the dataset, keeping the first 75 principal components from all of the images, which captured $> 84\%$ of the variance from the original dataset. The target model for teaching was obtained by fitting PLDA to the pre-processed data.

Participants

105 participants (62 male, 43 female) were recruited from Amazon Mechanical Turk and paid $1.50 for completing the task, which took roughly 10 minutes to complete. The mean age of participants was 35.3 years (SD = 10.0), ranging from 18 to 64 years. 13 participants were not included in the analysis for completing the experiment too quickly (less than one second per trial).

Design

On each trial, participants were presented with a target image and asked to classify it into one of two categories (A or B), where one category matched the category of the target image and the other was randomly selected from one of the other five emotion categories. These categories were chosen and matched based on the ground-truth labels at this stage. The participants were presented with a teaching set of three example images to represent each category. These images were chosen not based on the ground-truth categories but from what the target model predicted to belong to each of the two categories respectively. These teaching sets varied in three between-subjects conditions which participants were randomly assigned to: HELPFUL ($N = 36$), RANDOM ($N = 36$) and UNHELPFUL ($N = 33$).

To generate the teaching sets for the HELPFUL and UNHELPFUL conditions, we applied Equation (5) to each of the six category means. Intuitively, this equation corresponds to the "goodness" of each teaching set, where a higher probability indicates the Bayesian teacher believes the learner will more likely infer the target model given that teaching set. Thus, for the HELPFUL condition, the teaching sets are the sets with the highest posterior probabilities as given by BT-PLDA, while in the UNHELPFUL condition, the teaching sets are sets with the lowest posterior probabilities instead. For the RANDOM condition, the teaching sets were randomly sampled from all possible sets for a particular category. This process was repeated for each category independently. Finally, if a selected teaching set contained the target image, the next best set not containing the target image was used instead.

According to the target model's predictions, there were 77 images for angry, 95 for disgust, 84 for fearful, 89 for happy, 59 for sad, and 80 for surprise. The number of possible teaching sets for each category is given by $\binom{M^k}{n^k}$, where M^k is the number of images the model predicts to be in category k and n^k is set to 3 for all categories, as we select three images in each teaching set.

Note that since the target model does not perfectly learn to correctly classify the image categories, the examples from the teaching sets generated in the various conditions sometimes included images that were from other categories according to ground-truth labels.

Procedure

Participants were randomly assigned to one of either HELPFUL, UNHELPFUL, or RANDOM teaching set conditions at the

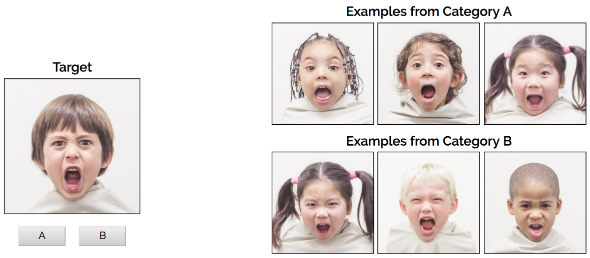

Figure 1: **Example of a trial from the task**. Participants were shown a target image (left), along with a teaching set of examples from both the **target** category (angry, bottom right) and the **other** category (surprise, top right), and asked to predict how the model would respond based on the examples provided. The teaching sets for the target and other categories were sampled from the set of examples that the *target model* considered to be in each category respectively.

beginning of the experiment. They were presented with instructions indicating that a robot had learned to categorize faces into different categories and that this robot would provide helpful examples to help them understand what the robot had learned. The goal for participants was to predict the robot's choice in categorizing the target images, using the examples provided on each trial to help them out.

On each trial, participants were presented with a target image on the left of the screen and asked "Does the robot think the following Target face on the left is a member of Category A or Category B?". On the right, participants were shown a row of three examples from the **target** category and a row of three examples from one of the **other** remaining five categories (based on the ground-truth labels). Again, the helpfulness of the examples as predicted by the teaching model varied based on which **teaching set** condition participants had been assigned to.

The position (upper or lower row, i.e., Category A or Category B) of the **target** category examples and the **other** category examples were randomized on each trial such that on half of trials, examples from the **target** category appeared as examples from Category A (upper row), and on the other half as examples from Category B (lower row), and vice versa for the **other** category. Participants did not receive any feedback after each response. During the experiment, they completed 120 categorization trials in total, 20 trials for each emotion category being the **target** category, while the **other** categories were selected randomly based on each trial; no target image was presented more than once.

Results

Because the target model's predictions differed from the ground truth labels of some of the target images, we removed the set of trials for which the target model's prediction of the target image did not match either the **target** category or the **other** category for that trial. This left 88 of 120 trials for analysis, 78 of which were cases where the prediction of the PLDA target model matched the **target** category, and 10 trials where the prediction matched the **other** category. The analysis presented here shows the extent to which participants' responses match the predictions of the target model on these 88 trials, and whether varying the "goodness" of teaching sets influenced whether participants responses matched the predictions of the target model.

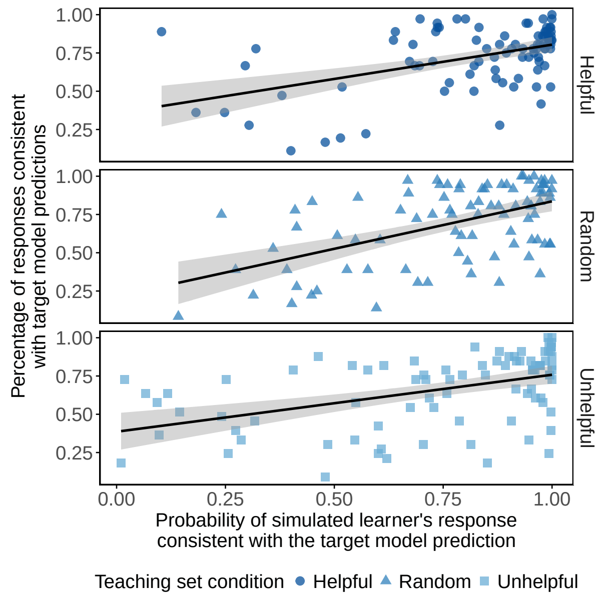

Figure 2: **How well do simulated learner's responses match human responses?** As human responses become more consistent with the target model's predictions, the simulated learner's certainty about the same predictions also increases.

First, did participant's judgments actually match the behavior of the simulated learner? If so, then the Bayesian teaching approach holds promise in generating teaching sets that influence human responses. To verify this, for each trial we examined the probability that the simulated learner would choose the correct category (correct is w.r.t. the target model's prediction of the target image) given the two sets of examples using Equation 6, and compared this to how well human behavior matched the target model, which is illustrated in Figure 2. The results indicate that the simulated learner matched how humans responded in the task ($r(262) = 0.49$, $p < .001$).

Second, did the various **teaching set** conditions lead to differences in how well participants' responses matched the model predictions? Mean performance across the three **teaching set** conditions are shown in Figure 3 on the left. Performance was highest for participants in the HELPFUL condition (M = 72.5%, SD = 2.1%), followed by the RANDOM condition (M = 69.3%, SD = 2.0%) and finally the UNHELPFUL condition (M = 66.6%, SD = 2.4%). We conducted a planned contrast across the different **teaching set** conditions (with HELPFUL = 1, RANDOM = 0 and UNHELPFUL = -1) and found a significant effect of **teaching set** condition on accu-

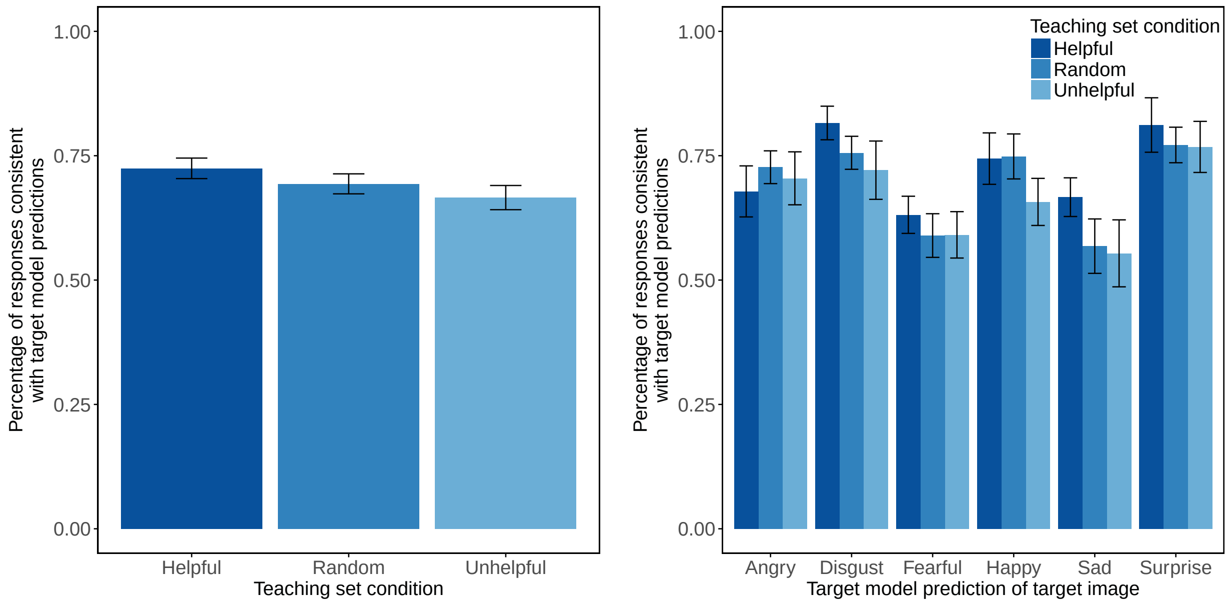

Figure 3: **Performance (percent of responses consistent with the target model's prediction) across different TEACHING SET conditions**. Mean performance of participants in the UNHELPFUL, RANDOM and HELPFUL conditions. Error bars depict 95% confidence intervals. Overall, results show that performance is best in the HELPFUL condition, followed by the RANDOM condition and then the UNHELPFUL condition. The same pattern of results holds when breaking down performance by each emotion category (as predicted by the target model), with performance varying depending on the category.

racy to the target model predictions ($F(1, 102) = 6.29$, $p = .013$).

Furthermore, we explored how well participants' responses matched the model's predictions for each emotion separately, as shown on the right in Figure 3. A two-way ANOVA revealed significant main effects for **teaching set** condition ($F(2, 612) = 8.87$, $p < .001$) and **model emotion** ($F(5, 612) = 32.48$, $p < .001$). There was a marginal, but not significant interaction between the two variables ($F(10, 612) = 1.84$, $p = .051$), suggesting that the effect of **teaching set** condition was consistent across emotion categories.

Discussion

This work asked two main questions: First, is it possible to scale the Bayesian teaching approach to more difficult learning problems such as image categorization? And second, can we use this approach to teach participants what a trained machine learning model has learned? We augmented a prototype-based model of categorization to generate teaching sets that varied in quality as predicted by the Bayesian teaching model and ran an experiment to compare how different teaching sets could teach participants the target model's knowledge about different image categories.

Overall, our results provide support to both of these ideas. The Bayesian teaching PLDA model allowed us to generate teaching sets of varying quality and the experimental results show that teaching sets with higher teaching probability in the BT-PLDA model produced a higher proportion of responses that matched the predictions of the PLDA target model.

However, the effects of different teaching sets was relatively small. How can we improve their effectiveness? One possibility is that the BT-PLDA model selected examples to teach from the target and other categories independently, ignoring both the target image and examples from the other category provided when generating its teaching sets. Alternatively, taking this information into account when generating teaching sets could potentially lead to generating more teaching sets that are actually helpful. For example, in Figure 2 many points in the top right are from the UNHELPFUL condition under the current BT-PLDA model but in fact helped both the simulated learner and human to perform well. Given the correlation between the human and simulated learner, one interesting research direction would be to design teaching sets that are based on the performance of the simulated learner for a particular task in a particular trial.

A second possibility is that participants may have relied on existing prior knowledge for this particular set of emotion image categories, and that the set of examples provided by the teaching model (regardless of the teaching set condition) may have been insufficient to shift humans from their prior. Further work exploring other image datasets, particularly for domains where people have less prior knowledge may be more fruitful in determining the effects of teaching sets in learning image categories.

This work provides a foundation for further exploration of using Bayesian teaching for teaching image categories. Further work could include extending the PLDA model to try and teach not only the category mean, but also the covariance of each category, or to optimize all of the presented stimuli simultaneously. This would allow for the testing and comparison of different kinds of teaching models to help determine what kinds of knowledge is most important to convey for effective learning. Another possibility would be to explore combining teaching examples with feedback. In the experiment presented in this work, participants were only given information about the modelâĂŹs knowledge implicitly through the examples provided, whereas presenting participants with feedback would allow one to measure learning over time and whether participants' knowledge begins to match the trained model based on which the examples are being generated. This research presents a first step toward programmatic approaches to scaleable methods of automating teaching of realistic domains of image categories.

Acknowledgments

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146 to P.S. and S.Y. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Eaves, B. S., Fledman, N., Griffiths, T., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review*, *123*, 758-771.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In *Advances in Neural Information Processing Systems* (pp. 3027–3035).

Ioffe, S. (2006). Probabilistic linear discriminant analysis. *Computer Vision–ECCV 2006*, 531–542.

LoBue, V., & Thrasher, C. (2015). The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, *5*, 1532.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via POMDP planning. *Cognitive science*, *40*(6), 1290–1332.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1632–1637).

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Tenenbaum, J. B., & Griffiths, T. (2001). The rational basis of representatives. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society.*

Yang, C.-H. S., Yu, Y., Givchi, A., Wang, P., Vong, W. K., & Shafto, P. (2018). Optimal cooperative inference. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics.*