
Interpretable Machine Teaching via Feature Feedback

Shihan Su, Yuxin Chen, Oisín Mac Aodha, Pietro Perona, Yisong Yue
California Institute of Technology

Abstract

A student’s ability to learn a new concept can be greatly improved by providing them with clear and easy to understand explanations from a knowledgeable teacher. However, many existing approaches for machine teaching only give a limited amount of feedback to the student. For example, in the case of learning visual categories, this feedback could be the class label of the object present in the image. Instead, we propose a teaching framework that includes both instance-level labels as well as explanations in the form of feature-level feedback to the human learners. For image categorization, our feature-level feedback consists of a highlighted part or region in an image that explains the class label. We perform experiments on real human participants and show that learners that are taught with feature-level feedback perform better at test time compared to existing methods.

1 Introduction

The goal of computer-assisted teaching of humans is to teach a new set of concepts to a learner as efficiently as possible. Efficiency is usually evaluated by the number of examples used during teaching. In the context of teaching visual categories, this is typically posed as selecting the most informative subset of images to show from a much larger set. However, one of the major limitations of existing approaches, e.g. [19, 10], is that they only give very limited feedback to students in the form of instance-level labels. In Fig. 1 (a) we see an example of this instance-level, i.e. class label, feedback. As opposed to only telling the student the correct class label, a knowledgeable human teacher would instead show them the specific parts and attributes that are informative for that particular class i.e. Fig. 1 (b).

We introduce an interpretable teaching algorithm that jointly selects informative images and features. This additional feature-level feedback provides more information than traditional instance-level only feedback, enabling the student to learn the task at hand more effectively. We validate the performance of our approach on two challenging image datasets with real human participants, and show superior results compared to existing methods.

2 Related Work

To date, a variety of different approaches have been explored for modeling the teaching of students such as assuming perfect learners [7, 23, 12], heuristic based approaches [2], Bayesian models [4, 6], recurrent neural networks [14], and reinforcement learning based approaches [16, 1, 21].

In the context of teaching binary visual classification tasks, Singla et al. [19] model the student as stochastically switching between a set of different hypotheses during learning. Their approach attempts to select the set of teaching examples offline that best guides the student towards the ground truth classification function. Johns et al. [10] propose an interactive approach, where the choice of future images to show is based on the individual’s past performance. However, the major limitation



Figure 1: (a) The majority of existing machine teaching algorithms that teach visual categorization tasks only give feedback at the image-level. (b) It is much more informative to display the discriminative regions that help determine the species. Here, by highlighting the important features we can see that the MacGillivray’s Warbler has a broken eye ring while the Connecticut Warbler’s is complete.

of these existing methods is that the feedback they provide to the student is not fully informative. In both cases, a student is shown a sequence of images and asked to estimate the object category they believe to be present in each image. They only receive the ground truth class label as feedback, and are not informed of the important features and parts in the image that indicate the correct label.

Clear and interpretable instructional material can improve a student’s ability to learn a new concept e.g. [8, 18]. For example, when a human teacher is unavailable, the most common way novices attempt to learn and perform species identification is by consulting expertly curated field guides. These field guides often come in the form of books or apps and contain descriptive text and example images highlighting important features for classifying different species e.g. [13]. Attempts have been made to automate the creation of these guides using highlighted part annotations [3] and image specific text descriptions [9]. [5] also showed that it is possible to collect distinctive image regions through gamification. An alternative approach that requires less additional annotations is to learn human interpretable models from the raw data e.g. [17, 11]. In the context of computer vision, there is some evidence to suggest that deep models commonly used in large scale image classification tasks learn features that are interpretable to humans [22].

Recently, Poulis and Dasgupta [15] outlined a method for incorporating additional supervised data from users which they call ‘feature feedback’. In addition to class level labels, their annotators provide information about the values of specific feature dimensions. Instead, our model selects the most informative image at each teaching iteration and presents it to the student. It then gives feedback to the *learner* regarding the importance of individual features and models how they incorporate this new information when updating their belief. In this work, we assume that an interpretable feature space is provided and leave the learning of these features for future work.

3 Teaching with Feature Feedback

Suppose we are given a set of m images $\mathcal{X} = \{x_1, \dots, x_m\}$, from which we can choose a subset (i.e., the *teaching set*) to teach the learner. Let $\mathcal{F} = \{f_1, \dots, f_n\}$ be the collection of all possible interpretable features for the images in the teaching set. For example, for the bird classification task, each $f_j \in \mathcal{F}$ could be an (attribute, value) pair that is visually meaningful, such as ‘blue wing’, ‘flat bill’, etc. In other words, each image $x_i \in \mathcal{X}$ is a realization of \mathcal{F} . The label set $\mathcal{Y} = \{y_1, \dots, y_c\}$ denotes all possible label classes. Each $x_i \in \mathcal{X}$ is associated with a class label $y_i \in \mathcal{Y}$, along with a set of up to n interpretable features $e_i = \{f_1, \dots, f_n\} \subseteq \mathcal{F}$ that explain why x_i is a member of class y_i . One important thing to note is that for $f_i \in \mathcal{F}$ to be in e_i , f_i must be realized in x_i . For discussion simplicity, we focus on binary classification tasks¹ where $y \in \{-1, 1\}$.

Let \mathcal{H} be a finite set of hypotheses. Each hypothesis $h : \mathcal{X} \mapsto \mathbb{R}^{n+1}$ represents a possible scoring rule of the learner, which maps any given image to a $(n + 1)$ -dimensional vector, where the first n entries $(h^{(1)}(x), \dots, h^{(n)}(x))$ represent its evaluation of the importance of the n interpretable features, and the $(n + 1)$ th entry $h^{(n+1)}(x)$ associates its class label. Intuitively, the importance of features can be viewed as an intermediate label of the image. For a given image x_i , hypothesis h

¹Note that our model can be readily extended to the c -class classification tasks as long as we have access to the prior over the learners’ hypotheses.

will predict the positive label if $\text{sign}(h^{(n+1)}(x)) > 0$. Furthermore, we assume that there exists an optimal hypothesis $h^* \in \mathcal{H}$ that correctly predicts both the *labels* and the *feature importance* of all $x \in \mathcal{X}$.

3.1 Image Only Learner Model

In our setting, we assume that the teacher has access to image set \mathcal{X} , label set \mathcal{Y} , prior hypothesis space \mathcal{H}_0 . During the teaching session, the teacher selectively shows a sequence of images from the teaching set to the learner. In particular, the teacher will select the most informative images so that the learner doesn't need to review all possible images to learn a particular concept. For each image shown, the learner will guess or apply their knowledge learned to identify its true class label. After the learner provides their answer, the teacher will reveal the ground truth to the learner. For this work, we focus on the non-interactive setting, where the teacher doesn't take the learner's response into account at the end of each teaching iteration.

We adopt the stochastic STRICT model of Singla et al. [19] as a basic model to characterize how learners adapt to the images shown by the teacher. The model was originally proposed for teaching with label only feedback. It assumes that learners carry out a random walk in the hypothesis space \mathcal{H} . At the beginning of teaching, the learner randomly pick a hypothesis $h \in \mathcal{H}$ according to the prior distribution $P_0(h)$. After receiving a new image, the learner will stick to the current hypothesis if the ground truth label is consistent with her own prediction; otherwise she randomly switches to a new $h \in \mathcal{H}$ according to $P_t(h)$ which is constructed in a way that reduce the probability of hypotheses that disagree with the true labels in the images that are taught so far

$$P_t(h_j) = \frac{1}{Z_t} P_0(h_j) \prod_{\substack{s=1 \\ y_s \neq \text{sign}(h_j^{(n+1)}(x_s))}}^t P(y_s | h_j, x_s) \quad (1)$$

where Z_t is a normalization factor. As in [19], we model how consistent the prediction of hypothesis h_j is with example (x_s, y_s) as

$$P(y_s | h_j, x_s) = \frac{1}{1 + \exp(-\alpha h_j^{(n+1)}(x_s) y_s)}, \quad (2)$$

where $\alpha > 0$ is a parameter that controls how degree of noise tolerance of the model. As $\alpha \rightarrow \infty$, hypotheses that are inconsistent with the class label are immediately discarded.

3.2 Feature Feedback Learner Model

In the above setting, the teacher limits its teaching power in that it does not fully utilize all available resources. By having access to the prior hypothesis space \mathcal{H}_0 , the teacher also knows the 'importance' each learner puts on each of the features. In other words, the teacher not only knows what each learner will predict for each image, it also knows why they make that decision. Since h^* is also in $P_0(h)$, the teacher knows what the optimal importance should be. In our feature feedback setting we make use of this additional knowledge. During each round of teaching, after the learner reveals their answer the teacher shows not only the ground truth class label but also the explanatory features indicating why the image has that particular ground truth label.

Here, we also assume that learners carry out a random walk in the hypothesis space. However, in addition to label feedback, the learners also incorporate the feature feedback by jumping to another hypothesis when observing any inconsistency according to probability distribution P_t , which is constructed such that the inconsistent hypotheses will have lower probability at the end.

In order to capture how the learners would adapt to feature feedback, we introduce an additional discount factor

$$p(e_{sk} | h_j, x_s) = \exp(-\beta |e_{sk} - h_j^{(k)}(x_s)| * w_k), \quad (3)$$

where e_{sk} is the 'true importance score' for the k^{th} feature selected for image x_s revealed by the teacher, and $h_j^{(k)}(x_s)$ is the importance score h_j assigns to feature k for image x_s . The parameter w_k defines the global importance known by teacher. It is initialized such that predictive features

will have higher weights than noisy features. This is achieved by first initializing the weights to be uniformly distributed and then adding a one to the predictive dimension and finally renormalizing. Similar to the noise parameter α from above, β captures the learner’s ability to adapt to the feature feedback (in other words, β represents the learner’s noise from the teacher’s perspective). At the limit $\beta \rightarrow \infty$ indicates that inconsistent hypotheses are completely removed from the hypothesis space. The posterior P after revealing t images becomes

$$P_t(h_j) = \frac{1}{Z_t} P_0(h_j) \prod_{\substack{s=1 \\ y_s \neq \text{sign}(h_j^{(n+1)}(x_s))}}^t P(y_s | h_j, x_s) \prod_{e_{sk} \in E_s} P(e_{sk} | h_j, x_s), \quad (4)$$

where Z_t is again a normalization constant, which sums over the hypothesis space and E_s is the set of predictive features selected for images $\{x_s\}_{s=1:t}$.

3.3 Teaching Objective

Given our model of the learner, our goal is to select a sequence of ‘explanatory’ teaching images A (including explanations E_A), such that after teaching, the learners are directed towards a distribution over the hypotheses that result in as few mistakes as possible. More concretely, we define the error of a single hypothesis as

$$\text{err}(h_j) = \frac{|x \in \mathcal{X} : \text{sign}(h_j^{(n+1)}(x)) \neq y|}{|\mathcal{X}|}.$$

This is the fraction of images that the hypothesis will incorrectly predict when compared to the ground truth labels. The expected error for the learner given teaching set A is then defined as

$$\mathbb{E}[\text{err}(h) | A] = \sum_{h_j \in \mathcal{H}} P(h_j | A) \text{err}(h_j),$$

and the learner’s posterior after seeing the teaching set A is

$$P(h_j | A) = \frac{1}{Z_A} P_0(h_j) \prod_{\substack{x \in A \\ y \neq \text{sign}(h_j^{(n+1)}(x))}} P(y | h_j, x) \prod_{e_k \in E_A} P(e_k | h_j, x).$$

Here, E_A is the sets of important features selected for examples $x \in A$. Our goal, is to find a set (of teaching images) A^* of minimal size, such that upon observing the labels and explanations for the images in $|A|$, the learner would achieve an expected error rate of at most ϵ . Formally, we want

$$A^* = \underset{A \in \mathcal{X}}{\text{argmin}} |A|, \text{ s.t. } \mathbb{E}[\text{err}(h) | A] \leq \epsilon, \quad (5)$$

where ϵ is a parameter that defines the allowed tolerance on learner error.

3.4 Optimization

Similar to the STRICT policy [19], we propose to adopt a greedy strategy for accommodating feature feedback. Let us define

$$R(A) = \mathbb{E}[\text{err}(h)] - \mathbb{E}[\text{err}(h) | A] = \sum_{h_j \in \mathcal{H}} (P_0(h_j) - P(h_j | S)) \text{err}(h_j),$$

as the reduction in expected error after observing the teaching images in A . Therefore, solving Eq. (5) is equivalent to finding the smallest set A achieving error reduction $\mathbb{E}[\text{err}(h)] - \epsilon$. As shown in [19], the problem of optimizing the original objective is NP-hard. We therefore replace the $R(A)$ with the following surrogate function

$$Q(A) = \sum_{h_j \in \mathcal{H}} (V(h) - V(h | A)) \text{err}(h_j), \quad (6)$$

making use of the unnormalized posterior of the learner

$$V(h_j | A) = P_0(h_j) \prod_{\substack{x \in A \\ y \neq \text{sign}(h_j^{(n+1)}(x))}} P(y | h_j, x) \prod_{e_k \in E_a} P(e_k | h_j, x). \quad (7)$$

In our model, we assume that the importance scores of different features across different images are *conditionally independent* given the underlying hypothesis h (and hence the joint distribution of the feature feedback factorizes). One immediate observation is that the surrogate objective $Q(A)$ is submodular. Hence, greedily optimizing the objective is guaranteed to output a near-optimal sequence of images for the problem of optimizing $Q(A)$ (Eq. (6)). Let $E = \sum_h P(h) \text{err}(h)$ be the expected error probability with respect to the prior distribution. It is easy to show that after receiving the teaching images in A , the learner’s expected error is at most a constant factor of $E - Q(A)$:

$$\mathbb{E}[\text{err}(h) | A] \leq \sum_h \frac{P(h | A)}{P_0(h^* | A)} \text{err}(h) = \sum_h \frac{V(h | A)}{P_0(h^*)} \text{err}(h) = \frac{E - Q(A)}{P_0(h^*)}.$$

Note, that here h^* represents the optimal hypothesis that agrees with the teacher’s feedback on both *labels* and *features* for all images in the teaching set, as opposed to the one that achieves 0 error rate on the labels only. Thus, greedily selecting teaching images according to $Q(A)$ until $Q(A) \geq E - P_0(h^*)\epsilon$ is *sufficient* to provide a solution that achieves expected error probability $\mathbb{E}[\text{err} | A] \leq \epsilon$. Our final selection approach is outlined in Algorithm 1. It follows from [19] that the performance of our greedy strategy (measured by the number of examples taught in the worst case) achieving error ϵ is within a logarithmic factor of the worst-case cost of the optimal algorithm OPT achieving error at least $\frac{P_0(h^*)\epsilon}{2}$. More specifically, the cost of our algorithm is bounded by $\text{OPT}(\frac{P_0(h^*)\epsilon}{2}) \cdot \log\left(\frac{1}{P_0(h^*)\epsilon}\right)$.

Algorithm 1: Teaching with Feature Feedback

- 1 **Input:** images $\mathcal{X} \{(x_i, y_i, e_i)\}_{i=1:m}$, hyps \mathcal{H} , prior P_0 , tolerance ϵ .
 - 2 **Output:** Selected images to teach A .
 - 3 $A \leftarrow \emptyset$;
 - 4 **while** $Q(A) < \mathbb{E}[\text{err}(h)] - P_0(h^*)\epsilon$ **do**
 - 5 | $i^*, k^* \leftarrow \text{argmax}_{i,k} Q(A \cup \{x_i, e_{ik}\})$;
 - 6 | $A \leftarrow A \cup \{x_{i^*}, e_{i^*k^*}\}$
-

4 Experiments

We perform two sets of experiments using 1) simulated learners and 2) real human participants from Amazon’s Mechanical Turk to validate the performance of our algorithm. Performance is measured using both learner classification error on the test set and the amount of time it takes the learners to complete the test set. Here, time acts as a proxy for measuring how confident the learner is in answering the test questions after completing the teaching phase, where faster is better. We compare our method against three baselines: 1) *random image* - random selection of each teaching image with image-level feedback i.e. only giving the class label, 2) *random feature* - random selection of image and feedback for a randomly selected interpretable feature i.e. class label with random feature highlighted, and 3) STRICT- the submodular image selection approach of [19] with image-level feedback.

4.1 Simulated Learners

For our first experiment we use the Breast Cancer dataset of [20] from the UCI Machine Learning repository. It contains a total of 569 examples from two classes with 30 positive real valued feature dimensions. We generate 200 linear hypotheses by randomly sampling pairs of points from the dataset and adding the hyperplane that bisects them. This ensures that the hypotheses span the space of the dataset. We do not use the ground truth class labels and instead pick a random hypothesis from the hypotheses set as the optimal hypothesis and use its predictions as the ground truth labels. We

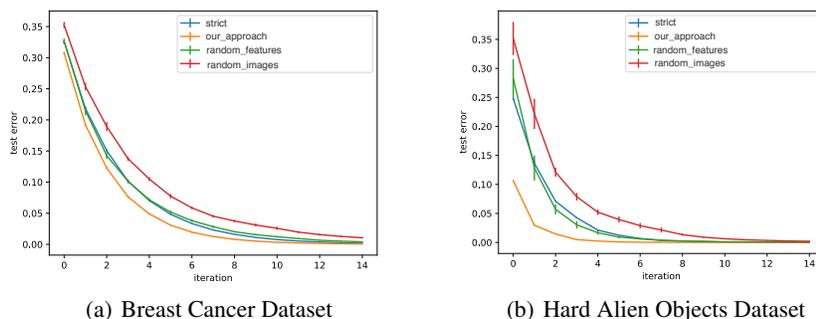


Figure 2: Test error for simulated learners. Our method performs best overall.

average ten trials with different train and test splits, different random prior hypothesis spaces, and different optimal hypotheses. For all experiments in the paper we use the same noise parameters with $\alpha = 1$ and $\beta = 1$.

4.1.1 Simulation Experiment Results

In Fig. 2 (a), we see that our approach outperforms the baselines converging to a lower test error faster. This implies that our approach selects more informative images and features to show to the learner.

4.2 Human Learners

For our experiments featuring real human participants, we generated two datasets of hypothetical alien objects coming from either ‘Jupiter’ or ‘Mars’. This simulated data gives us full control over the generating distribution and hypothesis space. Both datasets contain a total of 128 images evenly distributed between the two classes. One of the datasets is more challenging than the other as it requires a more complicated decision rule to separate the data. Example images from both datasets can be seen in Figs. 3 and 4.

For both datasets, each of the images has a fixed number of parts, where each part has up to two attributes such as color, size, and shape that vary from image to image. Features are represented in binary format, indicating the presence or absence of a particular dimension. Each dimension of our ten dimensional binary feature vector corresponds to a particular attribute of a part e.g. ‘red head’, ‘square neck’, etc. By altering the set of features that form the ground truth classification rule for a given class we can control the difficulty level of the learning task for the student. The easy dataset has six different visual parts in total and the hard dataset has eight. The predictive features for easy dataset also occupy larger area in the image and thus are more visible to learners. The ground truth decision rule for the hard dataset is more challenging to learn.

The hypothesis space \mathcal{H} consists of a set of linear functions $h(x) = w^T x$ with $w \in \{0, 1\}^d$, that place a weight of $\{0, 1\}$ on any given feature. The interpretation of $w_k = 0$ is that learners ignore the k^{th} feature when making their classification decision. Similarly, $w_k = 1$ implies that learners perceive the k^{th} feature as an indicator that the example belongs to the positive class. We generate 200 prior hypotheses by randomly sampling from all hypotheses that have non zero weights on at most four features. The optimal hypothesis is determined by design and is described in the captions in Figs. 3 and 4 where $w^* = 1$ for predictive features associated with the positive class and $w^* = 0$ everywhere else. We inject the optimal hypothesis into the prior hypothesis space to ensure that the learner can reach optimal performance i.e. the task is realizable.

Results with simulated learners for the hard dataset can be seen in Fig. 2 (b).

4.2.1 Experimental Setup

Experiments were performed on Amazon’s Mechanical Turk. Participants were randomly assigned one of the four different teaching strategies, where each strategy received on the order of 30 par-

ticipants. Participants were shown 10 training images for the ‘easy’ task and 15 for the ‘hard’ task during teaching and 12 images at test time for both. In order to motivate the participants, we paid a bonus to the top 10% of participants based on their performance at test set.

Experiments were conducted using the same protocol as [10]. Learners were shown a sequence of images during the ‘teaching’ phase and after each image asked to estimate the correct class label for the image they just saw. After estimating the class label for each image, they were given feedback in the form of the ground truth class label. For the teaching strategies that utilize feature feedback they were also shown a single interpretable feature, displayed as an arrow pointing to a single feature per image. Teaching was then followed by a ‘testing’ phase, similar to teaching, except no feedback was provided to the learners. We used the same randomly selected testing images when comparing all methods, where images from the test set were not contained in the teaching set.

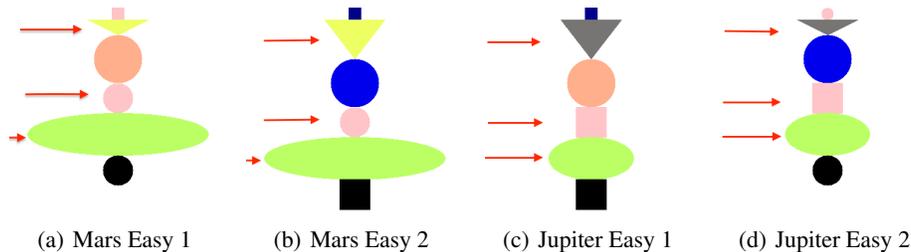


Figure 3: Sample images from the ‘easy’ alien objects dataset. Here, arrows indicate the predictive features. The ground truth decision rule is that Mars should have a yellow color, pink circle, and large size in the three locations highlighted by arrows, from top to bottom. Any other combination of features means the object is from Jupiter.

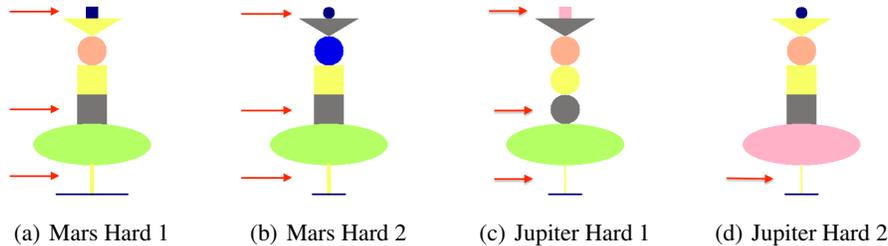
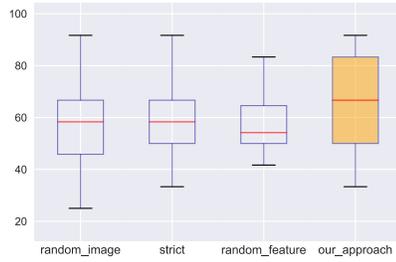


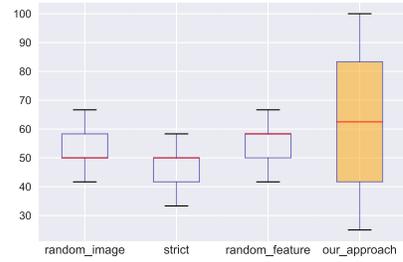
Figure 4: Sample images from the ‘hard’ alien objects dataset. Again, the arrows indicate the predictive features. The ground truth rule is that Mars must have blue color, grey square shape, and thick size on the three dimensions highlighted by the arrow, from top to bottom. Not having this combination of parts and attributes means the objects are from Jupiter.

4.2.2 Human Experiment Results

In Fig. 5 we see the median performance at test time for the four different teaching strategies on both alien objects datasets. Our approach performs best overall but has higher variance than some of the other baselines. In Fig. 6 (a) and (b) we can see that this variance is explained by two distinct modes in the test time performance histograms. Unlike the other baselines which have a peak around 50% accuracy i.e. close to random guessing from the noisy participants, our method is clearly capable of teaching some of the learners as there is a greater number of learners that achieve higher performance. In Fig. 7, we observe that participants also answer questions faster on average at test time. This implies that learners are more confident with the decision rule learned during teaching phase. Overall, the result suggests that our method does a better job at selecting teaching examples and providing feedback as the learners are able to answer the test questions both faster and more accurately.



(a) Easy alien objects dataset

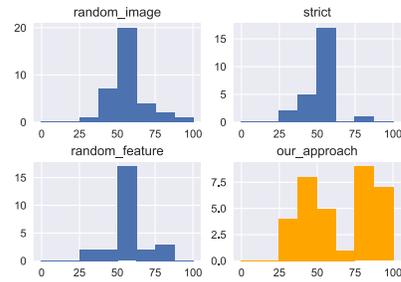


(b) Hard alien objects dataset

Figure 5: Test time accuracy on the the two alien objects datasets for real human participants, where higher values are better. Workers perform better on average when taught with our approach.

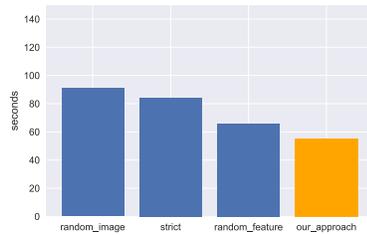


(a) Easy alien objects dataset

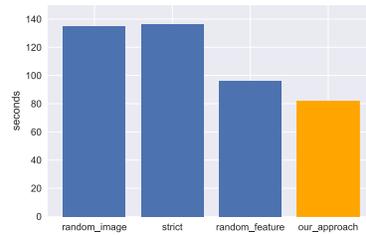


(b) Hard alien objects dataset

Figure 6: Histograms of learner test time accuracy on the two alien objects datasets.



(a) Easy alien objects dataset



(b) Hard alien objects dataset

Figure 7: Average time it takes the participants to complete the testing phase for both datasets. As expected, it takes the learners more time on average to complete the hard dataset in (b) compared to the easy one in (a). We see in both cases that our approach results in workers that are able to answer test questions faster.

5 Conclusion

We presented a method for the teaching of visual categories to human learners with interpretable, feature-level, feedback. Our experiments show that teaching with interpretable feedback generates more informative teaching sequences, resulting in faster learning. Our approach assumes that we have access to an interpretable feature space for teaching. In future, automatically discovering these informative features from weakly labeled datasets would allow us to significantly reduce the amount of annotation required to teach new concepts.

Acknowledgments

The authors thank Google for supporting the Visipedia project, and kind donations from Northrop Grumman, Bloomberg, and AWS Research Credits. Yuxin Chen was supported in part by a Swiss NSF Mobility Postdoctoral Fellowship.

References

- [1] Ji Hyun Bak, Jung Yoon Choi, Athena Akrami, Ilana Witten, and Jonathan W Pillow. Adaptive optimal training of animal behavior. In *NIPS*, 2016. 1
- [2] Sumit Basu and Janara Christensen. Teaching classification boundaries to humans. In *AAAI*, 2013. 1
- [3] Thomas Berg and Peter N Belhumeur. How do you tell a blackbird from a crow? In *ICCV*, 2013. 2
- [4] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 1994. 1
- [5] Jia Deng, Jonathan Krause, Michael Stark, and Li Fei-Fei. Leveraging the wisdom of the crowd for fine-grained recognition. *PAMI*, 2016. 2
- [6] Baxter S Eaves Jr, April M Schweinhart, and Patrick Shafto. Tractable bayesian teaching. 2015. 1
- [7] Sally A Goldman and Michael J Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 1995. 1
- [8] Elizabeth R Grant and Michael J Spivey. Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 2003. 2
- [9] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *ECCV*, 2016. 2
- [10] Edward Johns, Oisín Mac Aodha, and Gabriel J Brostow. Becoming the expert-interactive multi-class machine teaching. In *CVPR*, 2015. 1, 7
- [11] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 2016. 2
- [12] Weiyang Liu, Bo Dai, James M Rehg, and Le Song. Iterative machine teaching. *ICML*, 2017. 1
- [13] RT Peterson. A field guide to the birds: a completely new guide to all the birds of eastern and central north america. vol. 1., 1998. 2
- [14] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *NIPS*, 2015. 1
- [15] Stefanos Poulis and Sanjoy Dasgupta. Learning with Feature Feedback: from Theory to Practice. In *AISTATS*, 2017. 2
- [16] Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. Faster teaching by pomdp planning. In *AIED*, 2011. 1
- [17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 2016. 2
- [18] Brett Roads, Michael C Mozer, and Thomas A Busey. Using highlighting to train attentional expertise. *PloS one*, 2016. 2
- [19] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014. 1, 3, 4, 5
- [20] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. 1992. 5
- [21] Jacob Whitehill and Javier Movellan. Approximately optimal teaching of approximately optimal learners. *Transactions on Learning Technologies*, 2017. 1
- [22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2
- [23] Xiaojin Zhu. Machine teaching for bayesian learners in the exponential family. In *NIPS*, 2013. 1