Becoming the Expert - Interactive Multi-Class Machine Teaching

Edward Johns Oisin Mac Aodha Gabriel J. Brostow
University College London
http://visual.cs.ucl.ac.uk/pubs/interactiveMachineTeaching

Abstract

Compared to machines, humans are extremely good at classifying images into categories, especially when they possess prior knowledge of the categories at hand. If this prior information is not available, supervision in the form of teaching images is required. To learn categories more quickly, people should see important and representative images first, followed by less important images later – or not at all. However, image-importance is individual-specific, i.e. a teaching image is important to a student if it changes their overall ability to discriminate between classes. Further, students keep learning, so while image-importance depends on their current knowledge, it also varies with time.

In this work we propose an Interactive Machine Teaching algorithm that enables a computer to teach challenging visual concepts to a human. Our adaptive algorithm chooses, online, which labeled images from a teaching set should be shown to the student as they learn. We show that a teaching strategy that probabilistically models the student's ability and progress, based on their correct and incorrect answers, produces better 'experts'. We present results using real human participants across several varied and challenging real-world datasets.

1. Introduction

Large, manually annotated image datasets have contributed to recent performance increases in core computer vision problems such as object detection and classification [16, 37, 25]. In cases where the visual categories of interest are generic everyday objects, annotation can be completed by crowd sourcing labels from the internet using services such as Mechanical Turk [1]. A typical image labeling task begins with a set of instructions to the annotator, showing them example images from the classes of interest. The annotator is then asked to assign class labels to new images where the ground truth is unknown.

But what happens if the annotator is unsure? This is a real problem when annotators are incorrectly assumed to have prior knowledge of the classes of interest from which they can generalize, either from everyday life or from specialized training. For many problems, highly specialized, domain specific knowledge, acquired through extensive training, is needed before someone can differentiate between potentially multiple, highly self-similar object categories.

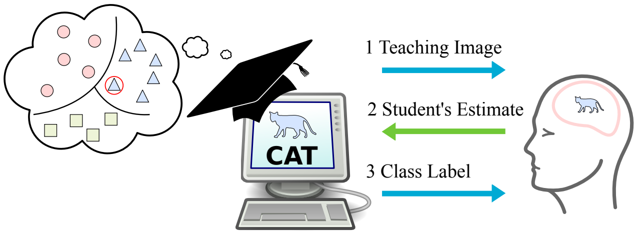

Figure 1: In Interactive Machine Teaching the computer teaches the human learner, one image at a time. It begins by showing them an image from the larger labeled image dataset while concealing the true class label. The learner/student responds with their estimate of the image's class. The teacher then updates their model of the student, and finally reveals the correct answer to them. This process is repeated with further images until teaching ends.

Designing the set of teaching images or 'teaching set' to show the annotators is challenging, because each annotator will have a different degree of expertise. While it is possible to model the uncertainty and noise generated from groups of annotators to improve their collective performance [34, 40, 27], these approaches tend to downweight votes from weak annotators by learning to trust the experts. In this paper, we pose the question - how does one become an expert? We posit that a human's discriminative ability for a given visual classification task can be improved by better modeling the teaching process required to make them experts.

The family of methods referred to as Machine Teaching offers a general solution to the problem of teaching humans [43, 18, 42, 39, 33]. Machine Teaching is not the same as Active Learning [38]. In Active Learning, the computer's goal is to learn more accurate models given the smallest amount of supervision. This is achieved by carefully selecting only the most informative datapoints to be labeled by the human. In Machine Teaching, the computer, rather than the human, is the (perfect) oracle and is tasked with delivering a teaching set to the human student to help them learn the given task more effectively. Teaching the human a new skill is useful in its own right. Further, they are now better positioned to accurately annotate the additional unlableled data outside the original teaching set. Automatic teaching algorithms have applications in many domains from education, language learning, medical image analysis, biological species identification [30], and more. Crucially, for automated teaching to be effective, it needs to be able to assess the student's current knowledge, and have a mechanism for selecting teaching examples to best improve this knowledge.

In this work, we focus on the task of image classification. Here, it is not possible for the teacher to directly 'teach' the high-dimensional decision boundary to the human learner, so instead, the student must learn this boundary by being shown teaching images. Our goal during teaching is to choose teaching images that will maximize the student's classification ability in the minimum amount of teaching time. Unlike computers, humans have both limited and imperfect memory for instance-level recognition, especially during the initial learning of a task [17]. However, humans have the advantage of possessing the ability to generalize to unknown examples and perform domain adaptation given only few instances. The majority of previous work in Machine Teaching has focused on non-interactive teaching, where one teaching set is computed offline, independent of feedback from each student [39, 33]. In this work, we address the under-explored problem of interactive teaching [14, 4]. Here, the teacher can adapt their teaching set online, based on the current performance of the individual student (see Figure 1).

We propose an algorithm that interactively teaches multiple visual categories to human learners. Our contributions are threefold: 1) Unlike computers, humans are not optimal learners. Our algorithm models student ability online, resulting in teaching sets that are adapted to each individual student. We make no assumptions regarding the internal learning model used by the students. Instead, we present them with teaching images that attempt to reduce their predicted future uncertainty based on an estimate of their current knowledge. 2) Our teaching algorithm reduces the amount of time it takes students to learn categorization tasks involving multiple classes. Experimentally we show that real human participants, using our algorithm, perform better than other baselines on several challenging datasets. 3) Finally, we provide a web based interface and framework for exploring new teaching strategies. Our intention is that this will encourage the development of new and diverse teaching strategies for a variety of human visual learning tasks.

2. Related Work

Here we cover the most closely related work in Machine Teaching. As we are concerned with the task of image categorization, we focus on research concerning teaching classification functions. However, it is worth noting that different types of teaching tasks have been explored in the literature, *e.g*. sequential decision tasks [8]. How humans acquire and represent categories is an active area of research in visual psychology. Many candidate models for category acquisition and representation in humans exist, and for an overview we direct the readers to [28, 35]. In this work, our goal is not to model these internal processes directly, but to instead treat the human as a stochastic black box learner. For convenience, we divide the related work in Machine Teaching into two areas - batch (fixed) teaching and interactive (adaptive or online) teaching. For a recent, and general, introduction to Machine Teaching, please see [43].

Machine Teaching - Batch (Fixed)

In batch-based teaching, the teacher's goal is to construct an optimal set of teaching examples offline, which are then presented to the student during teaching. Early work in this area focused on the theoretical analysis of the teaching dimension [18]. The teaching dimension is defined as the minimum number of examples required from a given concept to teach the concept to a student. Like many other works in teaching, [18] makes the simplifying assumption that the student has perfect memory (*i.e*. once shown an example the student will remember it in the future) - an assumption that is violated in real world teaching. Other theoretically motivated works, while interesting, provide little validation on real human subjects [3, 13, 46].

More recently, Zhu [42] attempted to minimize the joint effort of the teacher and the loss of the student by optimizing directly over the teaching set. The proposed noise-tolerant model assumes that the student's learning model is known to the teacher, and that it is in the exponential family. In follow-on work, Patil *et al*. [33] maintain that unlike computers, which have infinite memory capabilities, humans are limited in their retrieval capacity. Motivated by real human studies [17], they show that modeling this limited capacity improves human learning performance on tasks involving simple one-dimensional stimuli.

Most related to our work, Singla *et al*. [39] teach binary visual concepts by showing images to real human learners. Their method operates offline and tries to find the set of teaching examples that best conveys a known linear classification boundary. Experiments with Mechanical Turkers show an improvement compared to other baselines, including random sampling. Their approach attempts to encode some noise tolerance into the teaching set, but is still unable to adapt to a student's responses online during teaching, because the ordering of the teaching images is fixed offline.

Machine Teaching - Interactive (Adaptive)

Real human students are often noisy, especially in the early stages of learning when the concepts to be learned are not formed in their minds. Additionally, students do not all learn at the same rate - concepts that are difficult for some students may be easier for others. In Interactive Teaching (Figure 1), the teacher receives feedback from the student as teaching progresses. Given this feedback, the teaching strategy can adapt to the current ability of an individual student over time.

Using a probabilistic model of the student and a noise-free learning assumption, Du and Ling [14] propose a teaching strategy called 'worst predicted'. This strategy is similar to uncertainty sampling, which is commonly found in Active Learning [38]. However, unlike Active Learning, in Machine Teaching the teacher has access to the ground truth class labels and can use this to assess the student's performance during teaching. Experimentally, we show that their strategy performs sub-optimally as it only seeks to show the student the image that they are currently most uncertain about, without regard for how informative that image may be in relation to others. As a result, it is very susceptible to teaching outliers, *i.e.* unrepresentative images at the fringes of the teaching set.

In one of the few interactive teaching papers that deal with visual concepts, Basu and Christensen [4] evaluate human learning performance in binary classification using three different teaching methods. Students were tasked with classifying simple synthetically generated (and linearly separable) depictions of mushrooms into one of two categories. They do not explicitly model labeling noise from the student, but instead investigate different interface designs and feature space exploration methods to help teach the students.

In this paper, we address the problem of interactive multi-class teaching with real images by directly modeling the student's ability as they provide feedback during teaching.

3. Machine Teaching

In this section we formally define our Machine Teaching task. Our teacher-computer has access to a labeled dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ where each $\mathbf{x}_i$ is an M dimensional feature vector encoding an image I_i, and $y_i \in \{1, ..., C\}$ is its corresponding class label. The teacher's goal is to 'teach' the classification task to the human learner by showing them images from the dataset $\mathcal{D}$. We refer to these teaching images as the 'teaching set', $\mathcal{D}_t$, a subset of images from $\mathcal{D}$ where $|\mathcal{D}_t| \ll |\mathcal{D}|$. In each round of interactive teaching, the teacher first selects an image represented by the feature vector $\mathbf{x}_t$ to show to the human learner. The teacher displays images to the students as it is not possible to directly show them the high dimensional feature vector $\mathbf{x}_t$. The selection of the image to show is based on a process we refer to as the 'teaching strategy', $\mathcal{S}$, used by the teacher. First, the teacher only shows the image and does not yet display the ground truth class label. By not revealing the class label, the teacher is able to ask the student to state which class they believe the image belongs to. After receiving the student's response, the teacher then updates its model of the student, and then reveals the ground truth label. Teaching proceeds for a set number of teaching rounds, and during each iteration, the teacher acquires a better understanding of the student's current ability. Figure 1 outlines one teaching iteration.

With access to ground truth, the teacher trivially knows the conditional distribution $P(y_i \,|\, \mathbf{x}_i)$ for each datapoint $\mathbf{x}_i$. The student learner has a corresponding distribution $P_l(y \,|\, \mathbf{x})$, based only on training examples they have seen so far. During teaching, the teacher seeks to minimize the student's expected loss

$$\mathbb{E}_{\mathbf{x}} = L\left(P(y \,|\, \mathbf{x}), P_l(y \,|\, \mathbf{x})\right), \tag{1}$$

over the dataset, where $L()$ is an appropriate classification loss function. However, the teacher has no way of directly observing the student's true class conditional distribution, $P_l(y \,|\, \mathbf{x})$, so instead must approximate it as $\hat{P}_l(y \,|\, \mathbf{x})$. In this paper, we represent $\hat{P}_l(y \,|\, \mathbf{x})$ using a probabilistic, semi-supervised, classifier.

3.1. Teaching Strategies

The optimal teaching strategy is the one that minimizes the student's expected loss from Equation (1). A simple strategy for choosing the next teaching image is to randomly sample from the dataset $\mathcal{D}$. Random sampling ($\mathcal{S}_{rnd}$) does not model the student and is therefore unable to adapt to their ability. This lack of adaptation can manifest itself in two ways - 1) redundantly presenting teaching examples of concepts that have already been learned by the student, and 2) not directly reinforcing concepts that the student has shown themselves (through feedback) to be uncertain about.

Du and Ling [14] proposed a strategy called 'worst predicted', here $\mathcal{S}_{wp}$, which is related to uncertainty sampling commonly used in Active Learning [38]. However, unlike in Active Learning, in Machine Teaching, the computer does have access to the ground truth labels. Their strategy selects the next teaching image as the one whose prediction deviates most from the ground truth,

$$\mathbf{x}_t = \operatorname*{argmin}_{\mathbf{x}} \hat{P}_l(\bar{y} \,|\, \mathbf{x}), \tag{2}$$

where $\bar{y} = \mathrm{argmax}_y \, P(y|\mathbf{x})$ is the ground truth class label known to the teacher. The disadvantage of this approach is that it is prone to proposing outliers as teaching images, as they tend to be highly uncertain under the current model.

One potential solution to this problem is to weight the datapoints by some measure of local density in the feature space *e.g.* [38, 15].

3.1.1 Expected Error Reduction Teaching

Our teaching strategy, which we refer to as $\mathcal{S}_{eer}$, takes inspiration from optimal sampling methods found in Active Learning [36, 45, 29]. Unlike $\mathcal{S}_{wp}$, $\mathcal{S}_{eer}$ chooses the teaching image which, if labeled correctly, would have the greatest reduction on the future error over the images that are not in the teaching set, $\mathcal{D}_u = \mathcal{D} \setminus \mathcal{D}_t$, where

$$\mathbf{x}_t = \operatorname*{argmin}_{\mathbf{x}_p} \sum_{\mathbf{x}_i, \bar{y}_i \in \mathcal{D}_u} (1 - \hat{P}_l^{+(\mathbf{x}_p, \bar{y}_p)}(\bar{y}_i \,|\, \mathbf{x}_i)). \quad (3)$$

Here, $\hat{P}_l^{+(\mathbf{x}_p, \bar{y}_p)}$ is the updated estimate of the student's conditional distribution if they were shown $\mathbf{x}_p$ and in turn labeled it correctly. This strategy has the advantageous property that it first concentrates on regions of high density in the feature space, and as the student improves, refines the boundaries between these regions. In the context of Active Learning, this is referred to as the exploration versus exploitation trade off. This is related to the approach to learning advocated by curriculum learning, which focuses on easy concepts first and progressively increases the difficulty [5].

3.2. Modeling the Student

In this work we approximate the student's conditional distribution given the teaching set, $\hat{P}_l(y \,|\mathbf{x}, \mathcal{D}_t)$, using graph based semi-supervised learning [41, 44]. Using the Gaussian Random Field (GRF) semi-supervised method of [44], we can propagate the student's estimate of the class labels for the current teaching set, $\mathcal{D}_t$, to the unobserved images $\mathcal{D}_u$ by defining a similarity matrix $W \in \mathbb{R}^{N \times N}$. The benefit of using a graph based approach is that we do not need to work directly in feature space, and can instead use the similarity, w_{ij}, between image pairs. This gives us the flexibility of allowing similarity to be defined using feature vectors extracted from the images, human provided attributes, or using distance metric learning [23].

If we are given a feature representation for our teaching set, one common approach for computing the similarity w_{ij} between two images uses an RBF kernel

$$w_{ij} = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2), \quad (4)$$

where γ is a length scale parameter that controls how much neighboring images influence each other. Using matrix notation of [44], we define an $N \times C$ matrix $F = \hat{P}_l(y \,|\, \mathbf{x}, \mathcal{D}_t)$, where each element $f_{ic} = \hat{P}_l(y_i = c \,|\, \mathbf{x}_i)$. We can propagate information from the labels provided by the student for the teaching set, encoded as a $|\mathcal{D}_t| \times C$ matrix F_t, to the unlabeled images F_u,

$$F_u = (S_{uu} - W_{uu})^{-1} W_{ut} F_t, \quad (5)$$

where S is a diagonal matrix with entries $s_{ii} = \sum_j w_{ij}$. All entries in F_t are 0, except where the human learner has estimated (correctly or incorrectly) the class label c for teaching image $\mathbf{x}_i$, which we set to $f_{ic} = 1$. W_{uu} is the similarity matrix for the unobserved images, a subset of the full matrix W. As in [44], we can efficiently evaluate Equation (3) using standard matrix operations for datasets featuring 2000 images in under one second using unoptimized Python code.

4. Experiments

To validate our proposed multi-class teaching strategy, we performed studies on real human subjects. Participants were recruited through Mechanical Turk [1], and interacted with our system remotely using our custom made web interface, built using the Python-based Django web framework [2].

4.1. Data

For our experiments, we selected four different datasets, summarized in Table 1. To ensure that the teaching tasks were challenging to participants and one-shot learning was not possible, we chose datasets with small inter-class variation and large intra-class variation. Example images from each of the classes are presented in Figure 2. Unlike standard classification datasets featuring everyday objects *e.g.* [19, 16], our datasets contain image categories that are challenging for non-domain experts to discriminate between, as they are made up of uncommon classes.

Two of the datasets, 'Butterflies' and 'Seabed', were collated by the authors of this paper from ongoing scientific studies into visual species identification. 'Butterflies' is a subset of a larger collection of British butterfly images from a museum collection captured over a period of 100 years. 'Seabed' is a set of images of underwater species taken from a study attempting to measure the effects of trawling on underwater bio-diversity. Both datasets were curated and annotated by domain experts.

Dataset	# Classes	# Images per Class	Origin
Chinese	3	237-240	[26]
Butterflies	5	300	-
Seabed	4	100	-
Leaves	4	102-150	[24]

Table 1: Summary of the datasets used, showing the number of classes and the minimum and maximum number of images per class.

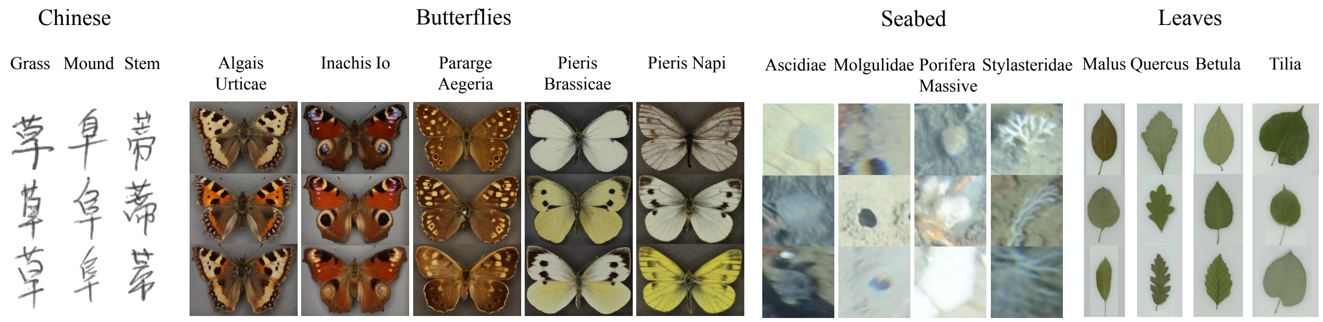

Figure 2: Example images from the four datasets used in our experiments. Each column shows three random images per class. Note, that these images are challenging to categorize as they exhibit a large amount of intra-class variation. Additionally, the ‘Seabed’ images are particularly difficult as they were captured ‘in the wild’ and contain occlusion and clutter.

Image features were extracted using the publicly-available ConvNet system of [21]. For each dataset, we computed features using a network pre-trained on the ImageNet 2012 challenge dataset [37]. We then fine-tuned the fully connected layers using the known ground truth class labels for each of our datasets, which produced a separate ConvNet for each dataset. To construct the similarity matrix W of (4), we reduced the dimensionality of the ConvNet features from 4096 to 50 using PCA, and set the length scale parameter γ to 0.025 for all datasets. In our initial experiments, we explored custom-designed HoG-based features [10] which we found to perform worse compared to our fine-tuned ConvNet. Here, the additional supervised information provided during fine-tuning produces a representation where images from the same class are more smoothly distributed in feature space. A feature space that is better aligned with the student’s view of similarity should benefit all probabilistic strategies equally. It would also be possible to compute similarity between teaching images by crowd-sourcing image rankings from a set of users, *e.g*. [32]. However, we found our ConvNet features to be a good balance between reducing the amount of additional supervised information required for each teaching task and real students’ performance. Code and data are available on our project website.

4.2. Experimental Design

To evaluate our teaching algorithm, we conducted experiments on participants recruited through Mechanical Turk [1]. Previously, Crump *et al*. [9] have shown that it is possible to replicate results from classic category learning experiments using Mechanical Turk. Using a similar experimental setup to [31], our participants were first presented with a sequence of teaching images, which were then followed by a sequence of testing images.

For each experiment, participants were first told how many classes they were being asked to learn. Then teaching commenced using the interactive teaching loop illustrated in Figure 1. For each teaching image, participants were first shown the image, asked to estimate its class label by clicking on the corresponding button in our web interface, and then provided with the correct answer. After receiving the estimated class label from the participant, the teaching strategy updates its model of the student and chooses the next image to be shown. In contrast to the teaching phase, no corrective feedback in the form of the true class labels was provided in the testing phase. The testing round was only used for evaluation purposes and is not necessary in real teaching scenarios. Test images were randomly chosen for each participant, with an equal from each class, and were excluded from the possible teaching set.

Each participant was presented with a random dataset from Table 1, combined with a random teaching strategy. For each dataset, the number of teaching images shown was set to three times the number of classes, and ten times the number for testing. In this way, the lengths of the teaching and testing rounds were proportional to the complexity of the task. We experimented with longer teaching rounds ($>$ 40 images) and testing at regular intervals between teaching images to achieve a learning curve. However, we found through feedback that students became bored and frustrated with the enforced delay, encouraging them to drop out.

It is worth noting that our teaching tasks are significantly more difficult than most crowd-sourced image annotation tasks. In typical annotation tasks, workers already possess strong prior knowledge of the concepts involved, whereas in our teaching tasks, the participants were unlikely to have prior domain expertise. We surveyed participants at the start of the task to ensure that they possessed no prior task knowledge and we rejected results for those who claimed to have even moderate familiarity of any of the classes. As such, the student’s answer to the first teaching image was always a random guess. To avoid workers who were seemingly clicking at random, we also rejected results for those whose average response time per image was too fast ($<$ 3 seconds) during testing. To encourage a conscientious effort in learning, we paid workers a bonus if they scored higher than a threshold during testing. After discarding noisy participants, we collected results from between 25 and 35 participants per strategy/dataset combination.

4.3. Baseline Strategies

In addition to the baseline teaching strategies outlined in Section 3.1, we also compared to two other baselines $\mathcal{S}_{cc}$ and $\mathcal{S}_{batch}$. For $\mathcal{S}_{cc}$, or class centroids, we computed the feature space centroids for each class for a given dataset, and students were only presented with the images represented by these centroids during teaching. Teaching images were selected by randomly choosing from one of these centroids. If there was little intra-class variation, if one-shot learning was possible, or if the classes were familiar to the student, we would expect this baseline to perform very well. The final baseline, $\mathcal{S}_{batch}$, is similar to offline batch teaching algorithms such as [39]. Here, the ordering of the teaching images was computed offline. We computed the ordering using the $\mathcal{S}_{eer}$ algorithm, but assuming that if shown an image, the student would always label it correctly. Given this assumption, the selection of teaching images is deterministic and is identical for all students regardless of their responses. Recent strategies for offline binary teaching, such as [39], are not directly applicable for comparison because we operate in the challenging interactive multi-class classification scenario.

4.4. Human Experiments

Results from human participants are summarized in Table 2. Results for individual datasets are depicted in Figure 3, where the average number of testing images answered correctly are shown for each dataset and strategy combination. We can see that our $\mathcal{S}_{eer}$ method outperforms the other teaching strategies on the ‘Chinese’, ‘Butterflies’, and ‘Seabed’ datasets. In these three, our method is consistently the best performing, while the other methods vary in performance depending on the specific dataset. As we can see from Table 2, there is no clear ‘second-best’ method, and the offline $\mathcal{S}_{batch}$ and uncertainty $\mathcal{S}_{wp}$ strategies are often outperformed by random $\mathcal{S}_{rnd}$. $\mathcal{S}_{eer}$’s performance is most pronounced on the ‘Seabed’ dataset, which also contains the most haphazard images, due to the acquisition of data from cameras ‘in the wild’, as opposed to neatly-framed imaging in controlled laboratory conditions.

Average timings during testing for the different strategies, calculated as the time between being shown a test image and submitting an answer, are presented in Table 2. Participants taught using our method tend to answer more quickly compared to the other strategies. $\mathcal{S}_{rnd}$ and $\mathcal{S}_{cc}$ also have low response times, but students’ poorer performance at test time possibly indicates a level of false-confidence.

Strategy		Ave. Time (ms)	Ave. Score
Random	$\mathcal{S}_{rnd}$	4876	0.67
Centroids	$\mathcal{S}_{cc}$	4706	0.58
Worst Pred.	$\mathcal{S}_{wp}$	5237	0.66
Batch	$\mathcal{S}_{batch}$	6216	0.64
EER (Ours)	$\mathcal{S}_{eer}$	**4659**	**0.73**

Table 2: Average participant response times during testing, and test set scores across all datasets.

Table 3 provides p-values for the statistical significance of our results. Two-tailed tests were conducted with a null hypothesis that the distributions of scores for our method across all datasets, and the competing method, are statistically similar, based on a Gaussian assumption. The p-values obtained are well within the standard measure of 0.05 for testing statistical significance, indicating that our results are not due to chance.

Figure 4 shows the average learning curves for the five teaching strategies obtained during teaching. The average score for each 10% progress interval (through the training set) is calculated by averaging the number of correct responses over all students and datasets at that point along the teaching phase. Note that this is not equivalent to the true learning curve, as images are chosen to actively teach the student, rather than to assess a snapshot of their performance. We see a general trend of improving recognition rates with further teaching images. However, $\mathcal{S}_{cc}$ gives a false sense of performance because the same centroid images are repeatedly shown, thus the student overfits to these images and typically fails to generalize during testing. Unlike the others, the uncertainty based $\mathcal{S}_{wp}$ strategy has a relatively flat learning curve, because the outlier images shown are challenging to learn. This underfitting gives students only a weak understanding of each class’s variability.

Figure 5 shows examples of the teaching images shown to students for each of the five strategies with the ’Chinese’ dataset. We see the capacity of $\mathcal{S}_{eer}$ to adapt to incorrect responses, where attention is given to the ’Stem’ class due to an incorrect previous answer, before returning to teach ’Grass’ due to its previous incorrect answer, and finally exploring the student’s understanding of ’Mound’. On the other hand, $\mathcal{S}_{batch}$ is unable to adapt its teaching set and focuses on teaching ’Mound’ and ’Stem’ despite the student’s

Strategy		P-value
Random	$\mathcal{S}_{rnd}$	0.0138
Centroids	$\mathcal{S}_{cc}$	< 0.0001
Worst Pred.	$\mathcal{S}_{wp}$	0.0027
Batch	$\mathcal{S}_{batch}$	< 0.0001

Table 3: Two-tailed p-values for hypothesis tests on the statistical significance of our method compared to all others.

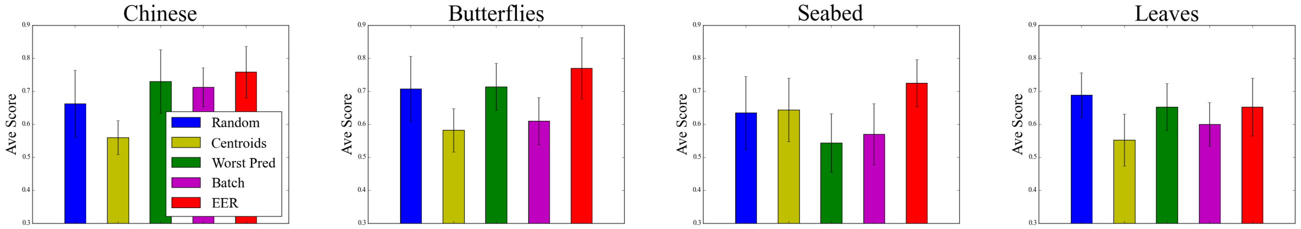

Figure 3: Human experiment results across the four datasets described in Table 1, showing the average scores after the testing phase across all participants. Human participants on Mechanical Turk using our Expected Error Reduction based teaching strategy (here EER) tend to have better recognition performance on average after teaching, compared to the other baselines.

poor performance with 'Grass'. $\mathcal{S}_{wp}$ begins by displaying reasonable examples, but ends by attempting to teach very unusual examples which are not representative of the dataset's distribution.

The performance of $\mathcal{S}_{eer}$ on the 'Leaves' dataset shows an example where we do not perform better than the random baseline, but come joint second. A property unique to this dataset is the multi-modal nature of the leaves present, where each class in fact represents an entire genus, composed of a number of different species that do not all look the same. We found that human learners typically assumed unimodal distributions during teaching and would often focus on only a single species within the entire genus.

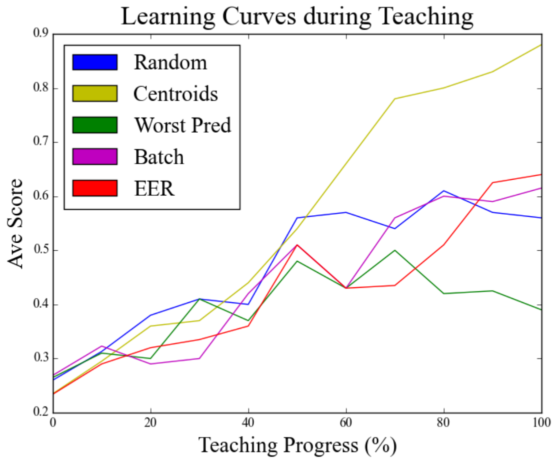

Figure 4: Average learning curves across all students and datasets during the teaching phase.

4.5. Limitations

Currently, our model does not attempt to directly recover from incorrect responses made by students in the past. If a student has previously given an incorrect answer, future teaching images will be selected from similar regions in the feature space. However, earlier incorrectly labeled images will still influence the label propagation. Allowing incorrectly labeled images to be relabeled could result in the teaching strategy continually presenting the same images until they are correctly labeled. This behavior would be appropriate for a machine learner, but a human would quickly learn to cheat the learning task. Any revision style strategy would have to be carefully designed to ensure that concepts that are already learned are not continually revisited due to a de-emphasizing of earlier teaching answers.

5. Conclusion

Machine Teaching has the potential to enable humans to learn concepts without human-to-human expert tutoring. By automatically adapting the curriculum to a student's ability and performance, teaching can be performed in situations where it is difficult or prohibitively costly to get direct access to domain-level expertise from a human teacher. In this work, we have taken a step in this direction by proposing an interactive multi-class teaching strategy. Its objective is to present to the student the teaching images that will be most informative, given an online estimate of their current knowledge. Unlike other proposed strategies, we are less likely to teach outliers, and as a result, do not waste time showing unrepresentative images. Similar to curriculum learning [5], our strategy initially focuses on representative images and then introduces more difficult ones over time, as the student's performance improves.

5.1. Future Work

Currently, we present teaching images to the students one at a time. In future, we plan to investigate different methods for displaying images. Visualizations such as pairwise comparisons [22], and highlighting local regions [11] or parts [6], may prove more effective at conveying discriminative details and characteristics of different categories. Some images are intrinsically more 'memorable' than others [12, 20], and incorporating such measures into teaching image selection may also improve test time performance.

In curriculum learning, task difficulty is increased as performance improves. In future work, we shall also investigate other teaching paradigms such as the spiral approach

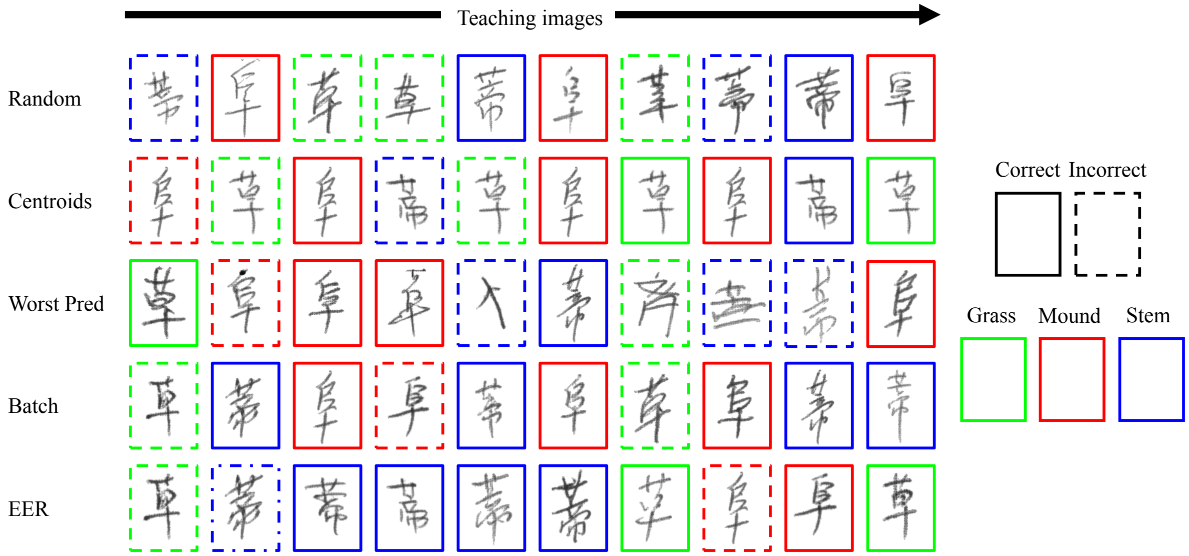

Figure 5: Example images and responses for the different teaching strategies from 5 sample individual students during teaching of the 'Chinese' dataset. Solid boxes indicate correct answers, dashed lines for incorrect answers, and box colors indicate the ground truth class labels.

to teaching [7]. In spiral learning, new categories are introduced over time while continually re-emphasizing the earlier concepts to ensure that they become committed to memory.

Given that we can now teach humans visual categorization tasks in an automated fashion, in future work we intend to investigate what additional information we can extract from our students during and after teaching. In contrast to machines, studies suggest that humans can learn with idealized versions of data that can have a different distribution from the test set [17]. Exploring teaching as a domain adaptation problem could allow us to acquire annotations for data which is very different from our teaching set. Finally, we have assumed that our feature space is correlated with a student's concept of similarity. It may be more effective to jointly estimate both the student's current ability and their notion of similarity during teaching.

Acknowledgements

Funding for this research was provided by EPSRC grant EP/K015664/1 and by Sustainable Fisheries Greenland. We would like to thank Maciej Gryka for his web development advice, and the Natural History Museum London for the butterfly data. Edward Johns' funding was provided by the Greenland Benthic Assessment project at the Institute of Zoology at ZSL, in association with the Greenland Institute of Natural Resources.

References

[1] Amazon Mechanical Turk. *https://www.mturk.com*, 2014.

[2] Django Web Framework. *https://www.djangoproject.com*, 2014.

[3] F. J. Balbach and T. Zeugmann. Recent developments in algorithmic teaching. In *Language and Automata Theory and Applications*. 2009.

[4] S. Basu and J. Christensen. Teaching classification boundaries to humans. In *AAAI*, 2013.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[6] T. Berg and P. N. Belhumeur. How do you tell a blackbird from a crow? In *ICCV*, 2013.

[7] J. S. Bruner. The Process of Education. *Harvard University Press*, 1960.

[8] M. Cakmak and M. Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI*, 2012.

[9] M. J. Crump, J. V. McDonnell, and T. M. Gureckis. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 2013.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[11] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 2013.

[12] A. Deza and D. Parikh. Understanding image virality. In *CVPR*, 2015.

[13] T. Doliwa, H. U. Simon, and S. Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In *Algorithmic Learning Theory*, 2010.

[14] J. Du and C. X. Ling. Active teaching for inductive learners. In *SDM*, 2011.

[15] S. Ebert, M. Fritz, and B. Schiele. RALF: A Reinforced Active Learning Formulation for Object Class Recognition. In *CVPR*, 2012.

[16] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 2010.

[17] G. Giguère and B. C. Love. Limits in decision making arise from limits in memory retrieval. *PNAS*, 110(19):7613–7618, 2013.

[18] S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 1992.

[19] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007.

[20] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *PAMI*, 2013.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

[22] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Breaking the interactive bottleneck in multi-class classification with active selection and binary feedback. In *CVPR*, 2010.

[23] B. Kulis. Metric Learning: A Survey. *Foundations & Trends in Machine Learning*, 2012.

[24] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*, 2012.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[26] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang. Casia online and offline chinese handwriting databases. *ICDAR*, 2011.

[27] C. Long, G. Hua, and A. Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *ICCV*, 2013.

[28] B. C. Love. Categorization. In *Oxford Handbook of Cognitive Neuroscience*, pages 342–358. 2013.

[29] O. Mac Aodha, N. D. Campbell, J. Kautz, and G. J. Brostow. Hierarchical Subquery Evaluation for Active Learning on a Graph. In *CVPR*, 2014.

[30] O. Mac Aodha, V. Stathopoulos, G. J. Brostow, M. Terry, M. Girolami, and K. E. Jones. Putting the scientist in the loop–accelerating scientific progress with interactive machine learning. In *ICPR*, 2014.

[31] D. L. Medin and M. M. Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207, 1978.

[32] P. O'Donovan, J. Lībeks, A. Agarwala, and A. Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)*, 2014.

[33] K. R. Patil, X. Zhu, Ł. Kopeć, and B. C. Love. Optimal teaching for limited-capacity human learners. In *NIPS*, 2014.

[34] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *ICML*, 2009.

[35] J. J. Richler and T. J. Palmeri. Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2014.

[36] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv:1409.0575*, 2014.

[38] B. Settles. *Active Learning*. Morgan & Claypool, 2012.

[39] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, and A. Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014.

[40] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010.

[41] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2004.

[42] X. Zhu. Machine teaching for bayesian learners in the exponential family. In *NIPS*, 2013.

[43] X. Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. *AAAI Conference on Artificial Intelligence (Senior Member Track)*, 2015.

[44] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.

[45] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *ICML workshops*, 2003.

[46] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *JMLR*, 2011.