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

Machine Teaching for visual classification is the process by which a machine teaches a human student to recognize object classes depicted in images. The task for the machine is to select images and display them to the student, such that they become an expert in visual recognition within a particular domain, in the shortest time possible. Unlike Active Learning, here the machine is the oracle, and the human is attempting to learn from the supervised data shown to them (see Figure 1). To learn more quickly, the student 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 image-importance varies with time and depends on their current knowledge.

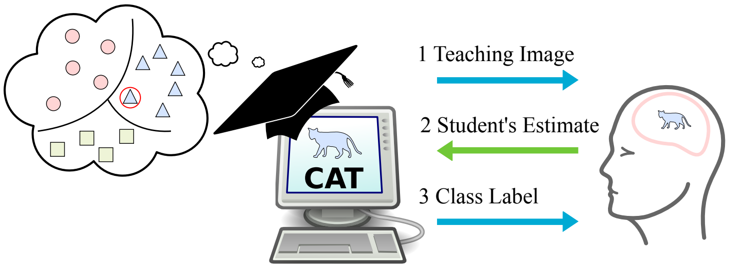

Figure 1: In Interactive Machine Teaching the computer teaches the student, one image at a time, by first showing them an image from a larger labeled image dataset, but concealing the true class label. The student responds with their estimate of the image's class. The teacher then updates their model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

Most existing methods for Machine Teaching are typically based on displaying fixed image sequences computed once offline [2], formulated only in theoretical simulation [3], or are evaluated only on simple synthetic data with binary classification [1]. We propose an interactive algorithm which probabilistically models the student's ability and progress, based on their correct and incorrect answers thus far. The student's understanding of an image is modeled by semi-supervised learning in a Gaussian Random Field [4], with the student's answers for observed images propagating their labels through the unobserved images. In this way, we model the scenario where the student answers incorrectly and their state diverges from the ground truth. The choice of the next image to display to the student is then the one which, if answered correctly, would have the greatest reduction on the future error [5]. By making no assumptions regarding the internal learning model used by the student, we allow for flexibility and adaptation to the specific needs of each individual.

To evaluate our teaching strategy, we performed experiments on real human subjects. Four datasets were used, with the number of classes varying from 3 to 5, and each consisting of at least 100 images per class. The datasets represent a range of image types that are challenging for non-domain experts to classify, including underwater species, leaves, butterflies, and Chinese characters. Image features were extracted using a Convolutional Neural Network. Experiments were divided into two phases. First, a teaching phase presented students with a sequence of teaching images, with the student giving an answer for each one, and the ground truth subsequently revealed. Second, a testing phase presented a sequence of testing images, where the ground truth was not presented, to evaluate the student's recognition ability after teaching.

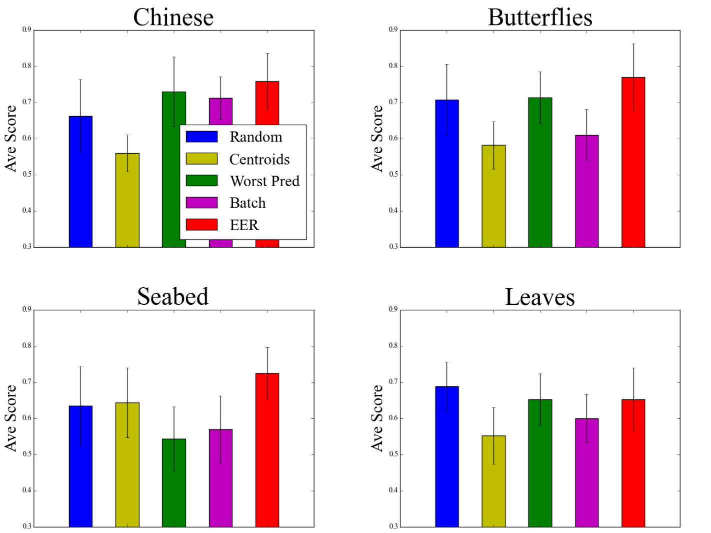

Figure 2: Human experiment results across the four datasets, showing the average score (higher numbers are better) during testing across all participants. Human participants on Mechanical Turk using our strategy (EER) tend to have better recognition performance on average after teaching, compared to the other baselines.

We compared our method (EER) to four other baselines, each with its own strategy for choosing the next teaching image to show: 1) Random - a random image, 2) Centroids - the centroid image of a random class, 3) Worst Pred - the image with the worst prediction given by the current model, and 4) Batch - the image that would reduce the future error the most based on all teaching images being answered correctly (a pre-computed, offline, non-adaptive strategy). Between 25 and 35 participants were recruited using Mechanical Turk for each dataset and strategy combination. Figure 2 shows the average scores from the testing round for all datasets and all strategies, where higher scores are better. Our strategy performs best on three of the four datasets, with no other strategy achieving consistently high scores. Statistical significance tests show our strategy to be the best to a high level of confidence, and the average time required for a student to respond during testing when being shown an image was lowest for ours.

Machine Teaching has the potential to enable humans to learn challenging visual concepts without human-to-human expert tutoring. In this work we have shown that by automatically adapting the curriculum to a student's knowledge online increases their ability to learn more effectively.

[1] Sumit Basu and Janara Christensen. Teaching classification boundaries to humans. In *AAAI*, 2013.

[2] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *ICML*, 2014.

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

[4] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.

[5] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *ICML workshops*, 2003.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.