Exploring Machine Teaching for Object Recognition with the Crowd

Jonggi Hong
Department of Computer Science
University of Maryland
College Park, MD, USA
jhong12@umd.edu

Kyungjun Lee
Department of Computer Science
University of Maryland
College Park, MD, USA
kjlee@cs.umd.edu

June Xu
Department of Electrical and Computer Engineering
University of Maryland
College Park, MD, USA
junexu@terpmail.umd.edu

Hernisa Kacorri
College of Information Studies
University of Maryland
College Park, MD, USA
hernisa@umd.edu

ABSTRACT

Teachable interfaces can enable end-users to personalize machine learning applications by explicitly providing a few training examples. They promise higher robustness in the real world by significantly constraining conditions of the learning task to a specific user and their environment. While facilitating user control, their effectiveness can be hindered by lack of expertise or misconceptions. Through a mobile teachable testbed in Amazon Mechanical Turk, we explore how non-experts conceptualize, experience, and reflect on their engagement with machine teaching in the context of object recognition.

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312873

Table 1: Characteristics of user studies and systems of prior work on behavioral studies and machine learning.

	User Study		System		
	#Participants (Avg.)	Crowdsourcing Platform	Machine Teaching	Hands-On Data	Crowdsourced Data
Thomz *et al.* [21]	14		•	•	
Jain *et al.* [11]	N/A		•		•
Yang *et al.* [23]	14		•		•
Guo *et al.* [8]	17		•		•
Our approach	100	•	•		•

KEYWORDS

teachable machines; interactive machine learning; object recognition; crowdsourcing.

INTRODUCTION

As the presence of machine learning and artificial intelligence increases in people's daily lives, so do efforts to better capture, understand, and imagine this coexistence. It is especially the case for systems that incorporate teachable interfaces (*e.g.*, [5, 7, 12, 14]), where end-users are called to consciously provide training examples and actively interact with the machine learning algorithm to increase its accuracy. By significantly constraining the conditions of the machine learning task to a specific user and their environment, these systems promise higher robustness in real world scenarios. However, they are also susceptible to the way that non-experts perceive machine teaching and their misconceptions.

With an intertwined goal of improving user experience while making learning more effective, teachable interfaces are fueled both by advances in machine learning (*e.g.*, transfer learning [18]) as well as human-computer interaction studies providing deeper insights into the users and their interactions [1]. This work contributes to the latter by exploring non-experts perception of machine teaching in the context of teachable object recognizers (TORs) [13, 15] with Amazon Mechanical Turk, a popular platform for studying user behavior at scale [3, 16], allowing us to recruit a large sample ($N = 100$) and collect data in realistic contexts using a performance-based payment scheme [10].

Participants in our study were asked to train and test an object recognizer using their mobile devices. We analyzed participants' photos qualitatively by identifying common patterns in their machine teaching strategies as well as quantitatively by reporting the performance of their recognition models.

RELATED WORK

Our study draws from prior work looking at how user interactions with an interactive machine learning system affect system performance, with a few representative papers in Table 1. While there is a rich HCI literature on crowdsourcing user perception (*e.g.*, graphical perception [9]) and user interactions with machine learning systems (surveyed in [22]), to our knowledge, this is the first study on crowdsourcing the perception of machine teaching in the context of teachable interfaces.

Thomaz *et al.* [21] characterized people's behavior of teaching robot actions by demonstration, and Yang *et al.* [23] conducted user studies on the perception of machine teaching for non-expert users in the context of general machine teaching. Those studies, however, include a relatively small participant pool compared to behavioral studies using crowdsourcing. Many studies have focused on building systems that combine data from a crowd to improve machine learning models in applications such as environmental sensing [8], identifying recidivism [20], and correcting traversing robots [11]. However, their goal is different from that of exploring machine teaching behaviors.

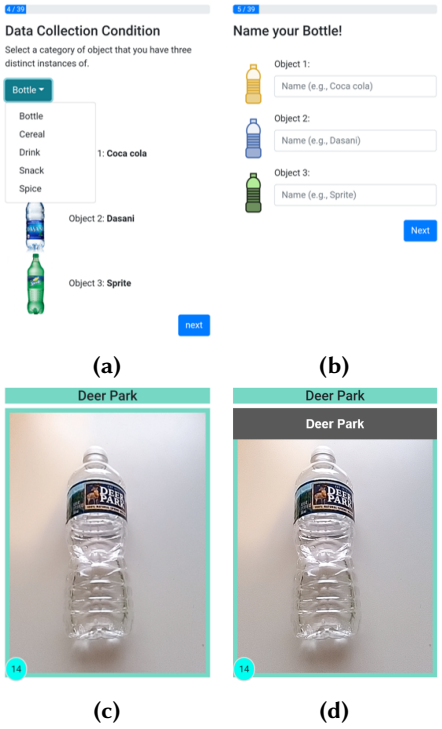


Figure 1: Interface for data collection.

Table 2: Variations across participants.

Code	Pre-test	Training	Post-test
Zoom in/out	49	65	57
Background	20	39	21
Side	11	36	29
Perspective	24	53	32
Position	6	39	17
Lightexposure	4	4	0
Light source	5	16	6

USER STUDY

We built a mobile web application that embeds the questionnaires and the testbed for collecting photos (shown in Fig. 1). The application communicates with a GPU server, where a TOR per participant is created and tested on the fly with the participant's photos. Similar to Kacorri *et al.* [13], each TOR model was built using transfer learning based on Google Inception (V3) [19].

Participants

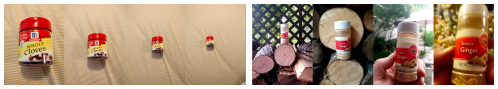
We recruited 100 participants from Amazon Mechanical Turk over an 11-day period. Their ages ranged 20-60 (M=32.6, SD=8.3); with 50 female, 49 male, and 1 non-binary participants. All participants reported having experience taking photos with their devices. When asked about their familiarity with machine learning, only few participants reported never having heard of it ($N = 7$) or having extensive knowledge on machine learning ($N = 1$). The majority reported having heard of it but not knowing what it does ($N = 46$) or having a broad understanding of what the machine learning is ($N = 46$).

Procedure

After providing demographic and technology experience information, participants selected a category of objects from five categories (Fig. 1a): *bottle*, *cereal*, *drink*, *snack*, and *spice*. They chose three objects belonging to their specific category and entered their names (Fig. 1b). The TOR module, with three stages (*pre-test*, *training*, and *post-test*), began once participants had identified the objects of choice in their environment. In pre-test (Fig. 1c), the TOR module randomly chose one of their objects, asked the participants to take a photo of it, and used the photo to test a generic model. The recognition result was then shown within two seconds (Fig. 1d). This process was repeated fifteen times (5 photos per object). The generic model was pre-trained on ImageNet [4] and then fine-tuned to objects in the GTEA dataset [6]. Since this model was not trained on participants' personal objects, almost all the predictions were wrong in this stage, which aimed to familiarize with the object recognition task.

During training, the module randomly selected one of the objects, and asked participants to take 30 photos of that object consecutively to train their own TOR; *i.e.* a total of 90 photos (30 photos per object). Participants were specifically asked to train their TOR to be robust so that it learns to *identify objects anywhere, anytime, for anyone*. The module trained the TOR with photos from the participants.

Participants were instructed that they would be given a bonus ($2) if their model passed a secret robustness test, a performance-based payment method based on Ho *et al.* [10]. After training, we asked participants to report their confidence in the robustness of their models and factors they perceived as important in training the TOR. Participants were then asked to select a subset of 20, 5, and 1 photo(s) that would make their model faster and more accurate. Last, in the *post-test*, participants performed the same procedure as in the pre-test with predictions coming from their newly trained TOR.



(a) Varying the distance between the camera and object (P1, P7).

(b) Varying the background (P7, P9).

(c) Varying the sides of objects (P4, P12).

(d) Varying the angles of camera and object (P8, P24).

(e) Varying the position of objects in photos (P26, P31).

(f) Varying illumination with different degrees of light exposure (P14, P65).

(g) Varying illumination with different sources of light (P25, P33).

Figure 2: Variation examples.

PATTERNS IN MACHINE TEACHING STRATEGIES AND MODEL PERFORMANCE

Two researchers examined the behavioral patterns within participants' photos with a thematic coding approach [2]. The researchers created initial codebooks independently and later resolved disagreements in their codebooks to generate the final codebook for inter-rater validation. The photos in different steps (*pre-test*, *training*, and *post-test*) were coded, separately. There is a substantial agreement between the coding data from the two researchers (Cohen's $Kappa = 0.80$).

Codebook

Our coding scheme is based on four dimensions that humans generalize across for visual recognition [17]: size, location, viewpoint, and illumination. Specifically, we focus on presence/absence of variation across factors within these dimensions for at least one of the objects.

Size (zoom in/out). We have four levels (0–0.25, 0.25–0.5, 0.5–1.0, higher than 1.0) for the ratio of the height of an object to the height of a photo. When there was more than one ratio level in photos of an object, we assumed that the participant varied the size factor.

Location (background). Variation in location was marked if a participant took photos of an object in different places as indicated by the background. Slight changes of background due to variation in viewpoint were not considered as variation in location/background.

Viewpoint (side, perspective, position). Viewpoint is captured by the side, angle, and position of the object relative to the camera. For example photos may include variation on object side if both front and back of the object is present; perspective if the angle between and camera and the object changes significantly; and position if the object is centered in some of the frames but not on others.

Illumination (light exposure, light source). This included variations in brightness due to change in light exposure (*e.g.*, in the amount of ambient light while the object remains at the same location) or due to change in light source (*e.g.*, use of flash light or another location).

Results

More than half of the participants took photos with variations in the size and viewpoint for at least one object when training their models (Table 2, Fig. 3). We found that participants tent to vary most the distance (Fig. 2a) and the angle (Fig. 2d) of the object from the camera. Fewer than 10 participants varied the light exposure on their training examples (Fig. 2f), which was the least common type of variation during training. While non experts, participants seem to understand the importance of variation in their training examples for robustness, with 77% of them including at least one type of variation. However, only 11% of them generalized across all four dimensions.

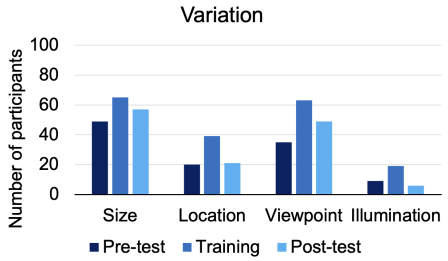


Figure 3: Number of participants whose images varied in at least one code per factor.

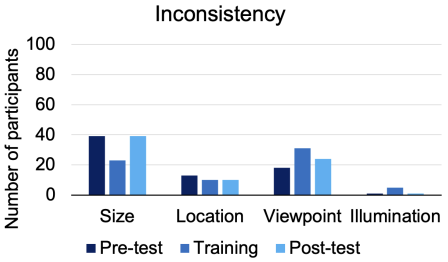


Figure 4: Number of participants whose images had inconsistencies in their variations for at least one code per factor.

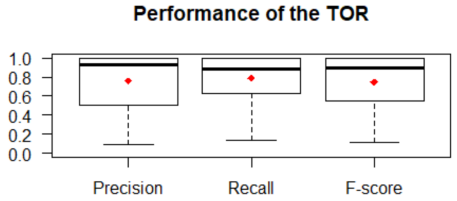


Figure 5: Precision, recall, and F-score of the TORs trained with photos from the Training ($N = 100$)

.

Even within the same dimension, variation is present across different factors; *e.g.*, in viewpoint, far fewer participants varied the side (Fig. 2c) and position (Fig. 2e) than perspective (Fig. 2d).

We observed that variations were not consistent across all objects in training for some of the participants, as shown in Fig. 4. After training, the majority of participants were uncertain ($N = 52$) or very uncertain ($N = 10$) about the robustness of their models. Though, a large portion stated as being certain ($N = 30$) or very certain ($N = 8$). When calculating precision, recall, and F-score of their models on their post-test images we found that on average they were 0.82, 0.79, and 0.83, respectively (Fig. 5). We did not observe a relationship between the performance of their models and participants reported expertise with machine learning or certainty in robustness of their models.

DISCUSSION

This work presents preliminary results from our analysis of non-experts perception of machine teaching in the context of teachable object recognizers. Our qualitative analysis of the collected training photos shows that more than half of the participants included some variation (in terms of object size, location, or viewpoint) in their training examples for their model to be robust. However, this often did not include variation in illumination (about 10%). Moreover, some of the participants (about 35%) incorporated different factors in variation across the three objects in the training stage. Models trained by the participants achieved a 83% accuracy on average. We did not observe a relationship between the model performance and the participants reported experience with machine learning.

We are currently investigating how participants alter their training if they are given a second chance to train their models. Also, we will look into any associations between model performance, participants' background (*e.g.* technology experience), and training strategies. We will explore clustering approaches for uncovering common behavioral patterns in an attempt to capture how non-experts may conceptualize robustness in machine teaching for object recognition. We will contextualize our quantitative results with qualitative analysis of participants text responses, feedback, and comments.

ACKNOWLEDGMENTS

This work is supported by NSF (Award: #1816380).

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[3] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. (2009).

[5] Christoph Evers, Romy Kniewel, Kurt Geihs, and Ludger Schmidt. 2014. The user in the loop: Enabling user participation for self-adaptive applications. *Future Generation Computer Systems* 34 (2014), 110–123.

[6] Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*. IEEE, 3281–3288.

[7] Rebecca Fiebrink, Perry R Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 147–156.

[8] Anhong Guo, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, and Jeffrey P Bigham. 2018. Crowd-AI Camera Sensing in the Real World. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 111.

[9] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 203–212.

[10] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 419–429.

[11] Ashesh Jain, Debarghya Das, Jayesh K Gupta, and Ashutosh Saxena. 2015. Planit: A crowdsourcing approach for learning to plan paths from large scale preference feedback. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 877–884.

[12] Hernisa Kacorri. 2017. Teachable machines for accessibility. *ACM SIGACCESS Accessibility and Computing* 119 (2017), 10–18.

[13] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5839–5849. https://doi.org/10.1145/3025453.3025899

[14] Google Creative Lab. 2017. Teachable machine. https://experiments.withgoogle.com/teachable-machine

[15] Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 12. https://doi.org/10.1145/3290605.3300566

[16] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[17] Thomas J Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience* 5, 4 (2004), 291.

[18] Sinno Jialin Pan, Qiang Yang, et al. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[20] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human+ Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123* (2018).

[21] Andrea L Thomaz and Maya Cakmak. 2009. Learning about objects with human teachers. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. ACM, 15–22.

[22] Jennifer Wortman Vaughan. 2018. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research* 18, 193 (2018), 1–46.

[23] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*. ACM, 573–584.