Demonstration of the EMPATHIC Framework for Task Learning from Implicit Human Feedback

Yuchen Cui,[1*] **Qiping Zhang,**[1*] **Sahil Jain,**[1] **Alessandro Allievi,**[1,2]
Peter Stone,[1] **Scott Niekum,**[1] **W. Bradley Knox**[1,2]
[1]University of Texas at Austin, Austin, TX 78712
[2]Robert Bosch LLC, Austin, TX
{yuchencui, qpzhang, sahiljain11}@utexas.edu, {pstone, sniekum}@cs.utexas.edu
{brad.knox, alessandro.allievi}@us.bosch.com

Abstract

Reactions such as gestures, facial expressions, and vocalizations are an abundant, naturally occurring channel of information that humans provide during interactions. An agent could leverage an understanding of such *implicit* human feedback to improve its task performance at no cost to the human. This approach contrasts with common agent teaching methods based on demonstrations, critiques, or other guidance that need to be attentively and intentionally provided. In this work, we demonstrate a novel data-driven framework for learning from implicit human feedback, EMPATHIC. This two-stage method consists of (1) mapping implicit human feedback to relevant task statistics such as rewards, optimality, and advantage; and (2) using such a mapping to learn a task. We instantiate the first stage and three second-stage evaluations of the learned mapping. To do so, we collect a dataset of human facial reactions while participants observe an agent execute a sub-optimal policy for a prescribed training task. We train a deep neural network on this data and demonstrate its ability to (1) infer relative reward ranking of events in the training task from prerecorded human facial reactions; (2) improve the policy of an agent in the training task using live human facial reactions; and (3) transfer to a novel domain in which it evaluates robot manipulation trajectories. In the video, we focus on demonstrating the online learning capability of our instantiation of EMPATHIC.

Introduction

People often react when observing an agent—whether human or artificial—if they are interested in the outcome of the agent's behavior. We have scowled at robot vacuums, raised eyebrows at cruise control, and rebuked automatic doors. Such reactions are often not intended to communicate to the agent and yet nonetheless contain information about the perceived quality of the agent's performance. A robot or other software agent that can sense and correctly interpret these reactions could use the information they contain to improve its learning of the task. Importantly, learning from such *implicit* human feedback does not burden the human, who naturally provides such reactions even when learning does not occur. We view learning from implicit human feedback (LIHF) as complementary to learning from explicit human teaching, which might take the form of demonstrations (Argall et al. 2009), evaluative feedback (Knox and Stone 2009; Knox, Stone, and Breazeal 2013), or other communicative modalities (Chernova and Thomaz 2014; Sadigh et al. 2017; Admoni and Scassellati 2017). The problem of **Learning from Implicit Human Feedback** (LIHF) asks how an agent can learn a task with information derived from human reactions to its behavior. We approach LIHF with data-driven modeling that creates a general **reaction mapping** from implicit human feedback to task statistics. A simplified overview of our proposed method is shown in Fig. 1.

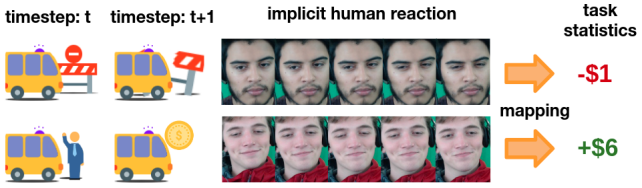

Figure 1: Illustrative overview of the proposed method.

The EMPATHIC Framework

We propose a data-driven solution to the LIHF problem that infers relevant task statistics from human reactions. The EMPATHIC framework has two stages: (1) learning a mapping from implicit human feedback to relevant task statistics and (2) using such a mapping to learn a task. In the first stage, human observers are incentivized to want an agent to succeed—to align the person's $R^{\mathcal{H}}$ with a known task reward function R—and they are then recorded while observing the agent. Task statistics are computed from R for every timestep to serve as supervisory labels, which train a mapping from synchronized recordings of the human observers to these statistics. Task state and action are *not* inputs to the reaction mapping, allowing it to be deployed to other tasks. In the second stage, a human observes an agent attempt a task with sparse or no environmental reward, and the human observer's reaction to its behavior is mapped to otherwise unknown task statistics to improve the agent's policy, either directly or through other usage of the task statistics, such as guiding exploration or inferring the reward function $R^{\mathcal{H}}$ that describes the human's utility. This demonstration presents one instantiation of EMPATHIC, using facial reactions as the modality for implicit human feedback.

[*]Equally contributing authors

We designed two task domains, one is a simulated driving domain named *Robotaxi* and the other is a robotic sorting task. Participants were recruited to observe agents performing tasks in both domains. To minimize explicit feedback (i.e., intended to influence the agent), participants were told that their "reactions are being recorded for research purposes", and nothing more was said regarding our intended usage of their reactions. This experimental setup contrasts with prior related work (Li et al. 2020; Veeriah, Pilarski, and Sutton 2016; Arakawa et al. 2018), in which human participants were explicitly asked to teach with their facial expressions, and aligns with a key motivation for the LIHF problem, which is to leverage data that is already being generated in existing human-agent interactions. We use data collected in the *Robotaxi* domain to instantiate the first stage of EMPATHIC and demonstrate its effectiveness in three different instantiations of the stage 2 task including: 1) offline learning in *Robotaxi*, 2) online learning in *Robotaxi* and 3) offline evaluation of trajectories in the robotic sorting task.

Reaction Mapping Design

A reaction mapping takes a temporal series of extracted features as input and outputs a probability distribution over reward classes. We use a pre-trained model to extract facial features from video data and train a deep neural network on predicting rewards with the extracted features in a supervised way. An open-source toolkit, OpenFace 2.0 (Baltrusaitis et al. 2018; Zadeh et al. 2017; Baltrušaitis, Mahmoud, and Robinson 2015), is used to extract features from raw videos of human reactions. For each image frame in the video, OpenFace extracts head pose and activation of facial action units (FAUs). For detecting head nods and shakes, we explicitly model the head-pose changes by keeping a running average of extracted head-pose features and subtract it from each incoming feature vector. Frequencies of changes in head-pose are then computed through fast Fourier transform, and the coefficients of frequencies are used as head-motion features. To allow the series of input features to cover a large enough temporal window of reactions, feature vectors of consecutive image frames are combined through max pooling of each dimension, resulting in temporally aggregated feature vectors of the same size. We include an optional auxiliary task of predicting the corresponding annotations as a single flattened vector, in which each binary element indicates whether a reaction gesture is occurring. This auxiliary task is intended to speed representation learning and act as a regularizer. We also use a binary classification loss that combines the two negative reward classes as one, which reintroduces the ordinality of the reward classes by additionally penalizing predictions with the wrong sign.

Evaluation Results

To validate that our instantiation of stage 1 effectively enables task learning in stage 2, we evaluated mappings learned in stage 1 on three different tasks. In what follows, we refer to observers from stage 1 who have created data in the training set as "*known subjects*".

Firstly, the learned reaction mappings are evaluated on a reward-ranking task in the *Robotaxi* domain. The maximum a posteriori reward ranking is chosen as the learned mapping's single estimation after incorporating mappings from all human reaction data in an episode. Using Wilcoxon Signed-Rank test, the mapping's performance on the *holdout* set is significantly better than uniformly random guessing ($\tau = 0$), supporting the hypothesis that our learned reaction mappings outperform uniformly random reward ranking using reaction data from *known subjects* watching the *Robotaxi* task; $p = 0.0024$ with the annotation-reliant auxiliary task and $p = 0.0207$ without it.

Secondly, the learned reaction mapping can be leveraged to interactively improve an agent's policy.[1] Specifically, the agent updates its belief over all possible reward rankings using human reactions to its recent behaviors and then follows a policy that is approximately optimal with respect to the most likely reward function. To test such online policy learning, all data collected in stage 1 trains a single reaction mapping, and this reaction mapping is used for single-episode sessions with human observers, none of whom created data within the stage-1 training set. 9 of the 10 participants' interactions achieved a better return than that of a random policy, and 7 of the 10 participants' interactions ended with the probability of reward mappings that lead to optimal behaviors being the highest, moderately supporting the hypothesis that the learned reaction mappings will improve the online policy of a *Robotaxi* agent via updates to its belief over reward functions, based on *online* data from *novel* human observers.

Lastly, we evaluate the reaction mappings on the robotic sorting task. We leverage a binary classification loss and interpret the output as a "positivity score". Human participants observed 8 total distinct trajectories. For each trajectory, we compute an overall (cross-subject) positivity score as the mean of the trajectory's per-subject positivity scores. After ranking the 8 trajectories by these scores, Kendall's τ independence test yields $\tau = 0.70$ ($p = 0.034$). This result shows that the learned reaction mappings can be adapted to evaluate robotic-sorting-task trajectories and outperform uniformly random guessing on return-based rankings of these trajectories, using reaction data from *known subjects*.

Conclusion

In this paper we introduce the LIHF problem and the EMPATHIC framework for LIHF. We instantiate the EMPATHIC framework with two different domains and demonstrate that our instantiation is able to interpret human facial reactions in both the training task and the deployment task. Our instantiation of EMPATHIC in this work is limited to a single training task and similar testing tasks. An important future extension is to generalize this method to tasks with varying temporal characteristics and reward structures. This work maps from facial reactions (to rewards). In future work, other forms of human implicit feedback, such as gaze, vocalizations, and gestures, could be included to get a more accurate mapping to different task statistics and better performance in a variety of real-world tasks.

[1] This online learning setting is demonstrated in the video.

References

Admoni, H.; and Scassellati, B. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6(1): 25–63.

Arakawa, R.; Kobayashi, S.; Unno, Y.; Tsuboi, Y.; and Maeda, S.-i. 2018. DQN-TAMER: Human-in-the-Loop Reinforcement Learning with Intractable Feedback. *arXiv preprint arXiv:1810.11748* .

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57(5): 469–483.

Baltrušaitis, T.; Mahmoud, M.; and Robinson, P. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, 1–6. IEEE.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. IEEE.

Chernova, S.; and Thomaz, A. L. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8(3): 1–121.

Knox, W. B.; and Stone, P. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*, 9–16. ACM.

Knox, W. B.; Stone, P.; and Breazeal, C. 2013. Training a robot via human feedback: A case study. In *International Conference on Social Robotics*, 460–470. Springer.

Li, G.; Dibeklioğlu, H.; Whiteson, S.; and Hung, H. 2020. Facial feedback for reinforcement learning: a case study and offline analysis using the TAMER framework. *Autonomous Agents and Multi-Agent Systems* 34(1): 1–29.

Sadigh, D.; Dragan, A. D.; Sastry, S.; and Seshia, S. A. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems*.

Veeriah, V.; Pilarski, P. M.; and Sutton, R. S. 2016. Face valuing: Training user interfaces with facial expressions and reinforcement learning. *arXiv preprint arXiv:1606.02807* .

Zadeh, A.; Chong Lim, Y.; Baltrusaitis, T.; and Morency, L.-P. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2519–2528.