Augmenting Evaluative Corrective Learning with Multimodal Feedback from Human Instructors

Akhil Goel, Vishnu Jaganathan, Bryan Zhao

Introduction

To summarize the goals and motivations behind this project, we are seeking to improve on TAMER [1]. When it was first introduced, TAMER provided a framework for enabling non-expert humans to easily train a learning agent to perform complex tasks by providing numerical feedback. Though it is a powerful idea, we recognize that feedback can be difficult to provide when restricted to only a numerical format, especially when one must anticipate the short-term intentions and long-term goals of the agent. Apart from evaluating a "good" or "bad" sign, the expert typically also must determine their confidence in order to assign a scaled numerical reward. As creatures of multiple senses, providing only numerical feedback is an unnatural way for humans to train an agent. We hope to improve on TAMER by enabling trainers to provide feedback with *multiple modes*, making it more accessible, engaging, and natural for humans to train an agent. Furthermore, by increasing the modalities in the provided feedback, we hope that the amount of information encoded in feedback will increase, enabling the policy to converge faster. By making feedback more engaging and expressive, we hope to make training learning agents more accessible to roboticists, industries that currently use robots, and even the everyday person who has robots in the home. Our work on incorporating multiple modalities into reward functions could be applicable to the work of people in the scientific learning community [2].

Current Progress & Next Steps

There are three main parts of the project that we are currently working on. The first is creating a game environment where users can train an agent using feedback. We have implemented the Atari games from OpenAI Gym [3] and are implementing a version of Deep TAMER [4] where users provide numerical feedback in real time as the agent plays the Atari game Breakout. The next steps are to finish the code for training the reward model as well as incorporate other modalities of feedback as the data pipelines become finalized.

The second part of the project is creating data pipelines to collect verbal and facial data. We have written code converting pre-recorded audio clips to text and are able to perform sentiment analysis on that text. The next step is to make this data pipeline real time. Additionally, we plan to use signal processing of the audio and derive a numerical rating based on the amount of excitement or disappointment shown in the user's voice. While other works [5] have used only textual meanings and sentiment to provide feedback to TAMER, we also plan to incorporate the tone of voice or emotion while deriving the overall feedback from verbal data. The hypothesis is that in times of frustration or excitement, the tone of voice can yield additional information about how the trainer feels about the robot performance which can affect the magnitude of the sentiment expressed through the text.

In terms of visual data, we are working on collecting static frames of facial expressions for sentiment analysis. Once this pipeline is built and we can verify its speed, the sentiment analysis could be extended to operate on real-time video data. Prior work on facial expressions [6] used a binary classification of expression, but we plan to classify on multiple emotions and extend this to predict the intensity of those emotions as well.

Finally, the third part of the project is human experimentation and evaluation of our framework. We are working on an IRB application and have completed CITI Training. The next steps are to complete the experimental design, surveys, and recruitment documents.

Risks

To address data security for user experiments, we plan to encrypt the user data and also enforce a retention policy where we will delete the recorded training feedback of participants (video, textual feedback, scores) after the completion of our experiments.

From a technical standpoint, there are two primary challenges that might cause roadblocks with our approach. While there is already a large existing body of work in interpreting language and facial expressions, it is still challenging to interpret these accurately and convert them to a reward signal *in real-time*. We are mitigating this risk by restricting the problem to only focus on basic sentiment classification for verbal data and emotion classification for facial data. If the latency is satisfactory, we can look to increase the predictive power of our models by increasing the number of classes, features, or parameters of our models. Another challenge lies in combining the different reward modalities to effectively train with TAMER. To address this, we aim to start with a simple linear combination of scores, and demonstrate an effective set of coefficients for our specific experiments. If this is successful, we could try to extend our work to a more general framework for combining the different modalities such as using neural networks or attention.

References

[1] W. Bradley Knox and P. Stone, "TAMER: Training an Agent Manually via Evaluative Reinforcement," in *2008 7th IEEE International Conference on Development and Learning*, Monterey, CA, Aug. 2008, pp. 292–297. doi: 10.1109/DEVLRN.2008.4640845.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent Advances in Robot Learning from Demonstration," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 3, no. 1, pp. 297–330, May 2020, doi: 10.1146/annurev-control-100819-063206.

[3] G. Brockman *et al.*, "OpenAI Gym." arXiv, Jun. 05, 2016. Accessed: Oct. 21, 2022. [Online]. Available: http://arxiv.org/abs/1606.01540

[4] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, "Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces." arXiv, Jan. 19, 2018. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/1709.10163

[5] P. Goyal, S. Niekum, and R. J. Mooney, "Using Natural Language for Reward Shaping in Reinforcement Learning." arXiv, May 31, 2019. Accessed: Nov. 18, 2022. [Online]. Available: http://arxiv.org/abs/1903.02020

[6] G. Li, H. Dibeklioğlu, S. Whiteson, and H. Hung, "Facial feedback for reinforcement learning: a case study and offline analysis using the TAMER framework," *Auton. Agents Multi-Agent Syst.*, vol. 34, no. 1, p. 22, Apr. 2020, doi: 10.1007/s10458-020-09447-w.