Augmenting Evaluative Corrective Learning with Multimodal Feedback from Human Instructors

Akhil Goel, Vishnu Jaganathan, Bryan Zhao

I. PROJECT MOTIVATION

Many existing frameworks such as TAMER [1] and COACH [2] rely on providing expert feedback to agents in order to learn a desired behavior. However, providing such feedback in an effective manner poses challenges and limitations. The TAMER framework queries an expert for a numerical score of an agent's behavior which indicates the quality of the action taken, and the agent seeks to maximize the expected reward over its trajectory. This is inherently limited because the expert must quantify their nuanced feedback of the actions into a singular scalar value. Furthermore, since expert feedback is queried periodically and the significance of actions taken during that time scale may vary, the expert may have different levels of confidence in their feedback due to varying amounts of information about the agent's behavior. In such situations, providing a neutral score and hence effectively providing no feedback might limit the quantity and quality of information an expert may be able to provide to the agent.

Incorporating multimodal data is a large area of potential growth in the field of learning from demonstration [3]. Thus, we seek to improve upon the feedback-giving framework by incorporating multiple modalities of human response, incorporating language and facial expressions. By doing so, we hope to solve the practical issues of TAMER and COACH by making feedback more accurate and easier to provide.

II. DESCRIPTION

Our algorithm would operate similarly to TAMER and COACH in that feedback will ultimately be captured by a single scalar value at regular time intervals over an agent's trajectory. However, we plan to incorporate verbal feedback and/or facial expressions into this score. We target these two modes of feedback because they are primary modes that humans naturally provide feedback in the real world. Furthermore, their ability to store large amounts of information in small amounts of data. Natural language in particular has been used to train a robot learner [4], and we hope to expand upon this with facial expressions to fully integrate verbal and visual cues with feedback based algorithms like TAMER and COACH. Experiments on the model will be performed on a suite of Atari games from OpenAI Model Gym [5].

The first major experimental axis will focus on how to quantify verbal statements and facial expressions into numerical scores. For the verbal statements, there are various sentiment analysis methods that could be used to determine the quality of the action taken [6]. For facial expressions, a classifier could be trained to determine the sentiment or mood, and a numerical value would be assigned to each sentiment. These could both be used in conjunction with a confidence rating to weight the magnitude of the final feedback given to the agent.

The second experimental axis would be how to combine the scores of different modalities. In the simplest case, all of the feedback could be combined into a single score. This could be performed linearly and would reduce to the original problem of simply quantizing multimodal feedback. This could also be done nonlinearly through a custom nonlinear function or the use of deep learning. Furthermore, we anticipate that some feedback will not be able to be incorporated into the score. For example, verbal directives such as "go forward more" encodes more information than just saying the actions that were taken were not desirable; it also includes a directive that a certain action must be taken instead. In this case, it would be more effective to make direct modifications into the policy based on the expert recommendation.

III. DATA

The data we want to collect is in three forms. One being the usual numeric score for TAMER and COACH given as feedback by the user. Additionally, we would want to collect data on facial expressions and their sentiment labels for the purpose of correlating a user's facial expression to feedback. Similarly, we would like to collect feedback on verbal statements rating the robot learner's feedback to use as input to train the model. We plan to use all three of these modes of feedback in combination to train the robot and identify what the most effective combination is. This data is to be collected mainly from us three group members and our peers in the class and student volunteers at Georgia Tech. The types of data we are collecting imposes a risk for the project. Since we are collecting voice and face data, we need to be mindful of the privacy of that data.

IV. EXPECTED OUTCOME

We expect to devise a way to collect multimodal feedback data in the form of numerical score, verbal speech, and facial expressions and combine them to train an agent to learn from demonstrations in a TAMER and COACH-like fashion. Increasing the modality of feedback that is provided to the agent will improve the quality of the feedback. With higher quality feedback we expect our algorithm to converge to an optimal policy faster when compared to other methods that use human feedback to train an agent. Furthermore, we expect that our method will make it easier for human instructors to provide feedback to the agent.

V. BENCHMARKS

Our primary benchmark will be vanilla TAMER [1], COACH [2], and ACTAMER [7] with only numerical human feedback and compare their performance and ease of use with multimodal feedback. We will also train a behavioral cloning agent using DAgger [8] to compare the ease of use of training an optimal agent against methods that require only demonstrations instead of feedback.

VI. TIMELINE

By our project update date on November 16th, we plan to have our multimodal pipelines of speech-to-text feedback and facial expression to TAMER/COACH score complete. We also will devise a way to combine these feedback and handle feedback that cannot be combined. Lastly, we aim to have the basic code of the TAMER/COACH algorithm complete and able to interact with our environment.

By the final project deadline on December 13th, we plan to collect data on our different feedback modes and use it to train the model. From this, we can make conclusions on the efficacy of our algorithm and finalize our final paper.

VII. 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] C. Celemin and J. Ruiz-del-Solar, “COACH: Learning continuous actions from COrrective Advice Communicated by Humans,” in *2015 International Conference on Advanced Robotics (ICAR)*, Jul. 2015, pp. 581–586. doi: 10.1109/ICAR.2015.7251514.

[3] 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.

[4] A. Silva, N. Moorman, W. Silva, Z. Zaidi, N. Gopalan, and M. Gombolay, “LanCon-Learn: Learning With Language to Enable Generalization in Multi-Task Manipulation,” *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1635–1642, Apr. 2022, doi: 10.1109/LRA.2021.3139667.

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

[6] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[7] N. A. Vien, W. Ertel, and T. C. Chung, “Learning via human feedback in continuous state and action spaces,” *Appl. Intell.*, vol. 39, no. 2, pp. 267–278, Sep. 2013, doi: 10.1007/s10489-012-0412-6.

[8] S. Ross, G. J. Gordon, and J. A. Bagnell, “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning.” arXiv, Mar. 16, 2011. Accessed: Oct. 21, 2022. [Online]. Available: http://arxiv.org/abs/1011.0686