ETH zürich

Organisationseinheit verbal
optional auf 2 Zeilen

RLHF using natural language feedback.

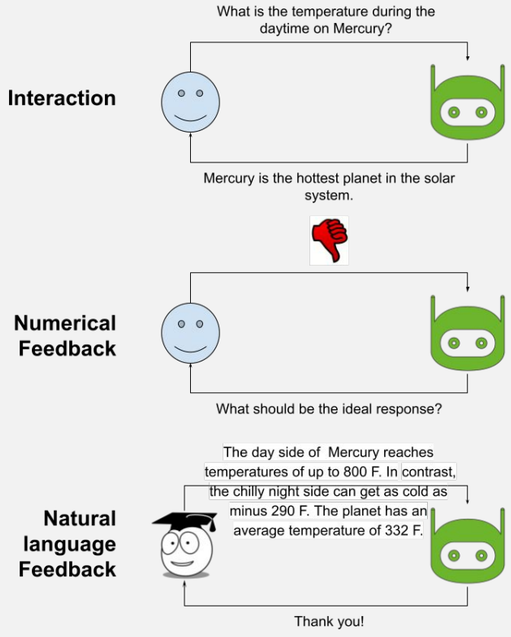
Tanay Gahlot[1,2]
[1]D-INFK, ETH Zurich; [2]Google

1 Introduction

Reinforcement learning from human feedback(RLHF) has proven beneficial in aligning language models[1], however, majority of these approaches rely on numerical feedback[2]. RLHF using natural language feedback remains largely unexplored[2]. We propose two techniques to factor in natural language feedback alongside numerical feedback.

2 Problem Overview

After deployment, the BOT can gather human feedback from the user in the following manner[4]:



3 Methods

1. Joint loss function

1.1 Policy gradient objective[3]:

$$J_{\mathrm{PG}}(\pi_\theta) = \mathbb{E}_{\tau\sim\rho_{\pi_\theta}}\left[\sum_{t=0}^{T-1}\gamma^t\cdot\hat{Q}_t\log\pi_\theta(a_t|s_t)\right],$$

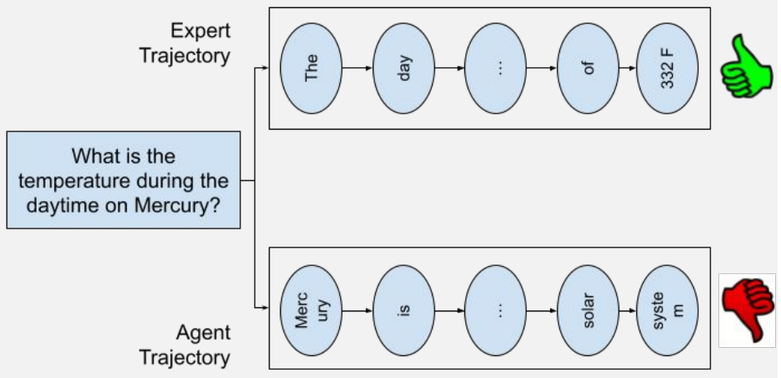
1.2 Interactive Imitation learning objective[2]:

$$J_{\mathrm{IL}}(\pi_\theta) = E_{s\sim d^{\pi_E},a\sim\pi_E(.|s)}[\log(\frac{\pi_E(a|s)}{\pi_\theta(a|s)})]$$

1.3 Joint objective:

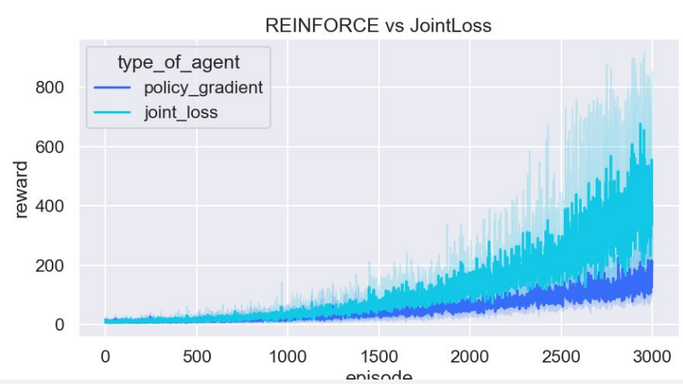
$$J(\pi_\theta) = \alpha J_{\mathrm{PG}}(\pi_\theta) + (1-\alpha)J_{\mathrm{IL}}(\pi_\theta)$$

2. Expert trajectory injection



4 Experiment setup and Results

- The proposed approaches are compared against REINFORCE algorithm(policy_gradient).
- Metric of comparison is cumulative reward after n episodes.
- The environment used for comparison is Mujoco's InvertedPendulum-v4. We will move to real world environment proposed in fits[4] once we get positive results in Mujoco.



References

[1] Long Ouyang Training language models to follow instructions with human feedback. 2022.
[2] Patrick Fernandes. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. Transactions of the Association for Computational Linguistics
[3] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.
[4] Jing Xu Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback, 2022.

Partner/Sponsor:

Project No 28