Enhancing Game Control Through Hybrid Reinforcement Learning

Danhua Yan [1]

1. Background and Related Work

Training Reinforcement Learning (RL) agents solely from exploration is usually data inefficient and likely converges to sub-optimal policies. The research field of bootstraping an RL agent's policy from demonstrations or imitation learning shows significant promise. Various hybrid paradigms that combine human guidance as offline RL and agent exploration as online RL have shown they can accelerate policy learning and achieve above-demonstration performance. (Hester et al., 2017; Nair et al., 2018; Song et al., 2023; Ren et al., 2024; Coletti et al., 2023).

This project investigates how hybrid RL can effectively enhance game control through guided explorations of the agent. It aims to evaluate the potential for achieving performance that surpasses the demonstration level.

2. Data

This project will leverage the `gym-super-mario-bros` library to create an OpenAI Gym environment for training an agent to play the NES game Super Mario Bros. The states are represented by in-game visual frames, the actions are discrete game controls, and the rewards are the game's scoring systems. Human demonstrations will be recorded into offline trajectories, and the agent will perform online exploration in the environment.

3. Methods

The methodology involves assessing RL agent performance on baseline and hybrid approaches on the trained game level and an unseen similar level to compare generalization capabilities.

- Baselines: Offline-only and online-only approaches are used for comparisons:
  - Imitation Learning Only: Train a policy via behavioral cloning (BC) using human demonstrations.
  - Online-only RL: Train an agent with online exploration only, leveraging Deep Q-Learning (DQN) with ϵ-greedy explorations.
- Hybrid RL: We propose two paradigms of hybrid approaches:
  - Following the DQfD (Deep Q-Learning from Demonstrations) framework by (Hester et al., 2017), which incorporates expert demonstrations into the replay buffer of DQN to control explorations.
  - Leveraging behavioral cloning (BC) as a warm-start, then further leveraging PPO (Proximal Policy Optimization) for policy fine-tuning. This approach is inspired by (Coletti et al., 2023).

4. Evaluations

We will evaluate the approaches using both quantitative and qualitative metrics. Quantitatively, performance will be measured via cumulative reward, level completion rate, and distance traversed per episode, plotted as learning curves against training episodes or timesteps. Multiple independent runs will ensure statistical significance. Sample efficiency will be analyzed by measuring interactions required to reach performance thresholds and wall-clock training time. Qualitatively, gameplay visualizations and trajectory overlays will provide insights into behavioral strategies.

References

Coletti, C. T., Williams, K. A., Lehman, H. C., Kakish, Z. M., Whitten, D., and Parish, J. Effectiveness of warm-start ppo for guidance with highly constrained nonlinear fixed-wing dynamics. *2023 American Control Conference (ACC)*, pp. 3288–3295, 2023. URL https://api.semanticscholar.org/CorpusID:259338376.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Dulac-Arnold, G., Osband, I., Agapiou, J., Leibo, J. Z., and Gruslys, A. Deep Q-learning from Demonstrations, November 2017. URL http://arxiv.org/abs/1704.03732. arXiv:1704.03732 [cs].

[1]Department of Computer Science, Stanford University. Correspondence to: Danhua Yan <dhyan@stanford.edu>.

CS234: Reinforcement Learning Winter 2025 - Final Project

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Overcoming Exploration in Reinforcement Learning with Demonstrations, February 2018. URL http://arxiv.org/abs/1709.10089. arXiv:1709.10089 [cs].

Ren, J., Swamy, G., Wu, Z. S., Bagnell, J. A., and Choudhury, S. Hybrid Inverse Reinforcement Learning, June 2024. URL http://arxiv.org/abs/2402.08848. arXiv:2402.08848 [cs].

Song, Y., Zhou, Y., Sekhari, A., Bagnell, J. A., Krishnamurthy, A., and Sun, W. Hybrid RL: Using Both Offline and Online Data Can Make RL Efficient, March 2023. URL http://arxiv.org/abs/2210.06718. arXiv:2210.06718 [cs].