Department of Computer Science and Engineering

University of Moratuwa

The Effectiveness of Proximal Policy Optimization in Multi-Agent Reinforcement Learning

Module: CS4681 - Advanced Machine Learning

Project ID: RL006

Domain: Reinforcement Learning

Research Area: Multi-Agent RL

Project Title / Sub-area: PettingZoo MAPPO

Benchmark Dataset: PettingZoo

SOTA Model: MAPPO

Supervisor: Dr. Uthayasanker Thayasivam

210404F - Nanayakkara A.H.M.

Table of Contents

Literature Review

Multi-agent reinforcement learning (MARL) is a rapidly evolving field focused on developing algorithms that enable multiple agents to learn and interact within a shared environment. A central debate in this domain revolves around the efficacy of on-policy versus off-policy algorithms, with many researchers historically favoring off-policy methods due to their perceived superior sample efficiency. However, a series of the recent papers challenges this conventional wisdom, demonstrating that the on-policy Proximal Policy Optimization (PPO) algorithm is not only a viable but a highly effective and the robust baseline in cooperative MARL.

The Foundational Work: The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games

The seminal paper by Yu et al. (2022) [1] directly confronts the notion that PPO is less suitable for multi-agent settings. The authors argue that PPO's underutilization in MARL is often based on the assumption of its poor sample efficiency compared to the off-policy counterparts like QMix and MADDPG. To empirically test this hypothesis , the researchers conducting an extensive study across four prominent cooperative multi agent testbeds,

- The multi-agent particle-world environments (MPE)
- The StarCraft multi-agent challenge (SMAC)
- Google Research Football
- The Hanabi challenge

The core finding of this work is that PPO based multi agent algorithms achieve surprisingly strong performance in these complex environments. The study introduced two primary PPO variants:

- MAPPO (Multi-Agent PPO): A variant employing a centralized value function that utilizes global state information for training.
- IPPO (Independent PPO): A decentralized approach where each agent uses PPO independently with only local observations.

Through rigorous experimentation, the authors demonstrated that MAPPO, in particular, was not only competitive with but often superior to its off-policy rivals in both final returns and sample efficiency. For example, in the Google Research Football domain , MAPPO consistently outperformed QMix across all the tested scenarios. The paper concludes by providing a set of concrete, practical suggestions for implementing PPO-based algorithms, including the use of value normalization and careful hyperparameter tuning for the clipping and the training epochs. This research firmly establishing PPO as a powerful and simple benchmark for future cooperative MARL studies.

Expanding the Framework: Recent Advancements and Enhancements

Building upon foundational insights of Yu et al., the subsequent research has focused on the further optimizing and extending PPO for MARL. For instance, the paper "Improving Proximal Policy Optimization Algorithm in Interactive Multi-Agent Systems" by Shang et al. (2024) [2] tackles the one of PPO's known limitations: sample inefficiency. While Yu et al. (2022) [1] showed PPO could be sample-efficient, this paper aims to accelerate its training. Shang et al. propose two novel methods:

- Shared parameters: All agents share the same policy and value network parameters, allowing the learning process to benefit from the experiences of all agents.
- Shared trajectories: Agents share their collected experiences trajectories in a common memory buffer , which is then used to update the shared parameters.

The study demonstrates that these enhancements significantly speed up the convergence of PPO, particularly in interactive multi-agent systems. This work underscores a new direction of research , not merely proving PPO's effectiveness, but actively develops methods to make it even more efficient.

Similarly, the paper "JointPPO: Diving Deeper into the Effectiveness of PPO in Multi-Agent Reinforcement Learning" by Mao et al. (2024) [3] further validates the efficacy of PPO within the Centralized Training with Decentralized Execution (CTDE) paradigm. The paper extends PPO to a CTDE framework, creating a novel algorithm called JointPPO. By using PPO to directly optimize a joint policy, the method simplifies the MARL problem and allows for better coordination among agents. The authors prove that JointPPO outperforms strong baselines on the challenging StarCraft Multi-Agent Challenge (SMAC) testbed, providing further evidence that PPO is a scalable and powerful algorithm for cooperative tasks.

Conclusion

In conclusion, the research community is increasingly recognizing the robustness and surprising effectiveness of PPO in cooperative multi-agent systems. The work by Yu et al. (2022) [1] served as a critical turning point, empirically demonstrating that a well-implemented PPO can match or exceed the performance of more complex off-policy algorithms. Subsequent research, such as the papers by Shang et al. (2024) [2] and Mao et al. (2024) [3], has built on this foundation by introducing novel enhancements and new implementations that further solidify PPO's position. This collective body of work firmly establishes PPO as a leading baseline and a fertile ground for future research in the pursuit of the building effective and scalable multi-agent systems.

Methodology

This methodology outlines a research project focused on advancing cooperative multi-agent control within the domain of Reinforcement Learning. The project's core is to explore and enhance the Proximal Policy Optimization (PPO) algorithm, specifically its multi-agent variant, MAPPO, using the PettingZoo benchmark dataset. The central innovation lies in developing a novel policy optimization enhancement to improve the cooperative capabilities of the agents.

1. Problem Formulation

The project tackles the challenge of cooperative multi-agent control, where multiple agents must learn to collaborate to achieve a shared objective. The problem is framed within the Centralized Training with Decentralized Execution (CTDE) paradigm. In this setting, agents are trained in a centralized manner with access to global state information to facilitate coordination, but once deployed, they act independently based on their local observations. This approach addresses the issue of non-stationarity, which arises from each agent's policy changing during training.

2. Baseline Model and Dataset

- SOTA Model (Baseline): The SOTA model chosen for this project is MAPPO (Multi-Agent Proximal Policy Optimization). As demonstrated in a key foundational paper, MAPPO is a strong and robust baseline that, contrary to previous beliefs, exhibits surprising effectiveness in cooperative settings.
- Benchmark Dataset: All experiments will be conducted using the PettingZoo library, an open-source framework for MARL. PettingZoo provides a variety of cooperative environments, including the popular MPE (Multi-agent Particle-World Environments), which will serve as the primary testbed. This dataset allows for direct comparison and validation against established benchmarks.

3. Key Enhancement: Policy Optimization

This project's key enhancement is a targeted improvement to the policy optimization component of the MAPPO algorithm. The proposed innovation aims to modify the standard PPO clipped objective function to more explicitly encourage cooperative behaviors among agents. The modification will be designed to:

- Improve Credit Assignment: The proposed method will better attribute global rewards to individual agent actions, addressing the fundamental challenge of credit assignment in cooperative MARL.

- Enhance Communication (Implicitly): By optimizing for a cooperative objective, the agents will implicitly learn the communication protocols and coordination strategies without the need for an explicit communication channel.

4. Experimental Setup

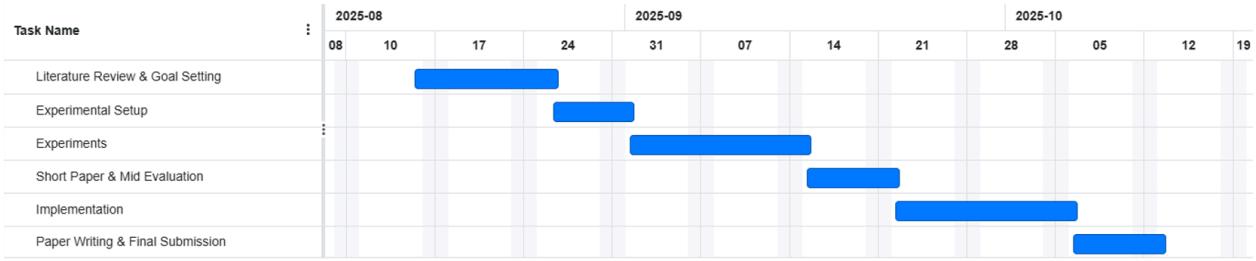
The methodology adheres to the rigorous experimental standards of the field to ensure reproducibility and valid comparisons.

- Environment: The MPE environments within the PettingZoo will be used for all experiments.
- Evaluation Metrics: Performance will be measured by the total team reward over time, with a focus on average episode returns and win rates, where applicable.
- Ablation Study: A critical part of the methodology will be a ablation study to isolate the impact of the proposed enhancement. The performance of the enhanced MAPPO will be compared against the standard MAPPO baseline.
- Hyperparameter Tuning: A systematic grid search will be conducted to finding optimal hyperparameters for the enhanced model, as per the guidelines in the original MAPPO research paper. The final paper will include a table detailing the chosen hyperparameters for all experiments.

5. Implementation

The entire methodology will be implemented in Python, utilizing those standard libraries for reinforcement learning such as PyTorch or TensorFlow, as well as the PettingZoo environment. The code will be fully documented, and a clean, organized Github repository will be maintained for code submission, as required by the assignment guidelines.

Project Timeline



Bibliography

[1] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *arXiv preprint arXiv:2103.01955*.

[2] Shang, Y., Chen, Y., & Cruz, F. (2024). Improving Proximal Policy Optimization Algorithm in Interactive Multi-Agent Systems. *2024 IEEE International Conference on Development and Learning (ICDL)*.

[3] Mao, Y., Wang, Z., & Chen, G. (2024). JointPPO: Diving Deeper into the Effectiveness of PPO in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2404.11831*.