Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

FEUDAL MADRL FRAMEWORKS: SYNCHRONIZING INDEPENDENT PPO AGENTS FOR SCALABLE MULTI-AGENT COORDINATION

Austin Starken[1]

[1]School of Modeling Simulation and Training, University of Central Florida, Orlando, FL, USA

ABSTRACT

Multi-agent deep reinforcement learning (MADRL) enables multiple agents to learn optimal strategies in intricate, ever-changing environments, utilizing deep reinforcement learning methodologies for cooperative and competitive scenarios. However, MADRL encounters challenges, including non-stationarity, shadowed equilibria, credit assignment issues, and communication overhead. This research draws on the hierarchical strategies of feudal reinforcement learning to integrate multiple independent Proximal Policy Optimization agents within a feudal framework to improve MADRL performance. Follower agents focus on learning specific tasks, while a leader agent synchronizes these tasks to achieve a unified goal. Initial findings demonstrate that this feudal approach improves training efficiency and scalability over traditional methods. These results indicate that feudal MADRL frameworks are well-suited for managing complex coordination tasks and hold promise for advancing research in more sophisticated environments and multi-team systems.

1 INTRODUCTION

Traditional simulation techniques fail to capture nuanced interactions and the adaptive behaviors required in complex *multi-agent systems*, often requiring human intervention, which can be expensive. *Multi-agent deep reinforcement learning* (MADRL) enables multiple fully automated agents to learn optimal strategies through interaction and feedback. MADRL can reduce costs and create more accurate simulations.

In cooperative MADRL, agents must learn to coordinate their actions to accomplish a common goal. However, MADRL is challenged by non-stationarity, shadowed equilibria, credit assignment, and communication overhead. This research addresses these challenges by extending the work of *feudal multi-agent hierarchies* (Ahilan & Dayan, 2019) by combining *feudal reinforcement learning* (FRL) (Dayan & Hinton, 1992) with *Proximal Policy Optimization* (PPO) (Schulman et al., 2017) to create *Feudal Leader Independent PPO* (FLIPPO) and answer the research question:

> *To what extent does a feudal hierarchy enhance the performance, scalability, adaptability, and generalizability of multiple independent PPO agents in high-dimensional environments compared to centralized training, centralized execution (CTCE), decentralized training, decentralized execution (DTDE), centralized training, decentralized execution with parameter sharing (CTDEPS) approaches?*

2 METHODOLOGY

PPO is a leading *deep reinforcement learning* (DRL) algorithm that has succeeded in single-agent DRL and MADRL environments (Yu et al., 2022). FRL is a hierarchical approach to DRL that involves managers and workers. FLIPPO arranges independent PPO models into a feudal hierarchy. The feudal leader accomplishes the environmental goal by designating an *area of responsibility* (AOR) and task to follower agents who receive a reward for completing their assigned task within the boundaries of their AOR.

A custom MADRL environment, inspired by the LOTZ optimization problem (Laumanns et al., 2002), was developed with a slight variation where success is measured by multiplying the leading ones and trailing zeros (LO*TZ). The LO*TZ state space contains over *7.5 billion* state-action pairs where agents

collaborate to transform a bit-string into an optimal configuration with the maximum number of leading ones and trailing zeros. The LO*TZ environment provides a controlled, simplified platform to develop and refine MADRL techniques, which can be applied to more complex environments.

Using the LO*TZ environment, FLIPPO was compared to CTCE, CTDE, and DTDE approaches. CTCE approaches included a single-agent PPO (SAPPO) and a single-model multi-agent (SMMA) approach. CTDE included a CTDE approach with parameter sharing, similar to MAPPO (Yu et al., 2022). DTDE included multiple SAPPO agents operating within the same environment. Each approach was trained and evaluated ten times to balance the need for thorough evaluation and resource constraints.

3 PRELIMINARY RESULTS

FLIPPO converged on an optimal policy with a mean (μ) = 1,729,930 training time steps (σ = 89,177), where all other approaches failed to learn any policy. During the evaluation phase, FLIPPO successfully solved $942/1000$ randomized bit strings with its deterministic model and $1000/1000$ with its non-deterministic model, achieving a higher average LOTZ score and significantly outperforming all baseline approaches (see Table 1).

Table 1: Preliminary Results

Approach	Model Type	Avg. LOTZ score	Solved Bit Strings (out of 1000)	Approach	Model Type	Avg. LOTZ score	Solved Bit Strings (out of 1000)
SAPPO	Deterministic	0.84 ± 2.13	0	CTDEPS	Deterministic	0.88 ± 2.53	0
	Stochastic	0.97 ± 2.78	0		Stochastic	0.83 ± 2.21	0
SMMA	Deterministic	0.86 ± 2.14	0	IPPO	Deterministic	0.99 ± 3.62	0
	Stochastic	0.92 ± 2.40	0		Stochastic	0.99 ± 3.01	0
FLIPPO	**Deterministic**	**186.06 ± 41.01**	**942**	**LOTZ Score, Kruskal Stat., p-value**	**Deterministic**	**2991.32, $p < .001$**	**Significant Difference Across Groups**
	Stochastic	**196 ± 0.0**	**1000**		**Stochastic**	**3164.68, $p < .001$**	

4 CONCLUSION

This study shows the effectiveness feudal hierarchies to enhance the performance of a multi-PPO agent system. FLIPPO's hierarchical framework simplifies individual agent learning tasks, fosters improved coordination, and significantly boosts system efficiency and scalability. By addressing non-stationarity, shadowed equilibria, and credit assignment, the leader coordinates follower agent actions to stabilize the learning environment, ensuring consistent progression toward equilibrium.

FLIPPO highlights the importance of hierarchical frameworks in the ongoing development of MADRL. By integrating more sophisticated communication protocols and adaptive learning mechanisms, FLIPPO's hierarchical approach can potentially set new benchmarks in coordination efficiency and performance across a broader range of MARL applications, particularly in dynamic and high-dimensional environments. Future research will focus on refining these hierarchical structures, exploring their adaptability to more complex environments, and further enhancing their scalability.

REFERENCES

Ahilan, S., & Dayan, P. (2019). *Feudal multi-agent hierarchies for cooperative reinforcement learning.* https://doi.org/10.48550/ARXIV.1901.08492

Dayan, P., & Hinton, G.E. (1992). Feudal reinforcement learning. In S. Hanson, J. Cowan, & C. Giles (eds.), *Advances in Neural Information Processing Systems* (Vol. 5, pp. 271-278). Morgan-Kaufmann. https://doi.org/10.5555/645753.668239

Laumanns, M., Thiele, L., Zitzler, E., Welzl, E., & Deb, K. (2002). Running time analysis of multi-objective evolutionary algorithms on a simple discrete optimization problem. In J. J. M. Guervós, P. Adamidis, H.-G. Beyer, H.-P. Schwefel, & J. L. Fernández-Villacañas (Eds.), *Parallel Problem Solving from Nature—PPSN VII* (Vol. 2439, pp. 44–53). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45712-7_5

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. https://doi.org/10.48550/ARXIV.1707.06347

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. *Advances in Neural Information Processing Systems*, *35*, 24611–24624.