Advancing Sample Efficiency and Explainability in Multi-Agent Reinforcement Learning

Doctoral Consortium

Zhicheng Zhang
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
zhichen3@cs.cmu.edu

ABSTRACT

Multi-Agent Reinforcement Learning (MARL) holds promise for complex real-world applications but faces challenges in sample efficiency and policy explainability. My dissertation aims to address these critical barriers, advancing MARL towards more practical and interpretable systems. To boost sample efficiency, it is crucial for agents to effectively learn from and generalize past experiences. We propose a meta-exploration technique to train meta-exploration policies that exploit the joint state-action space structure from meta-training tasks. This approach can be integrated with any off-policy MARL algorithm to improve learning efficiency. Complementing the efficiency gain, my research also focuses on augmenting the explainability of neural network policies' decision-making processes using techniques such as decision-tree extraction from MARL networks. In this extended abstract, I will summarize my research so far and outline promising future directions to further the deployability of MARL in complex real-world environments.

KEYWORDS

Multi-Agent Reinforcement Learning; Sample efficiency; Explainability

ACM Reference Format:
Zhicheng Zhang. 2024. Advancing Sample Efficiency and Explainability in Multi-Agent Reinforcement Learning: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

In Multi-agent Reinforcement Learning (MARL), two pivotal challenges stand out: enhancing sample efficiency and improving explainability. My dissertation seeks to bridge these gaps and aims to contribute to the development of MARL systems that can be practically deployed in complex real-world environments by making them more efficient in learning from past experiences and more transparent and interpretable to humans. This extended abstract delves into my research so far in these areas, including a meta-exploration method for learning efficiency and one decision-tree extraction method for multi-agent policy explainability.

To enhance sample efficiency in MARL, we first ask: how can we enable agents in MARL to explore their environment more effectively and enhance sample efficiency, especially when faced with sparse reward signals? To tackle this question, the first aspect of my research introduces an innovative approach *Cooperative Meta-Exploration in Multi-Agent Learning through Exploiting State-Action Space Structure* (MESA) [13]. MESA arises from the observation that in multi-agent scenarios, the sheer scale of the joint state-action space can make conventional exploration methods, which often rely on encouraging visits to less frequented and hence more novel states, increasingly inefficient. To alleviate this issue in multi-agent settings, MESA employs a meta-exploration framework that first identifies a high-reward joint state-action subspace. Then MESA trains a set of diverse exploration policies to sufficiently cover this identified subspace using a reward scheme that is based on the proximity to the high-rewarding regions. These meta-exploration policies can afterward be combined with any off-policy MARL algorithm to facilitate learning during the meta-testing phase. Empirical experimental results show that MESA performs significantly better in both low-dimensional matrix games and high-dimensional multi-agent environments.

While MESA focuses on addressing the learning efficiency problem in MARL, how can we also make sure that the decision-making process of RL policies is transparent and interpretable to humans? In many multi-agent reinforcement learning applications, such as air traffic control [3], cyber defense 22 [9], and autonomous driving [2], the real-world risk necessitates learning interpretable policies that people can inspect and understand before deployment. Therefore, the second stream of my research focuses on learning interpretable policies for MARL. We look at decision-tree policies, which are considered to be an intrinsically interpretable model familiy [10]. We develop IVIPER and MAVIPER that extract decision-tree policies from MARL-trained neural networks. These algorithms enable us to interpret the underlying decisions made by agents, with a particular focus for MAVIPER on coordination and collaboration happening in the multi-agent joint policies.

By synthesizing the strength of the two streams of my research, my dissertation aims to further the deployability of MARL in complex real-world environments. Moving forward, there is a wealth of potential, as outlined at the end of this extended abstract, in further refining these methods and exploring new avenues to fully unlock the capabilities of MARL.

2 MULTI-AGENT META-EXPLORATION

In this section, we delve into the Meta-Exploration in Multi-Agent Learning through Exploiting State-Action Space Structure (MESA)

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

method, which aims to improve cooperative multi-agent learning. MESA consists of two stages: meta-training and meta-testing.

The meta-training stage serves a dual purpose: identifying high-rewarding state-action subspaces and training a diverse set of exploration policies utilizing these rewards. This stage is executed in two steps. Initially, we discern the high-rewarding joint state-action subspace by accumulating experiences and storing high-reward joint state-action pairs into a stored dataset. Subsequently, we train the exploration policies to “cover" the identified high-rewarding joint state-action subspace, employing a distance metric to ascertain the proximity of state-action pairs. An important aspect of the meta-training stage is addressing the reward sparsity problem. To mitigate this issue, we assign positive rewards to specific joint state-action pairs that subsequently lead to valuable pairs within the trajectory. To encourage a broader coverage of the subspace and to avoid mode collapse, the reward assignment scheme ensures that repeated visits to similar joint state-action pairs within one trajectory would result in a decreasing reward for each visit.

In the meta-testing stage, MESA capitalizes on the meta-learned exploration policies to facilitate learning in previously unseen tasks. The exploration policies are deployed in an annealing schedule, supplying a greater quantity of exploration rollouts during the initial stages of training and gradually decreasing thereafter. This diminution strategy ensures the efficacy of the exploration process during the early stages of training when it is most crucial.

In conclusion, the MESA method addresses the challenges inherent in cooperative multi-agent learning through a two-stage process. MESA pinpoints high-rewarding state-action subspaces and trains meta-exploration policies during meta-training to effectively tackle the reward sparsity problem.

3 LEARNING INTERPRETABLE DECISION-TREE POLICIES

To learn interpretable policies for MARL, we focus on decision tree models as the policy representation due to their intrinsic interpretability [10]. In this work, we aim to extract decision-tree policies from neural network policies learned with MARL algorithms. Specifically, we adopt the DAgger framework [11] to iteratively collect new samples on which to query expert labels. We propose two algorithms: IVIPER and MAVIEPR. IVIPER is the direct extension of VIPER [1] to the mutli-agent settings, while MAVIPER focuses more on extracting decision-tree policies that address the issue of coordination.

The MAVIPER algorithm adds a prediction module that predicts what the tree would be for the other agents, so that when building one tree, the current status of the other trees can be taken into account. Additionally, MAVIPER designs a new weighting scheme in the multi-agent settings to weigh the samples collected by the DAgger framework with the expected difference in Q values by taking a suboptimal action. This new weighting scheme arrives from the observation that agents should focus more on the critical states where a good joint action can make a difference.

Empirically, we find that the learned decision trees perform relatively well in three multi-agent particle environment tasks [8], and that MAVIPER leads to an increase in coordinated joint performance comparing to the baselines and IVIPER.

4 FUTURE DIRECTIONS

The future directions of my dissertation still revolve around the topic of improving sample efficiency and enhancing the explainability of the model. One important future research direction would be to study the potential utilization of large language model (LLM) in MARL. Specifically, LLMs, due to their diverse ability, can be a great addition in the following aspects.

(1) *Sample Efficiency.* The rise of LLM-based agents offers the hope of using LLM as a starting point for designing AI agents that can adapt to diverse scenarios [12]. LLMs also show remarkable reasoning capabilities and the ability to work as a world model [6]. These preliminary results are exciting directions to further improve the sample efficiency of the reinforcement learning agents since the agent training would not be from scratch, but rather from an agent that already has abundant information about the environment as a prior. Particularly, in multi-agent reinforcement learning, these capabilities of LLM and LLM-based agents would be even more useful for achieving intelligent exploration, human-level cooperation and coordination, etc.

(2) *Explanability.* Using decision-tree policies as a form of interpretable policies has a lot of merits, but the approach could be made more general by looking at model transforms [5] or natural language as forms of explanations [4, 7]. In my research, I have obtained preliminary results that demonstrate the effectiveness of using LLM for enhancing explainability. Firstly, we developed three user-friendly explanation methods utilizing Large Language Models (LLMs) to generate interpretable insights for volunteers to better understand the task difficulty predictions. The models and findings from the complete study are in the process of being adopted at Food Rescue Hero, a large food rescue platform serving over 25 cities across the United States. Secondly, we are employing LLMs to generate explanations for the decision-making process of chess AI by grounding the output on the Monte Carlo search tree, which the AI utilizes to determine the suggested move.

5 CONCLUSION

In conclusion, I believe that advancing sample efficiency and explainability would be critical in the deployment of (multi-agent) reinforcement learning methods in the real world. I have developed methods to help combat these two pivotal issues, and I hope that the developed methods as well as future work in my dissertation can help further the possibility of deployment of such MARL policies.

ACKNOWLEDGMENTS

This research is supported in part by NSF IIS-2046640 (CAREER) and Sloan Research Fellowship. We thank NVIDIA for providing computing resources. Zhicheng Zhang is supported in part by SCS Dean’s Fellowship. The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. 2018. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems* 31 (2018).

[2] Sushrut Bhalla, Sriram Ganapathi Subramanian, and Mark Crowley. 2020. Deep multi agent reinforcement learning for autonomous driving. In *Canadian Conference on Artificial Intelligence*. Springer, 67–78.

[3] Marc Brittain and Peng Wei. 2019. Autonomous air traffic controller: A deep multi-agent reinforcement learning approach. *arXiv preprint arXiv:1905.01303* (2019).

[4] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic generation of natural language explanations. In *Proceedings of the 23rd international conference on intelligent user interfaces companion.* 1–2.

[5] Mira Finkelstein, Lucy Liu, Yoav Kolumbus, David C Parkes, Jeffrey S Rosenschein, Sarah Keren, et al. 2022. Explainable Reinforcement Learning via Model Transforms. *Advances in Neural Information Processing Systems* 35 (2022), 34039–34051.

[6] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992* (2023).

[7] Sawan Kumar and Partha Talukdar. 2020. NILE: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116* (2020).

[8] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

[9] Kleanthis Malialis and Daniel Kudenko. 2015. Distributed response to network intrusions using multiagent reinforcement learning. *Engineering Applications of Artificial Intelligence* 41 (2015), 270–284.

[10] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[11] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 627–635.

[12] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).

[13] Zhicheng Zhang*, Yancheng Liang*, Yi Wu, and Fei Fang. 2024. MESA: Cooperative Meta-Exploration in Multi-Agent Learning through Exploiting State-Action Space Structure. In *Accepted to Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems.*