Multi-Expert Distillation for Few-Shot Coordination (Student Abstract)

Yujian Zhu*, Hao Ding*, Zongzhang Zhang†

National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{zhuyj, dingh}@lamda.nju.edu.cn, zzzhang@nju.edu.cn

Abstract

Ad hoc teamwork is a crucial challenge that aims to design an agent capable of effective collaboration with teammates employing diverse strategies without prior coordination. However, current Population-Based Training (PBT) approaches train the ad hoc agent through interaction with diverse teammates from scratch, which suffer from low efficiency. We introduce Multi-Expert Distillation (MED), a novel approach that directly distills diverse strategies through modeling across-episodic sequences. Experiments show that our algorithm achieves more efficient and stable training and has the ability to improve its behavior using historical contexts. Our code is available at https://github.com/LAMDA-RL/MED.

Introduction

Reinforcement Learning (RL) is a fundamental paradigm of learning from evaluative feedback for building intelligent decision-making agents (Littman 2015; Wu and Zhang 2023). Multi-Agent RL (MARL), as a natural extension of RL for learning agents in multi-agent systems, has attracted a broad attention in recent years. Current MARL approaches face the challenge of generalizing to multiple tasks (Zhang et al. 2023) or various teammates (Ding et al. 2023). Ad hoc teamwork (Stone et al. 2010) refers to the challenge of training a generalist agent capable of efficient collaboration with diverse teammates without prior coordination. In this paper, we explore an extended variant of ad hoc teamwork named "few-shot coordination", where the generalist agent is required to establish cooperation with previously unseen partners and progressively adapt its behavior throughout multiple continuous episodes.

One widely adapted method towards the problem is Population-Based Training (PBT). Initially, a population is generated consisting of diverse cooperative strategies. Subsequently, the generalist agent interacts with agents from the population and undergoes training using RL. However, this method requires the generalist agent to learn from scratch through trial and error, preventing it from acquiring coordination knowledge from the population. This leads to reduced efficiency and instability in training. Additionally, the generalist agent lacks the ability to adapt its behavior based on interaction histories.

To tackle these limitations, we present a novel method that facilitates faster and more stable training while enabling the generalist agent to adapt its policy to accommodate its teammates. Rather than RL from scratch, we utilize the policies within the population not only as teammates but also as experts who provide advised actions for coordination. And we train the generalist through supervised learning with expert actions as targets. To enable the generalist agent to adjust its behaviors, we employ transformer as our policy model and sample sequences comprising multiple game episodes for training. We evaluate our approach in various scenarios that emphasize distinct cooperation styles and achieve exceptional performance.

Approach

We consider the problem in a two-player Dec-POMDP, defined by $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{O}^i\}_{i=1}^2, \{\mathcal{A}^i\}_{i=1}^2, \Omega, P, R \rangle$, where $\mathcal{N} = \{1, 2\}$ is the set of agents, $\mathcal{S}$ is the state space. At each step, agent $i \in \mathcal{N}$ observes $o^i \in \mathcal{O}^i$ with observation function $\Omega(s, i)$ and takes action $a^i \in \mathcal{A}^i$. The environment then transits from s to s' with probability $P(s'|s, a^1, a^2)$ and returns a shared reward $r = R(s, a^1, a^2)$ to each agent.

Our method consists of two parts. The initial step involves generating a population consisting of diverse cooperative strategies, which is the same with other PBT methods. Once the population is generated, MED lets the generalist agent interact with agents from the population in the environment, sampling across-episodic trajectories. Simultaneously, the corresponding agent within the population generates the advised action for the generalist to adopt. Subsequently, we employ the advised actions as labels and train the generalist through supervised learning.

Generating Population Following the common PBT paradigm, a population of diverse and high-quality teammates is necessary. The population is defined as a set of joint-policies $\Pi = \{(\pi_i^1, \pi_i^2)\}_{i=1}^n$. We apply a recent proposed method LIPO (Charakorn, Manoonpong, and Dilokthanakul 2022) to generate the diverse teammates population for MED and all the baselines for a fair comparison.

*These authors contributed equally.
†Corresponding author: Zongzhang Zhang.

PBT on Across-episodic Trajectories For each joint-policy (π_i^1, π_i^2) in the population, we substitute one member π_i^1 with the generalist policy π_{M} to sample trajectories. As π_i^1 and π_i^2 are jointly trained to maximize performance, the optimal response to π_i^2 is already contained within π_i^1. Thus, inspired by recent success in distillation (Zheng et al. 2021), we treat π_i^1 as an expert and let π_{M} distill its policy.

During sampling, we collect across-episodic trajectories, denoted as $h_t = (o_0, a_0, r_0, ..., o_t, a_t, r_t)$. Concurrently, we acquire the advised actions $(a'_0, a'_1, ..., a'_t)$ using π_i^1. Employing the history $(h_{<t}, o_t)$ as the feature and the advised action a'_t as the label, MED conducts supervised training by minimizing the cross-entropy loss function:

$$\mathcal{L} := -\sum_{t=0}^{T} \log \pi_{\mathrm{M}}\left(a'_t | h_{<t}, o_t\right), \tag{1}$$

where T is the length of the whole across-episodic sequence.

We employ transformer as our policy model. It takes historical contexts $(h_{<t}, o_t)$ as input and generates a predicted distribution for the next action $\pi_M(a_t|h_{<t}, o_t)$. Because of the sequence modeling capabilities of transformer, our algorithm can analyze teammate's cooperation style and align its behavior with that of the corresponding expert.

Experiments

To evaluate the effectiveness of our algorithm, we perform experiments on three environments that emphasize coordination and policy diversity: (i) One-Step Cooperative Matrix Game; (ii) Gridworld MoveBox, where two agents collaborate to move a box to one of the eight exits; and (iii) Overcooked, where agents have to prepare and serve one of the six recipes as fast as possible. A detailed introduction about these environments can be found in the appendix.[1]

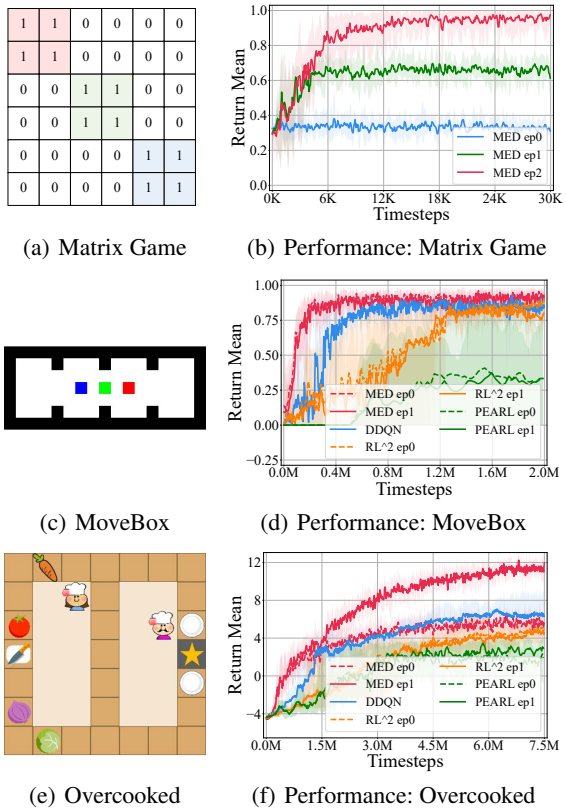
One-Step Cooperative Matrix Game The payoff matrix in the game is shown in Fig. 1(a). It presents three distinct groups of solutions, making cooperation among agents impossible without prior information. Nevertheless, our generalist agent can invest episodes in exploring the teammate's strategy, thereby enabling coordination. As shown in Fig. 1(b), in the initial episode, the agent is limited to selecting a random action, resulting in an average return of $\frac{1}{3}$. However, in subsequent episodes, the MED agent can eliminate incorrect actions and increase the expected return.

Gridworld MoveBox and Overcooked To investigate the effectiveness of MED in more complex settings, we conduct experiments in these two environments and compare MED to methods that employ RL from scratch: (i) DDQN (Van Hasselt, Guez, and Silver 2016); (ii) RL^2 (Duan et al. 2016); and (iii) PEARL (Rakelly et al. 2019).

As can be seen in Fig. 1(d) & 1(f), MED demonstrates remarkable learning speed, requiring far fewer total timesteps in sampling. This supports our claim that directly distilling experts' policies from the population is a considerably more efficient approach compared to starting from scratch.

It is worth noting that the training curves of the baseline methods exhibit significant variation under different random seeds. In contrast, the training process of MED exhibits significantly greater stability compared to that of the baselines.

Finally, we evaluate the performance of MED and examine whether it possesses the capability to leverage historical contexts for enhancing coordination in complex scenarios. Fig. 1(f) illustrates that while MED already has comparable results to all baselines in the first episode, its performance experiences a remarkable surge in the second episode. This provides evidence that MED can adapt its behavior based on interaction histories and enhance cooperation.



(a) Matrix Game (b) Performance: Matrix Game

(c) MoveBox (d) Performance: MoveBox

(e) Overcooked (f) Performance: Overcooked

Figure 1: Environments and the training timesteps-mean return curves on them. The 'ep' stands for different episodes.

Conclusion and Future Work

This paper presents a novel approach to achieve few-shot coordination by distilling policies from a diverse set of experts into a transformer model. The experiments show the effectiveness of our algorithm. For future work, we may focus on integrating the benefits of supervised learning and reinforcement learning in ad hoc teamwork settings.

Acknowledgments

This work is supported by the National Science Foundation of China (No. 62276126) and the Natural Science Foundation of Jiangsu (No. BK20221442).

[1] https://www.lamda.nju.edu.cn/zhuyj/MEDAppendix.pdf

References

Charakorn, R.; Manoonpong, P.; and Dilokthanakul, N. 2022. Generating Diverse Cooperative Agents by Learning Incompatible Policies. In *ICLR*.

Ding, H.; Jia, C.; Guan, C.; Feng, C.; Yuan, L.; Zhang, Z.; and Yu, Y. 2023. Coordination Scheme Probing for Generalizable Multi-Agent Reinforcement Learning.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv preprint arXiv:1611.02779*.

Littman, M. L. 2015. Reinforcement Learning Improves Behaviour from Evaluative Feedback. *Nature*, 521: 445–451.

Rakelly, K.; Zhou, A.; Finn, C.; Levine, S.; and Quillen, D. 2019. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *ICML*, 5331–5340.

Stone, P.; Kaminka, G. A.; Kraus, S.; and Rosenschein, J. S. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. In *AAAI*, 1504–1509.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep Reinforcement Learning with Double Q-Learning. In *AAAI*, 2094–2100.

Wu, C.; and Zhang, Z. 2023. Surfing Information: The Challenge of Intelligent Decision-Making. *Intelligent Computing*, 2: 1–7.

Zhang, F.; Jia, C.; Li, Y.; Yuan, L.; Yu, Y.; and Zhang, Z. 2023. Discovering Generalizable Multi-agent Coordination Skills from Multi-task Offline Data. In *ICLR*.

Zheng, Y.; Hao, J.; Zhang, Z.; Meng, Z.; Yang, T.; Li, Y.; and Fan, C. 2021. Efficient Policy Detecting and Reusing for Non-Stationarity in Markov Games. *Autonomous Agents and Multi-Agent Systems*, 35: 1–29.