A Novel Multi-Agent Deep Reinforcement Learning Approach

Dong Yin¹, Zhe Zhao¹, Yinglong Dai²,³, Han Long²,*

¹ College of Intelligence Science and Technology, National University of Defense Technology, Changsha, 410073, China
² College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, 410073, China
³ Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, 410081, China

*lonyhan@nudt.edu.cn

Abstract: Borrowing the power of deep neural networks, deep reinforcement learning achieved big success in games, and it becomes a popular method to solve the sequential decision-making problems. However, the success is still restricted to single agent training environment. Multi-agent reinforcement learning still is a challenge problem. Although some multi-agent deep reinforcement learning methods have been proposed, they can only perform well when the number of agents is very limited. In this paper, by analyzing the dynamic changing observation space and action space of multi-agent environment, we propose a novel multi-agent deep RL method that compress the joint observation space and action space as the time goes on. The proposed method is potential for a large number of agents cooperative or competitive tasks.

Keywords: Multi-agent, Deep Reinforcement Learning, Decision Space Compression, Decision-making

1. Introduction

Since Mnih et al. [1] proposed deep Q-network (DQN) by combining Q-learning with deep neural networks that achieved human-level control in Atari games, deep reinforcement learning (deep RL) become a popular research hotspot in sequential decision-making problems. Many advanced algorithms have been proposed, such as deep deterministic policy gradient (DDPG) [2] proposed for continuous action space training, asynchronous advantage actor-critic (A3C) [3] for concurrent training, proximal policy optimization (PPO) [4] for robust training. Deep RL methods have also been successfully applied in various fields beyond games, such as robotics [5], natural language processing [6], recommender system [7], and healthcare [8]. Nevertheless, the successes of deep RL methods are confined to single agent learning problems. Multi-agent systems are much more prevalent than single-agent system in real world problems. Therefore, deep RL methods would have much more potential applications if they could manage multi-agent systems.
Recently, researchers start to pay attention to multi-agent deep RL. Multi-agent deep RL is much harder because the environment is no longer stationary. Individual agent has to interact with other agents and the historical experience may not be trusted in the future so that will lead the training process become unstable. To solve the multi-agent credit assignment problem, Foerster et al. [9] proposed counterfactual multi-agent policy gradients (COMA) that computed a counterfactual baseline by marginalizing out an agent’s action while fixing other agents’ action. Lowe et al. [10] proposed a multi-agent deep RL method based on DDPG, called MADDPG. Mao et al. [11] also extended DDPG by introducing an attention mechanism in a principled way and presented ATT-MADDPG to enhance the centralized critic network. They used a strategy of centralized training with decentralized execution, which made the training process more efficient and stable than distributed training. However, their scalability is still very limited that can only be applied to several agents. The performance of the multi-agent deep RL methods will degenerate sharply when the joint observation space and action space become very large as the number of agents increases.

In this paper, we present a novel approach for multi-agent deep RL according to the dynamic changing joint observation space and action space. Because the joint observation space and action space of multiple agents can be shrunk as the training process goes on, we propose to model the multi-agent system to be dynamic observation space and action space. Then, we can contract the search space in the training process to promote the multi-agent training efficiency. As the contraction of the exploratory space, the method also can make the training process become stable.

The remainder of this paper will be organized as follows. First, we have some preliminary knowledge about the multi-agent reinforcement learning problem in Section 2. Then, we discuss the dynamic joint observation space and action space modeling method in Section 3. And we present how to contract the search space of multi-agent system to promote the training efficiency in Section 4. Finally, we conclude the paper in Section 5.

2. Problem of Multi-Agent System

As we know, reinforcement learning can be formalized as Markov decision process (MDP). The deep reinforcement learning framework is illustrated in Fig.1. The agent is to learn a policy \( \pi_\theta(a_t|o_t) \), which typically uses deep neural networks (DNNs) to approximate the function. In the training phase, the agent interacts with the environment represented as a transition model \( p(o_{t+1}|a_t, o_t) \), which will change its observation to \( o_{t+1} \) in the case of the previous observation \( o_t \) and action \( a_t \) at time \( t \). For the agent, it will receive a reward \( r_t \) at time \( t \), and the target of it is to obtain a maximum return \( R = \sum r_t \) in the future. When the observation can only partially represent the environmental state, the problem is known as partially observable Markov decision process (POMDP).

![Fig.1 The illustration of deep reinforcement learning framework](image)

The multi-agent framework is illustrated in Fig.2. The problem of multi-agent reinforcement learning belongs to decentralized partially observable Markov decision process (Dec-POMDP), which is a much harder problem than the single-agent reinforcement learning problem. Similar to the MDP formalization, Dec-POMDP can also be defined by a tuple

\[
M = \langle D, O, A, T, R \rangle,
\]
where $D = \{1, \ldots, n\}$ is the set of $n$ agents, $O = \{o^{(1)}, \ldots, o^{(n)}\}$ is the set of joint observation space, $A = \{a^{(1)}, \ldots, a^{(n)}\}$ is the set of joint action space, $T$ is the transition function, and $R$ is the reward function. The main difficulty of Dec-POMDP is that the environment is non-stationary for each agent. From the perspective of an agent, the other agents are also included in the environment. And the policies of the agents will continually change to impact on the environment. As a result, the state transition model with respect to each agent becomes non-stationary. Due to the complexity of multi-agent system, totally distributed training is typically infeasible.

3. Dynamic Joint Observation Space and Action Space Contraction

The strategy of centralized training with decentralized execution promotes the efficiency of the training process to some extent. However, the joint observation space and action space are still too large to efficiently train the multi-agent system when the number of agents increases. We conceive that the joint observation space and action space can be shrunk as the training process goes on. The sequential joint observation space and action space can be represented as $\langle O_1, \ldots, O_T \rangle$ and $\langle A_1, \ldots, A_T \rangle$ respectively. And the observation space and action space will be monotonically shrunk as

$$|O_1| \geq \gamma_1 |O_2|, \ldots, |O_{T-1}| \geq \gamma_1 |O_T|,$$

(1)

$$|A_1| \geq \gamma_2 |A_2|, \ldots, |A_{T-1}| \geq \gamma_2 |A_T|.$$

(2)

As the exploration space shrinks as Eqs. (1) and (2), the training process will accelerate gradually. In order to analyze the accelerate rate of training, we assume the joint observation space is shrunk by a ratio of $\gamma_1 \in (0,1)$, and the joint action space is shrunk by a ratio of $\gamma_2 \in (0,1)$,

$$|O_1| \geq \gamma_1 |O_2|, \ldots, |O_T| \geq \gamma_1 |O_{T-1}| \geq \gamma_1 |O_{T-1}| \geq \gamma_1 |O_T|,$$

(3)

$$|A_1| \geq \gamma_2 |A_2|, \ldots, |A_T| \geq \gamma_2 |A_{T-1}| \geq \gamma_2 |A_{T-1}| \geq \gamma_2 |A_T|.$$

(4)

Where $|$ denotes an operator of set dimension. According to Eqs. (3) and (4), the dimension of observation space $|O_t|$ and action space $|A_t|$ at training epoch $t$ will less or equal than initial observation space $|O_1|$ and action space $|A_1|$ to a certain scale, expressed as Eqs. (5) and (6) respectively.

$$|O_t| \leq \gamma_1^t |O_1|,$$

(5)

$$|A_t| \leq \gamma_2^t |A_1|.$$

(6)

In other words, the search space will be shrunk exponentially with respect to training epoch $t$. Therefore, the training process will converge quickly. And the training process can be accelerated, because the time cost of training process is proportionate to the scale of search space.
4. A Novel Multi-Agent Deep Reinforcement Learning Approach

Based on the analysis of dynamic joint observation space and action space, we propose a novel approach to dynamically contract the joint observation space and action space in the multi-agent training process. As we can see, multi-agent deep RL can achieve exponential convergence and training acceleration if the joint observation space and action space meet Eqs. (3) and (4) as the training process goes on.

To contract the search space of the multiple agents as the increasing training epochs, a simple method is to fix a portion of dimensions of observation space and action space after some extent of training exploration. Assume a multi-agent system with initial joint observation space $O_{t=1}$ and action space $A_{t=1}$, the main learning procedure can be presented as following steps.

**Step 1:** Explore in the joint observation space $O_t$ and action space $A_t$.

**Step 2:** Discover a part of nonsignificant dimensions or optimized dimensions of observation space $O_t$ and action space $A_t$, and fix the values of the dimensions to form new observation space $O_{t+1}$ and action space $A_{t+1}$.

**Step 3:** Repeat Step 1 and Step 2 to form new contracted observation space and action space until the actions of all the agents converge.

In the learning procedure, it may be a hard task to reduce the search space according to a scaling factor as Eqs. (5) and (6). Sometimes, the nonsignificant dimensions or optimized dimensions are difficult to judge, and we cannot guarantee some dimensions are optimized in combinatorial optimization problem. Therefore, we have to tradeoff between the training efficiency and the optimization performance. For instance, we must fix some dimensions of search space to promote the training efficiency even if we cannot ensure their importance in the future exploration. For this novel multi-agent deep reinforcement learning approach, there would be a great deal of work left to be done.

5. Conclusion and Future Work

For the multi-agent systems, deep RL methods face a big challenge of stability and scalability. The difficulty of multi-agent deep RL methods mainly comes from the huge dynamic exploration space of the observation space and action space combination of many agents. We analyzed the joint observation space and action space that could change dynamically as the training process goes on. Based on the analysis, we propose a novel multi-agent deep RL approach by contracting the joint observation space and action space. If the search space can be contracted by a factor at each training epoch, the multi-agent deep RL methods can achieve exponential convergence and training acceleration.

We believe that search space contraction will be an important approach for multi-agent deep RL to improve the training efficiency and stability. Nevertheless, it is still some technique problems in future work. For example, how to contract the search space at certain ratio that can guarantee the exponential convergence, and how to guarantee the agents performance bound when we restrict the search space of the agents.

Acknowledgments

This work is supported in part by Hunan Provincial Natural Science Foundation under Grant Number 2020JJ5367, Hunan Provincial Science and Technology Project Foundation under Grant Numbers 2018TP1018 and 2018RS3065.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski 2015 Human-level control through deep reinforcement learning Nature 518 (7540) pp529–533.

[2] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra 2016 Continuous control with deep reinforcement learning International Conference on Learning Representations pp529–533.

[3] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavuk-
cuoglu 2016 Asynchronous methods for deep reinforcement learning Proceedings of The 33rd International Conference on Machine Learning pp1928–1937

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov 2017 Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347

[5] S. Gu, E. Holly, T. Lillicrap and S. Levine 2017 Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates IEEE International Conference on Robotics and Automation pp 3389-3396

[6] J. Li, W. Monroe, A. Ritter, J. Dan, M. Galley and J. Gao 2016 Deep reinforcement learning for dialogue generation. Conference on Empirical Methods in Natural Language Processing pp1192-1202

[7] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang and Z. Li 2018 DRN: A Deep Reinforcement Learning Framework for News Recommendation. World Wide Web Conference pp167–176

[8] Y. Dai, G. Wang, K. Muhammad and S. Liu 2020 A closed-loop healthcare processing approach based on deep reinforcement learning Multimedia Tools and Applications

[9] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli and S. Whiteson 2018 Counterfactual multi-agent policy gradients Conference on Artificial Intelligence pp2974-2982

[10] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch 2017 Multi-agent actor-critic for mixed cooperative-competitive environments Proceedings of the 31st International Conference on Neural Information Processing Systems pp 6382–6393

[11] H. Mao, Z. Zhang, Z. Xiao and Z. Gong 2019 Modelling the Dynamic Joint Policy of Teammates with Attention Multi-agent DDPG. Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems pp1108–1116