Optimal Exploration Algorithm of Multi-Agent Reinforcement Learning Methods (Student Abstract)

Wang Qisheng,[1] **Wang Qichao,**[2] **Li Xiao***
[1]School of Information Engineering,Southeast University,Nan jing, 210096
qishengw@seu.edu.cn

Abstract

Exploration efficiency challenges for multi-agent reinforcement learning (MARL), as the policy learned by confederate MARL depends on the interaction among agents. Less informative reward also restricts the learning speed of MARL in comparison with the informative label in supervised learning. This paper proposes a novel communication method which helps agents focus on different exploration subarea to guide MARL to accelerate exploration. We propose a predictive network to forecast the reward of current state-action pair and use the guidance learned by the predictive network to modify the reward function. An improved prioritized experience replay is employed to help agents better take advantage of the different knowledge learned by different agents. Experimental results demonstrate that the proposed algorithm outperforms existing methods in cooperative multi-agent environments.

Introduction

Multi-agent reinforcement learning (MARL) coordinately uses concurrent exploration of the state-action space of the common environment in the early stage of RL to avoid all agents exploring the same or similar subspace simultaneously, which in turn will result in insufficient diversity of samples. If not constrained, the behaviour polices learned by different agents will lead to a unstable environment and seriously obstruct the convergence of the coordinated policy. Previous literatures such as MADDPG (Lowe et al. 2017) uses a centralized critic to deal with this problem allowing each agent to learn its own behavior strategy, which may overfit to other agents.

In this work, we proposes a novel communication method to guide MARL to improve the learning efficiency and use an extra network to obtain a more informative reward. Inspired by the tangible that many soldiers can effectively conduct military operations under the organization of army leader, we use a centralized critic network as a commander to coordinate behaviors of multiple RL agents, which means all agents share a Q-value network called Commander, and each maintains its own action network. Furthermore, a reward prediction network called PrecoderNet is used to relieve the problem of insufficient information on reward in RL[1]. We improve the prioritized replay buffer in (Schaul et al. 2015) by regarding the normalized priority as the agent's contribution weight to help them better communicate and avoid repeated exploration. The proposed multi-agent coordinative exploration algorithm with prioritized guidance is called MAEPG.

Algorithm and Process

We consider a system consisting of multiple joint actor networks $A_i{}_{i=1}^{M}$ with an improved priority experience replay R, a centralized PrecoderNet providing reward guidance σ and a critic network Commander as shown in Figure 1.

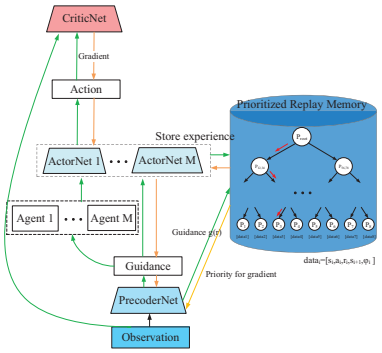

Figure 1: Architecture of MAEPG algorithm.

Partial observations from environment is fed into the PrecoderNet to generate the guidance σ_i for increasing the information of reward function to provide a more accurate ongoing direction and accelerate the exploration so that the exploration focuses on the effective subspace. ActorNet A_i se-

[1]Which is similar to the centralized critic in MADDPG, but our commander use guidance to predict the distribution of reward. Meanwhile, the action network of each agent perform different degrees of association according to the similarity of their experience pool priorities.

lects an action to interact with environment under the guidance σ_i, obtains feedback r_{t+1} and the new state s_{t+1} from the surrounding. Meanwhile, our Commander access the chosen action by the approximate Q-value $Q^C(s_t, a_t)$. The sample priority is shown in (1) by the TD-error and a discount factor η is used to determines how much the guidance is used to correct reward, i.e. $r_t = r_t + \eta\sigma_t^i$.

$$\varphi_t^i = Q^C(s_t, a_t) - r_{t+1} + \delta, \tag{1}$$

where the micro account $\delta > 0$ ensures a positive priority. Then φ_t^i is also stored in the latest transition in the prioritized experience replay. Since larger TD-error means greater contribution when conducting the gradient descent, the information , which is communicated between agents, in the learning process is further effectively increased.

The improved prioritized experience replay is stored in the form of "sum-tree" to improve sampling efficiency and save storage space. The lowest level leaf node stores the transition and priority while the remaining nodes only store the sum of the priority of their children nodes. Inspired by (Vezzani et al. 2019) and considering the tangible that visiting time of a sample can also reflect the importance of this specimen, we store the visiting time called ρ_t^i of the leaf node in the tree-structured experience replay and update the priority via

$$\varphi_t^i = \varphi_t^i + \frac{\rho_t^i}{\sum_j \rho_t^j}. \tag{2}$$

The Boltzman distribution of the *i-th* agent can be defined according to the sum of priority Φ_t^i stored in the root node of the experience replay to further take the different agents' contribution into consideration as $q_i = \frac{exp(\Phi_t^i)}{\sum_j exp(\Phi_t^i)}$.

Each agent uses (2) as weights to better improve the evaluation of Commanders. That is, the loss function of the Commander considering the contribution of different agents is (3) and (4):

$$L(\theta^C) = \frac{1}{M}\sum_{j=1}^{M} q_j (Q^C(s_t, a_t) - y_j)^2, \tag{3}$$

$$y_j = r + \gamma Q^{C'}(s_{t+1}, a_{t+1})|_{a_{t+1}=A'(s_{t+1})}. \tag{4}$$

Then we do the gradient descent to update on the PrecoderNet as shown in (5) where L represents the gradient:

$$L_{\theta^P} J(\theta^P) = E_{(s,a\sim R)}[L_{\theta^P}(C(s,a) - r(s,a)) L_{\theta^C} C(s,a)]. \tag{5}$$

The gradient flow from Commander through ActorNet and PrecoderNet, and the final gradient of Commander is the sum of the two gradient calculations of ActorNet and PrecoderNet. Multi-agent joint PrecoderNet and Commander's update will be more efficient and faster. To sum up, the more informative rewards combined with the improved priority experience replay leads to more coordinated communication and efficient explorations among MARL.

Simulation Result and Discussion

We verify the proposed algorithm in the RL environment Gym Pendulum-v0 and use DQN and MADDPG as benchmarks. We compare the performance of three algorithms when the number of agents $M = 2, 3$. As shown in Fig. 2, DQN is slow to converge while both MADDPG and MAEPG can converge, and MAEFG can converge in 110 episodes while MADDPG converges in about 250 episodes.

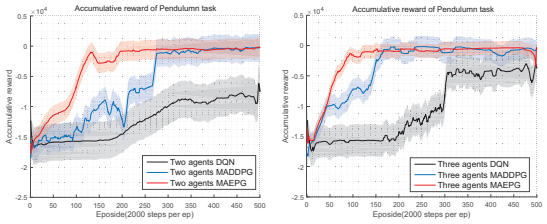

Figure 2: Convergence speed comparison among MAEPG, DQN and MADDPG with two and three agents.

It can be seen that the convergence speed of MAEPG is greatly improved with the increase of the number of agents and is significantly better than MADDPG in terms of stability. Thus, our MAEPG considers the importance of different agents and different samples to the overall learning progress in a multi-agent scenario, learning is faster, and at the same time, pendulum can be maintained in a vertically upward ideal position with a smaller swing. The experimental results prove that the proposed algorithm performs better than the benchmark under multi-agent collaborative exploration.

Conclusion

In this paper, we propose a novel cooperative MARL algorithm to achieve efficient and effective exploration by using knowledge learned by a centralized Commander and guidance perceived from previous experience. We assist the multiple agents to better communicate via improving the prioritized experience replay. The centralized PrecoderNet enriches the information of reward in RL tasks to accelerate the learning process in MARL. The experiments demonstrates that the proposed algorithm outperforms existing methods in cooperative multi-agent environments.

Acknowledgements

The work of X. Li was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0121500, and in part by the National Natural Science Foundation of China under Grant 61971126, Grant 61831013, and Grant 61571112.

References

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, O. P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 6379–6390.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2015. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Vezzani, G.; Gupta, A.; Natale, L.; and Abbeel, P. 2019. Learning latent state representation for speeding up exploration. *arXiv preprint arXiv:1905.12621*.