Deep Reinforcement Learning via Past-Success Directed Exploration

Xiaoming Liu, Zhixiong Xu,* Lei Cao, Xiliang Chen, Kai Kang

College of Command and Control Engineering, Army Engineering University

No. 1 Guang Hua Road, Nanjing, Jiangsu, China, 210007

lxm.nj@foxmail.com xu.nj@foxmail.com caolei.nj@foxmail.com 383618393@qq.com 36442181@qq.com

Abstract

The balance between exploration and exploitation has always been a core challenge in reinforcement learning. This paper proposes "past-success exploration strategy combined with Softmax action selection"(PSE-Softmax) as an adaptive control method for taking advantage of the characteristics of the online learning process of the agent to adapt exploration parameters dynamically. The proposed strategy is tested on OpenAI Gym with discrete and continuous control tasks, and the experimental results show that PSE-Softmax strategy delivers better performance than deep reinforcement learning algorithms with basic exploration strategies.

Introduction

The goal of reinforcement learning (RL) is to maximize the cumulative rewards during the ongoing interaction between the agent and the environment. The agent begins with no knowledge about the environment until it takes a sequence of actions to explore and learn from experience, which means that the agent can hardly maximize the cumulative reward unless it has sufficiently explored the environment. This leads to a fundamental trade-off of exploration versus exploitation (Tang et al. 2017).

In this paper, we proposes a novel and simple past-success directed exploration strategy called PSE-Softmax exploration strategy, which combines past-success directed exploration strategy and Softmax action selection method. Our proposed method not only utilizes the characteristics of the agent's online learning process to tune the exploration parameter dynamically, but also selects potential optimal action more accurately to guide the process of learning. Moreover, PSE-Softmax exploration strategy can also be applied to solve both discrete and continuous control tasks.

We evaluated the proposed method on OpenAI Gym with both discrete and continuous control tasks. Our proposed method was combined into two different types of representative reinforcement learning algorithms for testing, and the empirical evidence show that PSE-Softmax exploration strategy indeed improves the performance of reinforcement learning algorithms with basic exploration strategies.

Methodology

Previously Value-Difference Based Exploration combined with Softmax action selection (VDBE-Softmax) (Tokic and Palm. 2011) takes advantage of the value difference produced during learning process as a measure of the agent's uncertainty about the environment to adapt exploration parameters online, which has been proved to lead to statistically-significant improvement in bandit problems. However, one of drawbacks of those exploration strategies is that they have to record exploration parameters for each state, it's inefficient when encountered with large-scale continuous state or action space. We propose a novel and generic method called PSE-Softmax, which adapts the exploration parameter dynamically according to the immediate rewards and the average discounted rewards during the interaction between the agent and environment.

The core idea of PSE-Softmax method is to bias exploration by the amount and rate of success the agent has in reaching the goal state, which we called past-success directed exploration. On the one hand, when the agent receives reward at an increasing rate, it should exploits more. On the other hand, when the agent stops receiving reward because of a change in the environment, exploration should again increase.

The overall framework of past-success directed deep reinforcement learning is shown in Figure 1, which combines past-success directed exploration module with deep reinforcement learning, besides, PSE-Softmax exploration strategy can also be applied to solve both discrete and continuous control tasks.

* Corresponding author

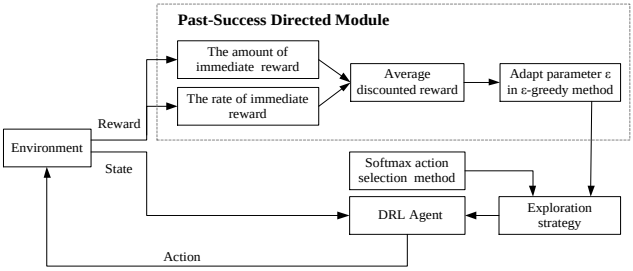

Figure 1: The framework of past-success directed deep reinforcement learning

We introduce the average discounted reward to help reflect both the amount and the frequency of received immediate rewards, which is defined as the following equation:

$$\mu_t^r = \frac{\sum_{k=1}^{t} v^{t-k+1} r_t}{\sum_{k=1}^{t} v^{t-k+1}} \tag{1}$$

Where $v \in (0,1]$ is the discount factor and r_t represents the reward received at time t. The average discounted reward is regarded as the hints given by the learning process itself to adapt the exploration parameter ε, which can be derived as follows:

$$\varepsilon_t = k \cdot e^{-\alpha \cdot \mu_t^r} + b \tag{2}$$

Where $k, b \in (0,1)$, α determines the slope of the sigmoid curve called inverse sensitivity.

In the end, we combine past-success directed exploration strategy with Softmax action selection method as PSE-Softmax method, which we redefine the PSE-Softmax exploration strategy as following equation:

$$\pi(s) = \begin{cases} \text{Softmax action } a & \text{if } \xi < \varepsilon_t \\ \arg\max_{a \in A(s)} Q(s,a) & \text{otherwise} \end{cases} \tag{3}$$

Where ε_t comes from equation(2).

Experiments

Environments

We choose Mountaincar-v0 and LunarLanderContinuous-v2 on OpenAI Gym as the test tasks. For Mountaincar-v0 task, the reward is given by $r(s,a) = -1 + h$, where h means the vertical position of car. For LunarLander-v2 task, See Box2D games in OpenAI Gym for details.

Baseline methods

We choose DQN (Mnih et al. 2015) and DDPG (Lillicrap et al. 2015) as the basic deep reinforcement learning algorithms with *ε-greedy* and Softmax method for comparison, and evaluate PSE-Softmax method on OpenAI Gym with high-dimensional state space, discrete and continuous action space.

Implementation Details

The network parameters of DQN and DDPG are same as the original paper. Moreover, for *ε-greedy* method, we set $\varepsilon = 0.05$, for the Softmax method, set $\tau = 3$, and for the PSE-Softmax method, set $k = 0.8$, $b = 0.1$, $\alpha = 0.9$, and $\tau = 3$. We call the improved DQN and DDPG algorithms as PSE-DQN, PSE-DDPG algorithms.

Experimental Results

We independently carried out each algorithm 10 times. The algorithms were tested 50 episodes per 100 training episodes to calculate the average scores of multiple runs.

Table 1 and 2 present the average scores and standard deviations. Compared to basic algorithms with *ε-greedy* and Softmax method, the PSE-DQN and PSE-DDPG exhibit superior performance and stability.

Table 1: Comparison of Mountaincar-v0

Task (AVG,STD)	Random	DQN with *ε-greedy*	DQN with Softmax	PSE-DQN
Mountaincar-v0	-178.2	(-87.8, 6.3)	(-78.2, 7.4)	(-70.3, 5.8)

Table 2: Comparison of LunarLanderContinuous-v2

Task (AVG,STD)	Random	DDPG with *ε-greedy*	DDPG with Softmax	PSE-DDPG
LunarLander-Continuous-v2	-246.2	(135.3, 43.5)	(149.1 34.1)	(210.3, 30.2)

Conclusions

This paper proposes a versatile exploration strategy to balance exploration and exploitation dynamically and avoid blind exploration. The experimental results show that PSE-Softmax method leads to improvements of performance while combining with basic reinforcement learning methods.

References

Lillicrap, T.; Hunt, J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; and Silver, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.; Alex, G.; Riedmiller, M.; Fidjeland, A.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature 7540:* 518-529.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Chen, O. X.; Duan, Y.; Schulman, J.; Turck, F. D.; and Abbeel, P. 2017. Exploration: A study of count-based exploration for deep reinforcement learning. In *NIPS*, 2753-2762.

Tokic, M., and Palm, G. 2011. Value-difference based exploration: adaptive control between epsilon-greedy and softmax. In *Annual Conference on Artificial Intelligence*, 335-346.