动态噪声的竞争深度Q网络 \*

张鑫，张席

(深圳大学 计算机与软件学院，深圳 518061)

摘　要：深度强化学习通常是通过在动作空间注入噪声来进行探索，但难以根据学习状况进行调整。另一种方法是在参数空间加入噪声，使得探索更加丰富，但会大幅减缓训练过程，使得收敛速度变慢。针对这个问题，提出了结合竞争网络结构和动态噪声的方法，即在动作空间与参数空间均加入噪声，动态地结合了各自优点。在训练前期以动作空间的噪声为主，减少神经网络的参数学习压力，从而更快的适应环境。利用测试平台OpenAI Gym进行对比实验，结果表明提出的方法取得了更好的学习性能，训练稳定性有明显提升。

关键词：深度Q网络；强化学习；探索与利用

中图分类号：TPxxx﹝查询请参考http://ztflh.xhma.com﹞

Dueling Deep Q Network with Dynamic Noise

Zhang Xin, Zhang Xi

(College of Computer Science and Software Engineering of Shenzhen University, Shenzhen 518061, China)

Abstract: Deep reinforcement learning generally engage in exploratory behavior through noise injection in the action space, but it is difficult to adjust according to the training situation. An alternative is to add noise directly to the parameter space, which can lead to a richer set of behaviors, but also slow down the training process and the convergence speed greatly. To address this problem, a method combining dueling network and dynamic noise is proposed, that is, adding noise both to the action space and parameter space. In the early stage of training, the noise in the action space is mainly used to reduce the parameter learning pressure of the neural network, so as to adapt to the environment more quickly. The OpenAI Gym was used for comparison experiment. The results show that the proposed method achieved better learning performance and significantly improved training stability.

Key words: DQN; Reinforcement Learning; exploration and exploitation

1. 引言

强化学习作为机器学习中的一种无监督式学习，基本学习方法是通过智能体在一定时间步中，不断的采取动作与所在环境进行交互，观测到环境对当前智能体状态、动作的反馈信号后调整下一步动作，提升自身的表现性能。在结合深度学习之前，传统的强化学习算法如Q-Learning[1]利用Q表来存储状态动作映射的Q值，由于计算复杂度的问题，不能处理高维度的输入，只能局限于低维问题。深度强化学习将深度学习的高维度感知处理能力与强化学习的决策能力相结合，取得了实质性的突破[2]。

Mnih V等人[3]首次提出将深度学习与强化学习算法Q-Learning结合的方法，即DQN (Deep Q-network) 深度强化学习，直接将预处理后的原始图片像素作为深度神经网络的输入，预测该输入对应的Q值，输出要选择的动作，在Atari 2600游戏中取得了超过人类玩家的分数[4]。双DQN[5] (Double DQN) 方法采用不同网络参数来对Q值进行估计，解决DQN的过估计问题。竞争网络[6] (Dueling Network)改进原本的网络结构，将原始DQN的输出解耦成两个分支，分别预测作为标量的状态值函数，与作为矢量的优势函数，后者中的每个值对应一个动作，两个函数相加输出每个动作的Q值，提升函数逼近效果。为了平衡强化学习中探索和利用之间的关系，通常是在动作空间加入噪声，如*ε*-greedy策略[1]，但参数*ε*是单调递减的线性函数，对于动态的学习过程，并不能完全适应。文献[7]和文献[8]分别提出两种在参数空间加入噪声的方法，引起网络输出变化，从而影响选择的动作。本文采用动作空间噪声与参数空间噪声动态结合的竞争深度Q网络针对控制问题进行研究，在OpenAI Gym提供的控制问题环境中取得了更高的分数。

1. 强化学习
   1. 强化学习算法

强化学习与传统的搜索策略不同，与环境进行交互的智能体能够通过学习环境中采集的样本，获得从状态到动作的映射，从而更快的得到最优的动作。强化学习的应用场景假设遵循马尔科夫性[9]，只需要当前步的状态信息，即可预测出一个合理的动作，执行动作到环境中得到反馈回报奖励，针对状态回报奖励为：



其中，智能体在在状态下执行的动作达到下一个状态，并获得奖励的概率为。状态动作值函数表示状态在时间步下执行动作获得的期望回报，折扣因子表示未来反馈回报对当前期望回报的影响程度，根据公式(1)递推，可得到递归公式[10]：



给定一个确定性策略，为获得更优策略，不断更新使其接近最优策略，记动作，则：



结合公式(2)(3)推导，可得*Q*-Learning算法[11]公式如下：



状态动作函数值通过映射表的方式进行存储，但在高维度问题下，映射表的大小占用大量的存储空间，并且有取值困难、效率低下等问题。

* 1. 深度Q网络
  2. 改进的竞争网络

1. 动态噪声模型
   1. *ε*-greedy策略
   2. 参数空间噪声
   3. 动态噪声
2. 实验结果与分析
   1. 实验设计
   2. 实验结果分析
3. 结束语

参考文献

1. Sutton R.S, Barto A.G. Reinforcement learning: An introduction [J]. IEEE Transactions on Neural Networks, 1998, 9(5):1054-1054.
2. 刘全,翟建伟,章宗长,等.深度强化学习综述 [J].计算机学报,2018, 41(1): 1-27.(Liu Quan,Zhai Jianwei,Zhang Zongchang, *et al*. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41 (1): 1-27. )
3. Mnih V, Kavukcuoglu K, Silver D, *et al*. Playing atari with deep reinforcement learning. arXiv preprint arXiv: 1312. 5602, 2013
4. Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518：529-533.
5. Van H, Guez A, Silver D. Deep reinforcement learning with double q-learning [C]// Thirtieth AAAI conference on artificial intelligence, 2016
6. Wang Z , Schaul T , Hessel M , *et al*. Dueling Network Architectures for Deep Reinforcement Learning [C]// Proc of the 33rd International Conference on Machine Learning, 2016.
7. Plappert M, Houthooft R, Dhariwal P, *et al*. Parameter space noise for exploration. arXiv preprint arXiv:1706.01905, 2017
8. Fortunato M, Azar MG, Piot B, *et al.* Noisy networks for exploration. arXiv preprint arXiv:1706.10295, 2017
9. 邱祎, 董彦彦. 基于马尔可夫过程的线性规划方法探讨 [J]. 统计与决 策, 2017, 10: 88-90. (Qiu Yi, Dong Yanyan. Linear programming method based on Markov process [J]. Statistics & Decision, 2017, 10: 88-90. )
10. Dolcetta IC, Ishii H. Approximate solutions of the Bellman equation of deterministic control theory[J]. Applied Mathematics and Optimization. 1984, 11(1):161-81.
11. Watkins C J C H，Dayan P. Q - learning[J]. Machine Learning，1992，8：279-292.