论文解读之 A novel DDPG method with prioritized experience replay

会议：IEEE SMC 2017
原文地址：https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8122622
代码：https://github.com/cardwing/Codes-for-RL-PER

本文[1]主要是将强化学习中基于优先级的经验回放机制(prioritized experience replay)从离散控制领域拓展到了连续控制领域，并指出该做法可以显著减少网络的训练时间，提高训练过程的稳定性及模型的鲁棒性。

论文内容简介：

在传统的连续控制（动作是实数值）领域里，让 agent 直接通过像素级的视觉输入来完成复杂的操纵任务是十分困难的[2, 3]。最近，一个新颖的深度强化学习(deep reinforcement learning)算法，DDPG (deep deterministic policy gradient)[2]，在很多仿真的连续控制任务里取得了很好的效果。DDPG 使用一个经验回放池(replay buffer)来消除输入经验(experience)间存在的很强的相关性。这里，经验指一个四元组(s_t, a_t, r_t, s_{t+1})[4, 5]。同时，DDPG 使用目标网络法来稳定训练过程。作为 DDPG 算法里的一个基本组成部分，经验回放极大地影响了网络的训练速度和最终效果。

经验回放机制具体如下：它用一个固定大小的内存空间（又叫 replay buffer）来存储之前的经验并每次从中随机选取一个固定数目的经验来更新网络。显然，存在于回放的经验间的时序相关性被极大减弱因为经验回放机制把新旧经验混合在了一起来更新网络。不过，经验回放机制是基于这样一个 intuition，即经验池中的所有经验都一样重要，因此它均匀地从经验池中选取一定数目的样本来训练网络。但是很显然，这种 intuition 有违常理。在人们学习做某件事情的时候，那些高回报，十分成功，或者痛苦的经历会在学习过程中在人们的大脑里不断回顾。因此，那些不断被回顾的经验和一般经验比更有学习价值。因此，基于这样的想法，我们提出区分经验池中不同经验的价值，即将 DDPG 中原本的随机经验回放替换为基于优先级的经验回放[6]。

基于优先级的经验回放具体如下：首先，我们使用时间差分偏差(temporal difference error)来衡量每个经验的学习价值；其次，通过时间差分偏差的绝对值来对经验池里的经验进行排序，我们更加频繁地回放那些高偏差的经验。这种做法会不可避免地改变状态访问频率从而引入 bias。为了修正 bias，我们使用了重要性采样权重(importance-sampling weight)[7]。我们在经典的倒立摆任务[8]中测试了我们提出的算法。实验表明，基于优先级的经验回放可以显著减少 DDPG 算法的训练时间，提高训练过程的稳定性，并提升模型的鲁棒性。（详细内容见原论文）

References:

[1] Hou, Y., Liu, L., Wei, Q., Xu, X. and Chen, C., 2017, October. A novel DDPG method with prioritized experience replay. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 316-321). IEEE.
[2] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D., 2015.

Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.
[3] Heess, N., Wayne, G., Silver, D., Lillicrap, T., Erez, T. and Tassa, Y., 2015. Learning continuous control policies by stochastic value gradients. In Advances in Neural Information Processing Systems (pp. 2944-2952).
[4] Lin, L.J., 1993. Reinforcement learning for robots using neural networks (No. CMU-CS-93-103). CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.
[5] Adam, S., Busoniu, L. and Babuska, R., 2012. Experience replay for real-time reinforcement learning control. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2), pp.201-212.
[6] Moore, A.W. and Atkeson, C.G., 1993. Prioritized sweeping: Reinforcement learning with less data and less time. Machine learning, 13(1), pp.103-130.
[7] Mahmood, A.R., van Hasselt, H.P. and Sutton, R.S., 2014. Weighted importance sampling for off-policy learning with linear function approximation. In Advances in Neural Information Processing Systems (pp. 3014-3022).
[8] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. and Zaremba, W., 2016. Openai gym. arXiv preprint arXiv:1606.01540.