

深度强化学习中的探索综述

田鸿龙

Software Institute, Nanjing University

November 25, 2020

Table of Contents

概述

传统强化学习中的探索问题
深度强化学习

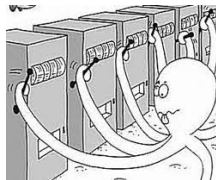
Table of Contents

概述

传统强化学习中的探索问题

深度强化学习

Bandit



MAB 问题：你进了一家赌场，假设面前有 K 台老虎机 (arms)。我们知道，老虎机本质上就是个运气游戏，我们假设每台老虎机 i 都有一定概率 p_i 吐出一块钱，或者不吐钱（概率 $1 - p_i$ ）。假设你手上只有 T 枚代币 (tokens)，而每摇一次老虎机都需要花费一枚代币，也就是说你一共只能摇 T 次，那么如何做才能使得期望回报 (expected reward) 最大呢？

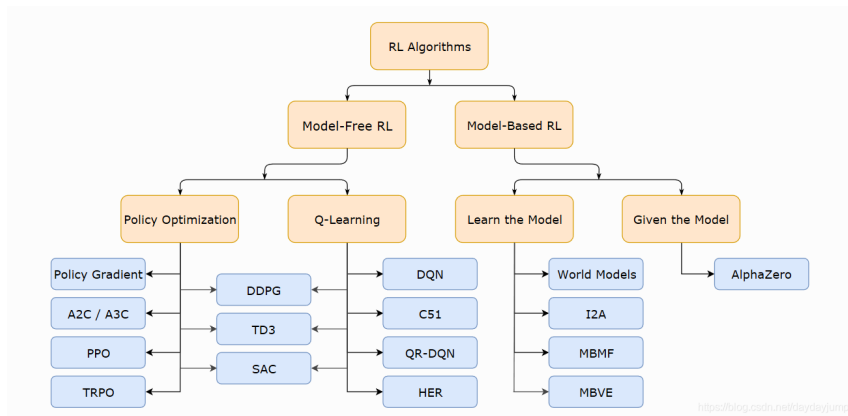
Expore or Exploit?

Table of Contents

概述

传统强化学习中的探索问题
深度强化学习

深度强化学习主流算法



两种思想

- Value-Based: 和传统强化学习一样，试图学到一个值函数（Q Function 或者 V Function），通过这个值函数贪心（或在贪心的基础之上探索）形成策略，理论基础是广义策略迭代。
- Policy-Based: 基于函数逼近的方法，因为深度学习强大的拟合能力而成为深度强化学习的主流方法，直接学习 $\pi: S \rightarrow A$ ，理论基础是策略梯度定理。
- 两种思想结合形成一大类 Actor-Critic 方法。其中 A2C, A3C, PPO, TRPO 等方法是 On-Policy 的，也被视为 Policy-Based，而 DDPG, TD3, D4PG, SAC 等算法是 Off-Policy 的。

Value-Based







- DQN[1] 是深度学习和强化学习结合的第一步，将 Q-learning[2] 和神经网络结合起来。
- DRQN[3] 将 DQN 和 RNN 结合起来，用来解决 POMDP 问题（但是因为训练成本和采样效率的问题，时序模型在强化学习中并不常见）。
- Dueling DQN[4] 将 Q Function 拆开分成 Advantage Function 和 Value Function 的和，提升了 DQN 的泛化性能。
- Double DQN[5] 结合 Double Q-learning[6] 的思想，有效的缓解了值函数低估的问题。
- PER[schaul2015prioritized] 在 DQN 的基础上加入优先经验回放，进一步提升了采样效率。
- Distributional DQN[7][8][9] 将值函数扩展到值概率分布，进一步提升函数估计的准确性。

Rainbow[hessel2017rainbow] 将上述方法结合，形成了有效的算法。




Policy-Based

- 策略梯度定理 [sutton1999policy] 给出了 Policy Gradient 算法的可行性和性能分析。
- [kakade2001natural] 说明了梯度下降的局限性，将自然梯度下降用于 Policy 的优化。
- A3C[mnih2016asynchronous] 使用异步并行缓解采样效率低下的问题。
- TRPO[schulman2015trust] 结合 [kakade2001natural] 给 Policy Gradient 的局部 Off-Policy 做出了理论分析。
- PPO[schulman2017proximal] 在 [schulman2015trust] 的基础之上做了近似估计。

References I

-  Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *nature* 518.7540 (2015), pp. 529–533.
-  Christopher JCH Watkins and Peter Dayan. “Q-learning”. In: *Machine learning* 8.3-4 (1992), pp. 279–292.
-  Matthew Hausknecht and Peter Stone. “Deep recurrent q-learning for partially observable mdps”. In: *arXiv preprint arXiv:1507.06527* (2015).
-  Ziyu Wang et al. “Dueling network architectures for deep reinforcement learning”. In: *International conference on machine learning*. PMLR. 2016, pp. 1995–2003.
-  Hado Van Hasselt, Arthur Guez, and David Silver. “Deep reinforcement learning with double q-learning”. In: *arXiv preprint arXiv:1509.06461* (2015).
-  Hado Hasselt. “Double Q-learning”. In: *Advances in neural information processing systems* 23 (2010), pp. 2613–2621.

References II

-  Will Dabney et al. “Distributional reinforcement learning with quantile regression”. In: *arXiv preprint arXiv:1710.10044* (2017).
-  Will Dabney et al. “Implicit quantile networks for distributional reinforcement learning”. In: *arXiv preprint arXiv:1806.06923* (2018).
-  Marc G Bellemare, Will Dabney, and Rémi Munos. “A distributional perspective on reinforcement learning”. In: *arXiv preprint arXiv:1707.06887* (2017).