doi:10.3969/j.issn.1672-1292.2022.01.004

基于 UCB 算法的交替深度 Q 网络

吴卿源[1],谭晓阳[1,2]

(1.南京航空航天大学计算机科学与技术学院,江苏 南京 211106)
(2.南京航空航天大学模式分析与机器智能工业和信息化部重点实验室,江苏 南京 211106)

[摘要] 在深度强化学习中,智能体需要与环境进行交互学习,这就需要智能体能够很好地去平衡利用与探索.因此如何提升算法的样本有效性,增加算法的探索能力,一直是深度强化学习领域中非常重要的研究方向. 结合已有研究成果,提出了一种交替使用多个不同初始化深度 Q 网络方法,使用网络随机初始化带来的探索性能.基于最大置信度上界算法先构造一种交替选择深度 Q 网络策略. 并将该调度网络策略与多个随机初始化的深度 Q 网络结合,得到基于最大置信度上界的交替深度 Q 网络算法. 在多个不同的标准强化学习实验环境上的实验结果表明,该算法比其他基准算法有更高的样本效率和算法学习效率.

[关键词] 强化学习,深度强化学习,深度 Q 网络,最大置信度上界
[中图分类号]TP18 [文献标志码]A [文章编号]1672-1292(2022)01-0024-06

Alternated Deep Q Network Based on Upper Confidence Bound

Wu Qingyuan[1],Tan Xiaoyang[1,2]

(1.College of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China)
(2.MIIT Key Laboratory of Pattern Analysis and Machine Intelligence,Nanjing University of Aeronautics and Astronautics,Nanjing 211006,China)

Abstract:The agent needs to learn interactively with the environment in the paradigm of deep reinforcement learning (DRL). The important dilemma of DRL is that the agent needs to balance exploitation and exploration. Therefore,how to improve the sample efficiency of algorithms and increase the exploration ability of the algorithm is a very popular research direction in the field of DRL. Different from existing works,we apply multiple DQNs with independent random initialization and use them to interact with the environment alternately. Using the generalized exploration abilities brought by random initialization of the networks,this paper proposes a method of alternately selecting DQN based on the maximum confidence upper bound(UCB) method,which is called Alternated DQN(ADQN). Experimental results on different standard reinforcement learning experimental environments show that ADQN has higher sample efficiency and algorithm learning efficiency than other benchmark algorithms.

Key words:reinforcement learning,deep reinforcement learning,deep Q-network,upper confidence bound

在强化学习(reinforcement learning,RL)问题中,智能体与环境进行交互,并学习如何在当前状态下采取最好的动作以最大化未来累计奖励[1]. 传统的强化学习算法,如 Q-learning、SARSA[1] 等都已经在简单强化学习问题上取得了不错的效果. DeepMind 公司将强化学习与深度学习相结合提出了深度 Q 网络[2] 在 Atari 游戏取得了不错的效果[3]. 其将复杂的高维图像输入神经网络提取特征,并与传统的 Q-learning 进行结合. 结合了深度 Q 网络和蒙特卡洛树搜索等著名 AlphaGo 的强化学习算法[4],其打败了人类围棋世界冠军. 且最新的 MuZero[5] 几乎能够学习所有的棋类游戏以及 Atari 游戏,并都能达到超越人类的水平.

在实际环境中,智能体常常面对的是未知的环境,所能获取的关于环境的信息很少,只能通过之前的交互经验来学习. 并且还常常面对着探索与利用的难题,智能体需要去探索不同的环境以提升未来的奖励,会牺牲一些眼前的奖励,以最终获得的奖励最大化. 并且智能体需要进行有效的探索,以使得样本采

收稿日期:2021-08-31.
基金项目:科技创新 2030 重大项目(2021ZD0113203)、国家自然科学基金项目(61976115).
通讯作者:谭晓阳,博士,教授,研究方向:强化学习. E-mail:x.tan@ nuaa.edu.cn

样有效性最大化,从而使得智能体训练效率提升. 而如何进行有效的探索一直是一个尚未解决的难题. 本文通过运用多个不同初始化的深度 Q 网络进行加速探索. 基于最大置信度上界算法提出了一种交替深度 Q 网络的策略算法,即在当前状态下根据 UCB 值来确定该使用哪个深度 Q 网络进行交互. 在 3 个不同的标准实验环境下,本文提出的方法比其他的基准算法效果更好.

本文的主要贡献包括以下 3 个方面:

(1)本文提出了一种基于最大置信度上界算法,根据环境来选择深度 Q 网络的策略算法 UCB-Q.

(2)本文将 UCB-Q 与多个不同初始化的深度 Q 网络结合,提出了基于 UCB 的交替深度 Q 网络.

(3)在不同的实验环境上的实验效果表明,本文提出的方法能够有效地提高样本效率,相比其他的基准算法更有优势.

1 相关工作

深度 Q 网络是一个非常经典的基于价值函数的深度强化学习算法,现有的许多工作都以其为研究背景. 如文献[6]为了解决深度 Q 网络中的过估计问题,将 Double Q-learning 与神经网络结合提出 DDQN. 文献[7]认为在历史经验回放池中,需要根据时序差分误差给每个样本进行概率加权,使得学习效率更高. 文献[8]提出了一种新的结构,其网络分别通过学习状态价值函数和状态动作优势函数,以提升算法训练的稳定性. 文献[9]结合了当前深度 Q 网络的各种技巧结合得到了 Rainbow,其在各个实验环境上得到最优的水平.

在当今的强化学习问题中,如何平衡利用与探索一直是一个尚未解决的难题. 并且由于强化学习需要根据在经验样本回放池中存放大量的历史交互经验来训练神经网络. 而传统的深度 Q 网络常常使用 ε-贪心算法来进行探索,但在面对复杂环境下,这样的探索有效性是非常低的. 因此如何进行有效的探索,提高样本有效性是当前一个非常重要的研究方向. 现有的很多研究都致力于去解决该问题,如文献[10]提出了使用噪声网络去增强网络的探索能力.

与本文使用多个网络的思想类似,文献[11]提出了使用随机初始化多头网络的 Bootstrapped DQN 去模拟价值函数分布. 文献[12]提出了基于最大置信度上界的集成深度 Q 网络,使用了随机初始化的多头网络,更关注于动作维度上的利用与探索. 文献[13]则是将对于采样的样本进行最大置信度上界策略的采样.

本文提出的基于最大置信度上界的交替深度 Q 网络算法. 根据当前状态下计算多个不同初始化深度 Q 网络的 UCB 值,并根据该值去选择深度 Q 网络与环境进行交互,关注于多个网络层面的利用与探索.

2 背景知识

2.1 马尔可夫决策过程

强化学习问题通常被建模为马尔可夫决策过程模型. 一个马尔可夫决策过程通常定义为一个五元组 $M=[S,A,P,R,\gamma]$,其中 $s\in S$ 为状态空间中的一个状态,$a\in A$ 为动作空间中的一个动作,$P(s'|s,a)$ 为智能体在状态 s 下采取动作 a 之后转移到状态 s' 的概率,$R:S\times A\to\mathbf{R}$ 是奖励函数且 $\gamma\in[0,1)$ 是折扣系数. 智能体希望学习一个策略 $\pi:S\to A$,即一个从状态空间到动作空间的映射函数,以最大化在给定状态 s 下的未来期望回报,即状态价值函数 $V^{\pi}(S)=E_{a\sim\pi}Q^{\pi}(s,a)$. 其中 $Q^{\pi}(s,a)$ 为状态-动作价值函数,其为根据策略 π 在状态 s 下采取 a 之后,直到交互结束所获得的累积奖励. 本文中采取算法是基于 Q 学习算法,其希望通过贝尔曼最优算子来学习最优策略,可表示为

$$Q(s,a)=R(s,a)+\gamma\underset{s'\sim P(\cdot|s,a)}{E}\left[\max_{a'\in A}Q(s',a')\right]. \tag{1}$$

2.2 深度 Q 网络

通过将深度学习与 Q 学习算法相结合得到了深度 Q 网络算法. 即通过参数为 θ 的神经网络函数来近似状态动作价值函数 $Q(s,a|\theta)\approx Q(s,a)$. 在与环境进行交互时,深度 Q 网络采取的是离策略,即在与环境交互时采取 ε-贪心策略进行交互. 在估计最优动作状态价值函数时采取的是贪心策略.

深度 Q 网络将所有的历史交互经验保存在一个具有固定长度的历史经验回放池中. 在更新 Q 网络时,从历史经验回访池中采样得到$<s_t,a_t,r_t,s_{t+1}>$,并根据式(2)对网络参数进行更新.

$$\theta_{t+1}\leftarrow\theta_t+\alpha(y-Q(s_t,a_t|\theta_t))\nabla_\theta Q(s_t,a_t|\theta_t). \tag{2}$$

式中,$y=r_t+\gamma\max\limits_a Q(s_{t+1},a_t|\theta^-)$为目标值,$Q(s,a|\theta^-)$为参数固定为 θ^-的目标 Q 网络. 其初始化参数与策略 Q 网络参数一致,每过一段时间就会将策略 Q 网络的参数赋予目标 Q 网络,$\theta^-\leftarrow\theta$. 由于在深度 Q 网络算法中存在最大化操作算子,其会导致智能体对与价值函数的估计偏大,即会导致过估计问题. 因此双重深度 Q 网络就采取式(3)对目标值进行估计,以减小过估计的负面影响.

$$y=r_t+\gamma\max_a Q(s_{t+1},\arg\max_a Q(s_{t+1},a|\theta_t)|\theta^-). \tag{3}$$

2.3　UCB 算法

在解决探索与利用难题上,现有的深度 Q 网络常用 ε-贪心来进行平衡. 但该方法在进行探索时是进行均匀探索,而在历史经验回放池中大量的历史信息都没有被使用. 而在传统的规划问题中,UCB 算法是一个有效利用了历史信息的探索利用算法. 其主要思想是希望通过构造一个 UCB 值,该值包括了利用与探索两项. 即其对于每一个动作都有一个置信度评价,反映了其期望平均收益和不确定性,在每次决策时都会选择 UCB 值最大的那一个动作. 经典的 UCB 策略为

$$y_i=V_i+\sqrt{\frac{2\ln N}{1+N_i}}. \tag{4}$$

式中,y_i 表示了第 i 个动作的 UCB 值,V_i 表示了选择了第 i 个动作的平均收益,N_i 为选择第 i 个动作的次数,N 为选择所有动作的次数之和. 其第一项为利用项表示了对于动作的性能评价,而第二项为历史动作选择比例,其表示了对于该动作的一个不确定度.

3　基于 UCB 的交替深度 Q 网络算法

不同与以往的深度 Q 网络算法,本文希望解决的是强化学习任务中样本效率与算法效率低下的问题. 而传统的深度 Q 网络只使用单个智能体与环境进行交互以获得经验,且其探索能力只能依靠 ε-贪心策略,所获取到的经验样本单一,从而使得网络样本效率低下. 本文希望将一个强化学习问题,即马尔克夫决策过程,转换为让多个智能体轮流进行单独决策,共同完成目标的决策问题. 因此本文采取($k\geqslant 2$)多个不同初始化的深度 Q 网络 $Q(s,a|\theta^i)(i=1,\cdots,k)$,简记为 Q_i. 并使用一个公共的历史经验回放池 B,所有的历史交互经验样本全都存于当中. 本文所研究的环境为单智能体的环境,因此每个时刻只会选择一个深度 Q 网络与环境进行交互. 而本文采用基于 UCB 的策略来选择交互的深度 Q 网络,并且每个网络在交互时只使用贪心策略进行探索. 因此在状态为 s 时,会选择一个最佳的 Q 网络. 每个 episode 会交替着选择多个不同的 Q 网络. 所有的交互经验信息都会存放到 B 中. 在每次更新网络时,每个网络都会从中各自采样一批样本来训练. 本文提出了一个基于 UCB 的交替选择深度 Q 网络的策略算法 UCB-Q,

$$Q_i=\arg\max_i\left(\max_a Q(s,a|\theta_i)+\sqrt{\frac{2\ln N}{1+N_i}}\right). \tag{5}$$

式中,$\max\limits_a Q(s,a|\theta_i)$为网络的最优状态动作价值,即在使用贪心策略时,智能体会选择的最优动作的状态动作价值. 而后一项为 Q_i 的历史选择次数比,代表了该网络的不确定性. 结合该选择深度 Q 网络的策略,本文提出了基于最大置信度上界的交替深度 Q 网络算法(alternated DQN,ADQN). ADQN 的算法流程图如图 1 所示.

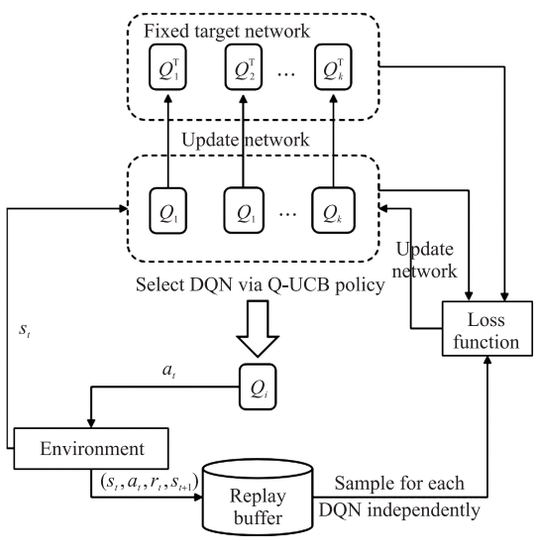

图 1　基于 UCB 的 ADQN 算法流程图

Fig. 1　The flow chart of ADQN algorithm based on UCB

本文提出的基于 UCB 的交替深度 Q 网络算法总结为.

算法 1　基于 UCB 的交替深度 Q 网络算法

输入:k 个各自独立初始化的深度 Q 网络 $\{Q_i\}_{i=1}^k$,公共历史经验回放池 B.

(1)将均匀随机选择 Q_i 与环境进行交互的经验样本

储存到 B 中

(2) For each episode do

(3) 获取环境的初始状态 s_0

(4) For each step t until end of episode do

(5) 根据式(5)选择网络 Q_i

(6) 贪心选择动作 $a_t=\arg\max_a Q_i(s_t,a)$

(7) 执行动作 a_t,得到状态 s_{t+1}和奖励 r_t

(8) 将(s_t,a_t,r_t,s_{t+1})储存到 B 中

(9) 更新网络选择次数 N_i 和 N

(10) 从 B 中独立随机采样更新网络$\{Q_i\}_{i=1}^{k}$

(11) End for

(12) End for

4 实验与结果

在本节中,将在不同的实验环境下对比现有算法,以证明本文提出算法的有效性. 为了公平对比,所有算法使用相同的深度学习框架,且参数等设置都一致,如历史经验回放池的大小、神经网络的结构、学习率、优化器等[14].

4.1 实验环境

本文选择在 3 个不同的标准强化学习实验环境 gym 验证算法的有效性. 选择的实环境有经典的强化学习问题环境 CartPole-v0 和 CartPole-v1,以及 MinAtar[15] 中的 Breakout 环境. 其游戏环境界面如图 2 所示.

(a) CartPole-v0/v1　　(b) MinAtar-Breakout

图 2　本文所使用的强化学习实验环境

Fig. 2　The RL environment used in this article

4.2 基准算法及参数设置

CartPole-v0 和 CartPole-v1 是经典的强化学习实验环境,因此该环境较为简单,因此本文在此环境上选择的对比基准算法为 DQN 与 DDQN. 在算法参数设置方面,使用激活函数为 ReLU 的 3 层全连接神经网络. 其训练长度分别为 50 和 125 个 episode. 具体的参数设置如表 1 所示.

在 MinAtar 的 Breakout 环境上选择不同的基准算法分别 Average DQN 与 Bootstrapped DQN[16]. 在算法参数设置方面,使用激活函数为 ReLU 的 5 层卷积神经网络. 其训练长度为 5×10^6 个 steps. 所有使用 k 个网络的算法都设置为一致.

本文提出的算法在较为简单的经典强化学习环境上使用 $k=2$ 个独立随机初始化的深度 Q 网络,而在 MinAtar 强化学习环境上使用 $k=5$ 个独立随机初始化的深度 Q 网络.

表 1　经典强化学习环境的参数设置

Table 1　The parameter setting in classical RL environment

参数名	数值
历史经验回放池长度	10 000
Batch size	32
目标网络更新次数	100
γ	0.99
学习率	0.001

表 2　MinAtar 强化学习环境的参数设置

Table 2　The parameter setting in MinAtar RL environment

参数名	数值
历史经验回放池长度	100 000
Batch size	32
目标网络更新次数	1 000
γ	0.99
学习率	0.002 5

4.3 实验结果及分析

在本文实验中,对每种算法在每个环境上进行了 5 次独立的实验,因此得到性能曲线如图 3、图 4 所示.

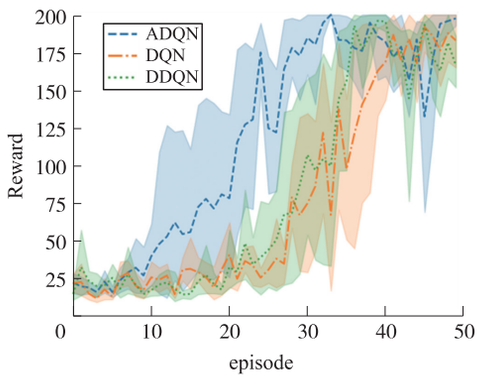

图 3　在 CartPole-v0 上算法的性能比较

Fig. 3　Performance comparison on CartPole-v0

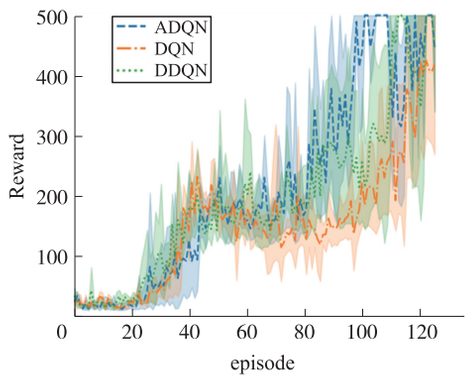

图 4　在 CartPole-v1 上算法的性能比较

Fig. 4　Performance comparison on CartPole-v1

根据分析在经典强化学习问题上的实验结果可知. 由图 2 可知,相比 DQN 和 DDQN,ADQN 在初始阶段的探索能力更强,其样本效率更高,并在 30 个 episode 时就趋近于收敛. 而 DQN 和 DDQN 其样本效率较低,分别在大约 35 和 40 个 episode 时趋近于收敛. 分析图 3 的结果可知,ADQN 在接近 100 个 episode 时就接近收敛到最高分. 相比之下,DQN 和 DDQN 样本效率低,需要在 120 个 episode 才能收敛. 由这两个经典的强化学习环境上的实验结果可知,ADQN 相比基准算法探索能力更强,且样本效率更高,性能更好.

分析图 5 中展示在 MinAtar-Breakout 环境上的实验结果可知,Alter DQN 和 Bootstrapped DQN 效果都明显优于其他基准算法. 并且在大约 3×10^6 个 steps 时 Alter DQN 开始优于 Bootstrapped DQN.

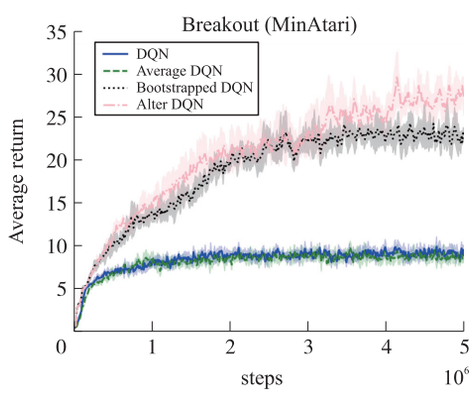

图 5　在 MinAtar 游戏 Breakout 上算法的性能比较

Fig. 5　Performance comparison on Breakout of MinAtar

5　结论

本文提出了一种基于 UCB 的交替深度 Q 网络算法. 不同于已有的深度 Q 网络算法常使用的 ε-贪心策略. 通过多个独立初始化的 DQN,并通过新构造的最大 UCB 策略算法来控制选择单个智能体与环境进行交互. 本文构造的 UCB 项希望能够平衡对当前网络的利用和探索,通过该策略能够提升算法的样本效率以提高性能. 通过在多个不同的标准强化学习环境上的实验结果证明了本文算法的有效性.

[参考文献](References)

[1] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054-1054.

[2] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv Preprint arXiv: 1312.5602, 2013.

[3] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[4] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[5] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering atari, go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588(7839): 604-609.

[6] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. arXiv Preprint arXiv: 1509.06461v3, 2016.

[7] SCHAUL T,QUAN J,ANTONOGLOU I,et al. Prioritized experience replay[J]. arXiv Preprint arXiv:1511.05952,2015.
[8] WANG Z,SCHAUL T,HESSEL M,et al. Dueling network architectures for deep reinforcement learning[C]//International Conference on Machine Learning. Lodon,UK,2016:1995-2003.
[9] HESSEL M,MODAYIL J,VAN HASSELT H,et al. Rainbow:Combining improvements in deep reinforcement learning[C]//Thirty-second AAAI Conference on Artificial Intelligence. Lousiana,USA,2018.
[10] FORTUNATO M,AZAR M G,PIOT B,et al. Noisy networks for exploration[J]. arXiv Preprint arXiv:1706.10295,2017.
[11] OSBAND I,BLUNDELL C,PRITZEL A,et al. Deep exploration via bootstrapped DQN[J]. Advances in Neural Information Processing Systems,2016,29.
[12] CHEN R Y,SIDOR S,ABBEEL P,et al. UCB exploration via Q-ensembles[J]. arXiv Preprint arXiv:1706.01502,2017.
[13] 朱斐,吴文,刘全,等. 一种最大置信上界经验采样的深度 Q 网络方法[J]. 计算机研究与发展,2018,55(8):1694-1705.
[14] WATKINS C,DAYAN P. Q-learning[J]. Machine Learning,1992,8(3/4):279-292.
[15] ANSCHEL O,BARAM N,SHIMKIN N. Averaged-DQN:variance reduction and stabilization for deep reinforcement learning[C]//International Conference on Machine Learning. Sydney,Australia,2017:176-185.

[责任编辑:陈　庆]