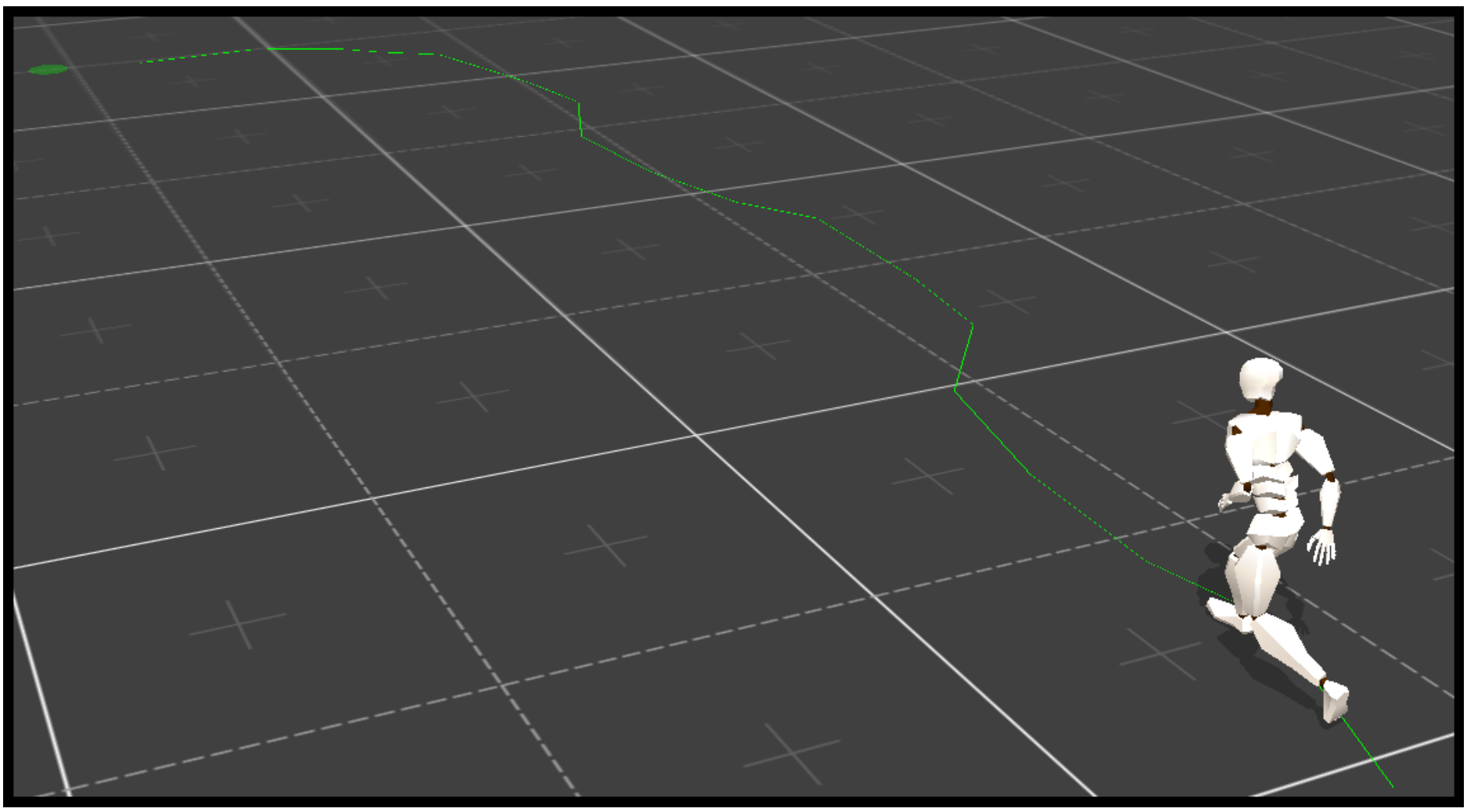
AlgorithmDog
freedom

强化学习系列之九:Deep Q Network (DQN)

发表于2016年9月12日由lili

文章目录 [隐藏]

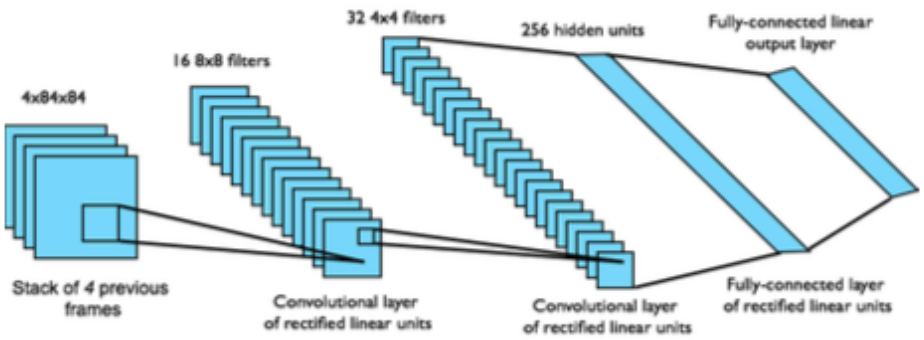
我们终于来到了深度强化学习。



1. 强化学习和深度学习结合

机器学习=目标+表示+优化。目标层面的工作关心应该学习到什么样的模型，强化学习应该学习到使得激励函数最大的模型。表示方面的工作关心数据表示成什么样有利于学习，深度学习是最近几年兴起的表示方法，在图像和语音的表示方面有很好的效果。深度强化学习则是两者结合在一起，深度学习负责表示马尔科夫决策过程的状态，强化学习负责把控学习方向。

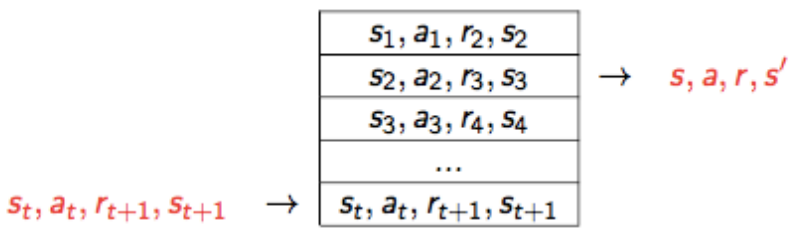
深度强化学习有三条线：分别是基于价值的深度强化学习，基于策略的深度强化学习和基于模型的深度强化学习。这三种不同类型的深度强化学习用深度神经网络替代了强化学习的不同部件。基于价值的深度强化学习本质上是一个 Q Learning 算法，目标是估计最优策略的 Q 值。 不同的地方在于 Q Learning 中价值函数近似用了深度神经网络。比如 DQN 在 Atari 游戏任务中，输入是 Atari 的游戏画面，因此使用适合图像处理的卷积神经网络（Convolutional Neural Network，CNN）。下图就是 DQN 的框架图。



2. Deep Q Network (DQN) 算法

当然了基于价值的深度强化学习不仅仅是把 Q Learning 中的价值函数用深度神经网络近似，还做了其他改进。

这个算法就是著名的 DQN 算法，由 DeepMind 在 2013 年在 NIPS 提出。DQN 算法的主要做法是 Experience Replay，其将系统探索环境得到的数据储存起来，然后随机采样样本更新深度神经网络的参数。



Experience Replay 的动机是：1）深度神经网络作为有监督学习模型，要求数据满足独立同分布，2）但 Q Learning 算法得到的样本前后是有关系的。为了打破数据之间的关联性，Experience Replay 方法通过存储-采样的方法将这个关联性打破了。

DeepMind 在 2015 年初在 Nature 上发布了文章，引入了 Target Q 的概念，进一步打破数据关联性。Target Q 的概念是用旧的深度神经网络 w^- 去得到目标值，下面是带有 Target Q 的 Q Learning 的优化目标。

$$J = min(r + \gamma max_{a'} Q(s', a', w^-)) - Q(s, a, w))^2$$

下图是 Nature 论文上的结果。可以看到，打破数据关联性确实很大程度地提高了效果。

Game	With replay, with target Q	With replay, without target Q	Without replay, with target Q	Without replay, without target Q
Breakout	316.8	240.7	10.2	3.2
Enduro	1006.3	831.4	141.9	29.1
River Raid	7446.6	4102.8	2867.7	1453.0
Seaquest	2894.4	822.6	1003.0	275.8
Space Invaders	1088.9	826.3	373.2	302.0

3. 后续发展

DQN 是第一个成功地将深度学习和强化学习结合起来的模型，启发了后续一系列的工作。这些后续工作中比较有名的有 Double DQN, Prioritized Replay 和 Dueling Network。

3.1 Double DQN

Thrun 和 Schwartz 在古老的 1993 年观察到 Q-Learning 的过优化 (overoptimism) 现象 [1]，并且指出过优化现象是由于 Q-Learning 算法中的 max 操作造成的。令 $Q^{target}(s,a)$ 是目标 Q 值；我们用了价值函数近似，Q^{approx} 是近似 Q 值；令 Y 为近似值和目标之间的误差，即

$$Q^{approx}(s,a) = Q^{target}(s,a) + Y_{s,a}$$

Q-learning 算法更新步骤将所有的 Q 值更新一遍，这个时候近似值和目标值之间的差值

$$
\begin{aligned}
Z &= r_{s,a} + \gamma max_{a1} Q^{approx}(s', a1) - r_{s,a} + \gamma max_{a2} Q^{target}(s', a2) \\
&= \gamma max_{a1} Q^{approx}(s', a1) - \gamma max_{a2} Q^{target}(s', a2) \\
&\geq \gamma Q^{approx}(s', a') - Q^{target}(s', a') = \gamma Y_{s',a'}
\end{aligned} \tag{1}
$$

其中 $a' = argmax_a Q^{target}(s', a)$。这时候我们发现，即使 $E[Y] = 0$ 也就是一开始是无偏的近似， Q Learning 中的 max 操作也会导致 E[Z] > 0。这就是过优化现象。为了解决这个问题，Thrun 和 Schwartz 提出了 Double Q 的想法。

Hasselt 等进一步分析了过优化的现象，并将 Double Q 的想法应用在 DQN 上，从而提出了 Double DQN。Double DQN 训练两个 Q 网络，一个负责选择动作，另一个负责计算。两个 Q 网络交替进行更新，具体算法如下所示。

Algorithm 1 Double Q-learning

1: Initialize Q^A,Q^B,s
2: **repeat**
3: Choose a, based on $Q^A(s,\cdot)$ and $Q^B(s,\cdot)$, observe r, s'
4: Choose (e.g. random) either UPDATE(A) or UPDATE(B)
5: **if** UPDATE(A) **then**
6: Define $a^* = \arg\max_a Q^A(s',a)$
7: $Q^A(s,a) \leftarrow Q^A(s,a) + \alpha(s,a)\left(r + \gamma Q^B(s',a^*) - Q^A(s,a)\right)$
8: **else if** UPDATE(B) **then**
9: Define $b^* = \arg\max_a Q^B(s',a)$
10: $Q^B(s,a) \leftarrow Q^B(s,a) + \alpha(s,a)(r + \gamma Q^A(s',b^*) - Q^B(s,a))$
11: **end if**
12: $s \leftarrow s'$
13: **until** end

下图是 Hasselt 在论文中报告的实验结果。从实验结果来看，Double DQN 拥有比 DQN 好的效果。

	DQN	Double DQN	Double DQN (tuned)
Median	47.5%	88.4%	116.7%
Mean	122.0%	273.1%	475.2%

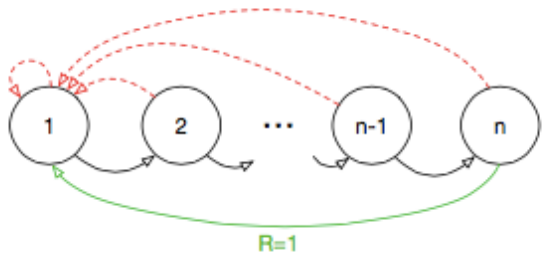
Table 2: Summary of normalized performance up to 30 minutes of play on 49 games with human starts. Results for DQN are from Nair et al. (2015).

3.2 Prioritized Replay

DQN 用了 Experience Replay 算法，将系统探索环境获得的样本保存起来，然后从中采样出样本以更新模型参数。对于采样，一个常见的改进是改变采样的概率。Prioritized Replay [3] 便是采取了这个策略，采用 TD-err 作为评判标准进行采样。

$$
TD - err = |r_{s,a} + \gamma max_{a'} Q(s', a') - Q(s, a)| \tag{2}
$$

下图是论文中采用的例子。例子中有 n 个状态，在每个状态系统一半概率采取 “正确” 或者一半概率 “错误”，图中红色虚线是错误动作。一旦系统采取错误动作，游戏结束。只有第 n 个状态 “正确” 朝向第 1 个状态，系统获得奖励 1。在这个例子训练过程中，系统产生无效样本，导致训练效率底下。如果采用 TD-err 作为评判标准进行采样，能够缓解这个问题。



论文报告了 Prioritized Replay 算法效果。从下图来看，Prioritized Replay 效果很好。

	DQN		Double DQN (tuned)		
	baseline	rank-based	baseline	rank-based	proportional
Median	48%	106%	111%	113%	128%
Mean	122%	355%	418%	454%	551%
> baseline	-	41	-	38	42
> human	15	25	30	33	33
# games	49	49	57	57	57

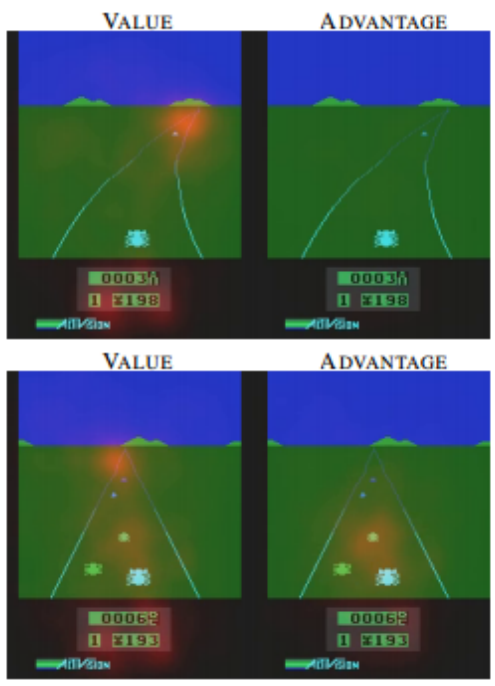
Table 1: Summary of normalized scores. See Table 6 in the appendix for full results.

3.3 Dueling Network

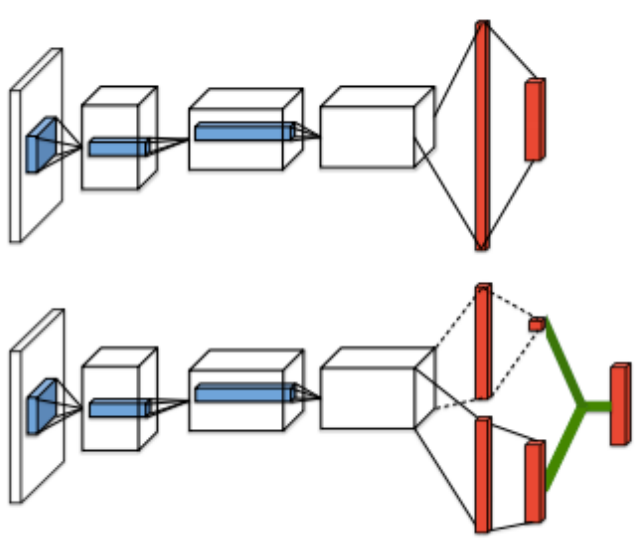
Baird 在 1993 年提出将 Q 值分解为价值 (Value) 和优势 (Advantage) [4]。

$$Q(s,a) = V(s) + A(s,a)$$

这个想法可以用下面的例子说明 [5]。上面两张图表示，前方无车时，选择什么动作并不会太影响行车状态。这个时候系统关注状态的价值，而对影响动作优势不是很关心。下面两张图表示，前方有车时，选择动作至关重要。这个时候系统需要关心优势了。这个例子说明，Q 值分解为价值和优势更能刻画强化学习的过程。



Wang Z 将这个 idea 应用在深度强化学习中，提出了下面的网络结构 [5]。



这种网络结构很简单，但获得了很好的效果。

	30 no-ops		Human Starts	
	Mean	**Median**	**Mean**	**Median**
Prior. Duel Clip	**591.9%**	**172.1%**	**567.0%**	**115.3%**
Prior. Single	434.6%	123.7%	386.7%	112.9%
Duel Clip	**373.1%**	**151.5%**	**343.8%**	**117.1%**
Single Clip	341.2%	132.6%	302.8%	114.1%
Single	307.3%	117.8%	332.9%	110.9%
Nature DQN	227.9%	79.1%	219.6%	68.5%

Dueling Network 是一个深度学习的网络结构。它可以结合之前介绍的 Experience Replay、 Double DQN 和 Prioritized Replay 等方法。 作者在论文中报告 Dueling Network 和 Prioritized Replay 结合的效果最好。

4. 总结

上次本来想把基于价值的深度强化学习的 Double DQN, Prioritized Replay 和 Dueling Network 也写了的，写到晚上 2 点。现在补上这部分内容。

从上面介绍来看，DQN、 Double DQN、Prioritized Replay 和 Dueling Network 都能在深度学习出现之前的工作找到一些渊源。深度学习的出现，将这些方法的效果提高了前所未有的高度。

文章结尾欢迎关注我的公众号 AlgorithmDog，每次更新就会有提醒哦~

欢迎关注

公众号讲述**机器学习**和**系统研发**的轶事，希望讲得有趣，**每周日更新**~

扫描二维码即可关注。您，不关注下么?

[1] S. Thrun and A. Schwartz. Issues in using function approximation for reinforcement learning. In M. Mozer, P. Smolensky, D. Touretzky, J. Elman, and A. Weigend, editors, Proceedings of the 1993 Connectionist Models Summer School, Hillsdale, NJ, 1993. Lawrence Erlbaum.
[2] Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep reinforcement learning with double Q-learning." CoRR, abs/1509.06461 (2015).
[3] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J]. arXiv preprint arXiv:1511.05952, 2015.
[4] Baird, L.C. Advantage updating. Technical Report WLTR-93-1146,
Wright-Patterson Air Force Base, 1993.
[5] Wang Z, de Freitas N, Lanctot M. Dueling network architectures for deep reinforcement learning[J]. arXiv preprint arXiv:1511.06581, 2015.

强化学习系列系列文章

- 强化学习系列之一:马尔科夫决策过程
- 强化学习系列之二:模型相关的强化学习
- 强化学习系列之三:模型无关的策略评价
- 强化学习系列之四:模型无关的策略学习
- 强化学习系列之五:价值函数近似
- 强化学习系列之六:策略梯度
- 强化学习系列之九:Deep Q Network (DQN)