

Abstract

经验回放技术能让人工智能（AI）有能力记住过去的游戏经历并可以重复使用，之前的研究都是从存储好的记忆空间中进行记忆采样。然而这么做却缺乏对游戏经历的重要性判断，可能游戏已经玩到第二关了，可是第一关的记忆还在影响现在的动作。这篇文章描述了我们新研究的经验优先级框架，能让更加重要的记忆更快频率的被使用，以提高学习效率。我们在DQN中使用这个记忆框架，让DQN得到了进一步升级，能泛化的玩耍41个游戏，我们一共实验了49个游戏。

1 INTRODUCTION

在线强化学习（RL）人工智能通过在经验中进行学习，并不断改善自己的参数。最简单的模式就是学习完后马上丢弃经验数据，这么做有两个严重的缺陷：一是强相关性更新会破坏随机梯度下降算法之中的一些假设，二是会快速遗忘以后可能用得到的罕见经验。

经验回放技术能够同时解决上述两个问题：经验会存储为可回放记忆，这样就能够通过混合近期回忆来更新，以打破时间上的相关性；罕见的经验也有机会能够保存下来再次使用。以上这些在DQN算法的文章中都是有描述过的，这些技巧能够稳定神经网络训练中的值函数更新。DQN使用了一个很大的记忆存储空间，然后从中进行随机采样。一般来说，经验回放能够降低学习用的经验总数，而换成更多的计算与内存消耗作为代价。

我们研究了如何进行经验的有选择回放可以使系统性能超越统一性回放。我们的主要思路就是AI能够从某些重要转换中更加有效的学习。游戏中的经历可能会让人感到惊讶、多余或与任务相关。某些经历可能暂时不会对AI有用，但当AI能力增强时可能就会用到这些经历（Schmidhuber，1991）。经验回放使AI按照经历的确切顺序处理记忆，优先经验回放则进一步使AI免于考虑与经历频率相同的记忆。

实际上我们绝对希望得分较高的那部分记忆能优先被采样，通常在DQN中这种高得分体现的TD（时间差分）误差上，TD误差越低得分越高。然而这种带优先级的采样可能会导致数据多样性的丧失，即某些得分较低的记忆其实也有它们存在的意义。所以我们使用随机优先级来减轻这种差异，并且引入偏置，然后通过重要性采样来不断优化。我们最终得到的结果是在Atari 2600 游戏环境下鲁棒性与扩展性的提高，这体现在更快的训练速度与绝无仅有的优秀成绩。

2 BACKGROUND

众多神经科学研究人员已经发现了足够的证据，证明啮齿动物的海马体存在经验回放机制，促使动物在清醒稳定或睡眠状态下回放大脑中所存储的先前记忆序列，尤其是那些跟奖励相关的记忆尤其容易被回放（例如有的犬类在遇到你的第一天会感觉到很害怕，但如果你在它附近休息一晚，它就会把这种事情看作一种奖励，第二天你会发现它对你友善的多）

众所周知，像值迭代这样的规划算法对于带有优先级的更新会显得算法效率很高。优先级交换（Prioritized sweeping）这种技术就能够根据值的变化来确定优先级，而TD误差很好的提供了这个处于变化的值。不过这类技术通常是适用于基于模型的规划，我们提出的方法与之类似，但可以泛化到无模型规划。此外，我们使用随机优先机制，当从样本中学习函数逼近时，这种方式更加稳健。

TD误差也被用作确定资源注意力位置的优先级机制，例如想要确定怎样的特征需要被探索或选择。

对于监督学习，有很多技术来处理那些分布不均匀的带标签数据，例如重采样、下采样与过采样。针对深度强化学习的重采样技术有一篇最新的论文(Narasimhan et al., 2015)，该方法将经验分为两个桶，一个用于正面奖励，另一个用于负面奖励，然后从每个桶中选取一个固定的记忆部分进行回放。这种方法适用于像我们一样的领域，即经验本身会具有好与坏的性质。此外，大名鼎鼎的Hinton在2007年介绍了一种基于误差的非均匀采样的方法，结合重要性采样的优化，能够让MNIST手写数字识别任务提高3倍的速度。

截至目前已经有很多关于Atari 2600 游戏的强化学习AI研究了，包括DQN与DDQN（Double DQN）等，都取得了不错的成果。和我们一同工作的团队也在研究将优势分析与值计算分解开来的方法，并已经能够让Atari基线测试有了重大的改进（意指决斗网络）。

3 PRIORITIZED REPLAY

实施经验回放意味着需要做好两种选择：哪些经验应该被储存，哪些经验应该被拿出来进行回放。本文主要针对后者，尽可能挑选出有价值的意义，提供给AI进行学习，尽管这一过程处于控制之外。

**3.1 A MOTIVATING EXAMPLE**

为了理解优先级机制带来的潜在收益，我们先建立一个形似图1左的名为盲道（Blind Cliffwalk）的环境，其奖励分布极为稀少，只有通过更多探索才有可能获得奖励。这个环境具有n个状态，AI往往需要行走指数级的时间步才能够获得第一次非零奖励。经过实践，如果仅采取随机行动，那么AI能够拿到奖励的概率为。与这种环境难度相似的实际案例就是双足式机器人在学会稳定行走前所需要经历的大量摔倒次数。

我们建立了两个AI，并使用上述环境来区别两个AI的性能。两个AI均使用同样的记忆存储，且都构建为同样的DQN模型。第一个AI使用纯随机记忆选取，第二个AI调用预言过程来获取优先级。预言过程能够选择出对当前状态造成最大梯度下降的记忆，具体参见附录B。图1右显示了我们的方法能够为训练过程提供明显的加速。

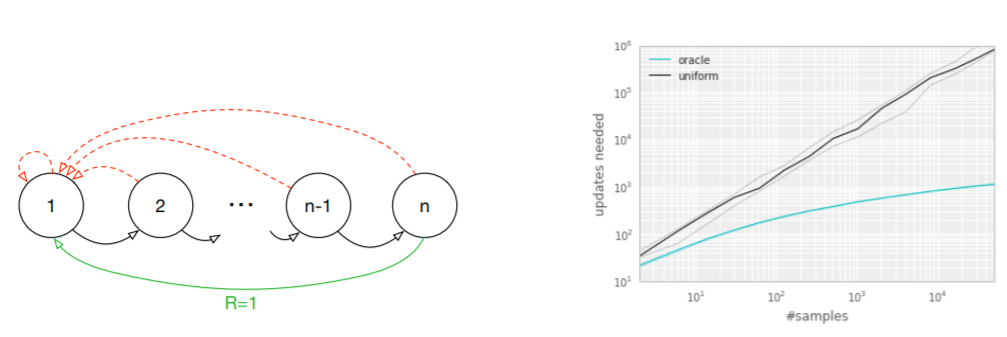


图1

**3.2 PRIORITIZING WITH TD-ERROR**

优先级重放的核心部分就是需要为衡量每个记忆的重要性制定标准，一个理想化的标准可以是评估AI在当前状态从某个记忆中能够学习的信息总量。当然这种理想化的标准是不可测量的，但是我们可以寻找替代物：测量这个记忆的TD误差，因为这个误差意味这AI对这个记忆的惊喜程度，在Andre等人在1998年的研究就发现这样能够测量出现在与下一时间步的自举估计的距离之差。这样的方法适用于在线强化学习算法。

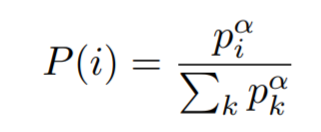
为了量化基于TD误差的优先级记忆采样的潜在实力，我们在上面的盲道环境中再次部署了同样两种AI，并使用了一种“greedy TD-error prioritization”算法。这种算法能及时的将TD误差与对应的记忆保存下来，相当于记忆与TD误差相互加权了，然后Q-learning算法就可以通过更新TD误差来为记忆配比权重，从而为记忆区分优先级。由于新到的记忆并没有对应的TD误差（针对这个时间步的反向传播还没开始），所以我们给予新记忆最高优先级，以保障所有的记忆至少都被AI看过一次。图2左可以看到这个方法也比较节省训练时间。

**Implementation:** 为了能够伸缩至内存大小N，我们使用二进制的堆作为优先级队列的数据结构，采样时就能在复杂度O(1)进行记忆搜索，更新为复杂度O(log N)，具体参加附录B.2.1。

**3.3 STOCHASTIC PRIORITIZATION**

上述方法可以称为基于贪婪优先级TD误差（greedy TD-error prioritization）采样方法，然而这个方法仍存在一些问题。首先为了避免记忆存放的内存溢出，TD误差仅在需要回放的记忆进行更新。这样做的结果就是新到的记忆在第一次访问时具有较低的TD误差，从而导致很长一段时间内这个记忆不会被回放。其次，如果奖励是随机的，那么使用这个方法的AI就会很容易收到噪音的影响，当估计的误差值看起来像一个噪音时，自举学习过程就会越来越恶化。最后，上述方法适用于经验的一个小范围子集，特别是误差变化较慢的那种，尤其是做函数逼近时，最初的高误差记忆会经常被重放，这时记忆采样会缺乏多样性，最终导致过拟合。

为了克服这些问题，我们引入了一种随机采样方法，该方法在纯贪婪优先级和均匀随机采样之间进行插值。我们确保采样的概率在转换优先级中是单调的，同时保证即使对于最低优先级转换也是非零概率。具体来说，我们将采样转换的概率i定义为下面这样：



上式中， 是记忆i的优先级，永远大于0。 则决定优先级的整体配比，其值为0时意味着无优先级。对于的选择我们有两种方式，proportional方式使用这个公式进行选择：

 ，即残差的绝对值加一个小范围浮动的非零随机数，以解决残差为0导致该记忆被选取概率为0的情况；而另一个方法rank-based则使用该记忆的秩：。详细实验结果见图2右，

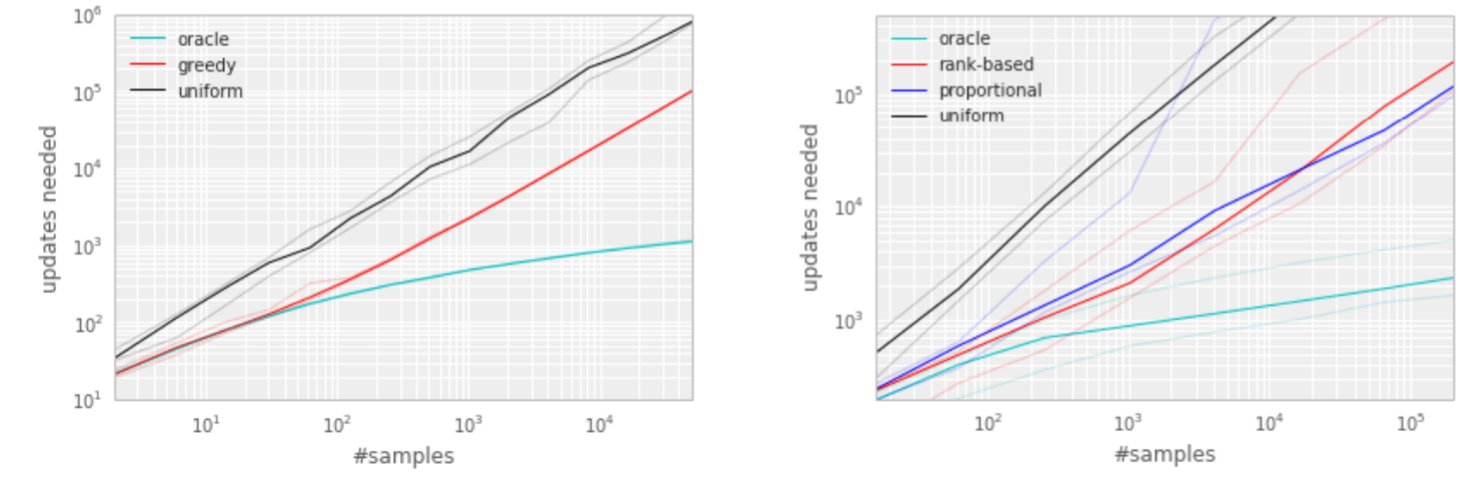
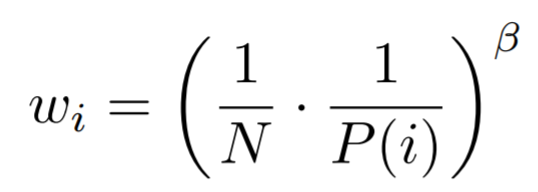


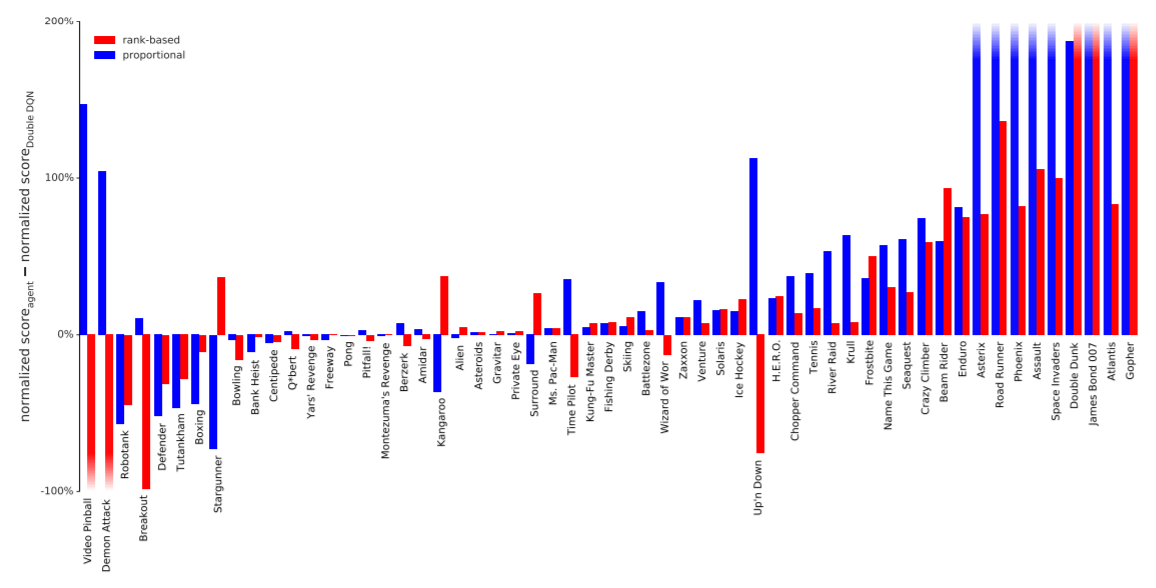
图2

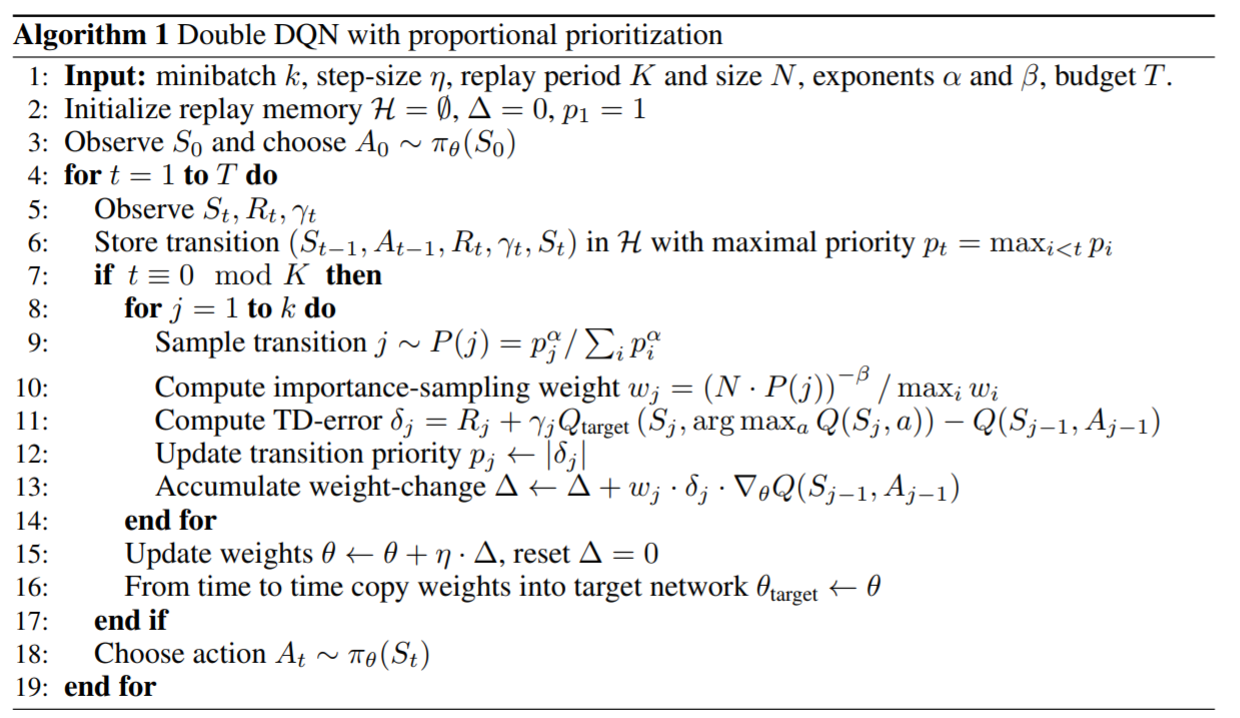
**3.4 ANNEALING THE BIAS**

随机更新对期望值的估计依赖于与其期望相同的分布的那些更新。优先级回放引入了偏置，由于它以不受控制的方式改变了这种分布，从而把函数估计的可收敛解也改变了（即使策略和状态的分布是固定的）。我们可以通过使用重要性采样（IS）权重来纠正这种偏差。



非线性函数逼近（例如深度神经网络）与优先级回放相结合时，重要性采样具有另一个好处：较大的步长可能非常具有破坏性，因为梯度的一阶近似仅在局部可靠，如果不用重要性采样则必须使用较小的全局步长来做防护。我们将两种采样方法结合DQN都做了实验，57个Atari游戏的结果如下图：





详细的实验结果就不再论述了，毕竟看论文的目的只是学习其核心技术，希望你学的开心，并勇于实践。