**强化学习在游戏AI中的应用调研报告**

计算机技术2班 2111812118 沈正圆

摘要：理论上，强化学习可用于尝试具有明确目标的算法模型，例如游戏AI。 然而，实际上，在游戏AI中应用强化学习更常用于验证和调试算法模型的准确性。 一些强化学习算法已经通过传统的简单游戏验证。 对于像英雄联盟和王者荣耀这样复杂的实时战略游戏，算法的设计相对复杂。 在本文中，我们讨论了基于Dota游戏AI的强化学习算法设计的一些想法。

关键字：强化学习 Dota游戏AI 算法思路

**1 强化学习**

强化学习受到心理学行为主义理论的启发，即生物体在环境奖励或惩罚的刺激下逐渐形成对刺激的期望，并产生能够获得更大利益的习惯性行为。该方法具有通用性，已在博弈论、控制论、运筹学、信息论、仿真优化方法、多智能体系统学习、群体智能、统计学和遗传算法等许多领域得到研究。强化学习也是多学科交叉的产物。其本质是解决“决策”问题，即学会自动决策。在计算机科学领域，它体现为一种机器学习算法；在神经科学领域，它体现在理解人脑如何做出决定。主要研究的是反馈系统；在心理学领域，如何对动物进行决策，是什么导致了动物在经济学领域的行为，这反映在博弈论的研究中。所有这些问题最终归结为一个问题：为什么人们可以做出最佳的决定，以及人类如何做。强化学习作为一个连续的决策问题，需要对行为进行连续的选择，并且作为一个更好的结果，这些行为会带来更大的好处。它不会告诉算法要做什么，先尝试做一些行为，然后得到结果。通过判断结果是对还是错来反馈以前的行为。此反馈用于调整以前的行为，通过连续调整算法，可以了解在什么情况下选择哪种行为可以获得更好的结果[1]。

强化学习不同于连接主义学习中的监督学习，主要表现在教师信号上，强化学习中由环境提供的强化信号是Agent对所产生动作的好坏作一种评价(通常为标量信号)，而不是告诉Agent如何去产生正确的动作。由于外部环境提供了很少的信息，Agent必须靠自身的经历进行学习。通过这种方式，Agent在行动一一评价的环境中获得知识，改进行动方案以适应环境。

强化学习系统学习的目标是动态地调整参数，以达到强化信号最大。若已知r/A梯度信息，则可直接可以使用监督学习算法。因为强化信号r与Agent产生的动作A没有明确的函数形式描述，所以梯度信息r/A无法得到。因此，在强化学习系统中，需要某种随机单元，使用这种随机单元，Agent在可能动作空间中进行搜索并发现正确的动作。

**2 强化学习方法**

**2.1 Model-free和Model-based**

如果不理解环境，环境给了什么就是什么，我们就把这种方法叫做 Model-free，这里的 Model 就是用模型来表示环境，理解环境就是学会了用一个模型来代表环境，所以这种就是 Model-based 方法。

Model-free 的方法有很多, 像 Q learning、Sarsa、Policy Gradients 都是从环境中得到反馈然后从中学习。而 Model-based RL 只是多了一道程序，为真实世界建模，也可以说他们都是 Model-free 的强化学习, 只是 Model-based 多出了一个虚拟环境，我们可以先在虚拟环境中尝试，如果没问题，再拿到现实环境中来。最终 Model-based 还有一个杀手锏，是 Model-free 所不具备的，那就是想象力.

Model-free 中, 计算机只能按部就班，一步一步等待真实世界的反馈，再根据反馈采取下一步行动。而 Model-based，能通过想象来预判断接下来将要发生的所有情况，然后选择这些想象情况中最好的那种，并依据这种情况来采取下一步的策略，这也就是围棋场上 AlphaGo 能够超越人类的原因[2]。

**2.2基于概率和基于价值**

基于概率是强化学习中最直接的一种，他通所处的环境，输出下一步要采取的各种动作的概率，根据概率采取行动，所以每种动作都有可能被选中，只是概率不同。而基于价值的方法输出则是所有动作的价值，根据最高价值来选着动作，相比基于概率的方法，基于价值的决策部分更为肯定，就选价值最高的，而基于概率的，即使某个动作的概率最高, 但是还是不一定会选到他.

**2.3基于概率和基于价值**

强化学习还能用另外一种方式分类，回合更新和单步更新。假设强化学习就是在玩游戏，游戏回合有开始和结束。回合更新指的是游戏开始后，需要等待游戏结束再总结这一回合，再更新我们的行为准则。而单步更新则是在游戏进行中每一步都在更新，不用等待游戏的结束，这样边玩边学习[3]。

**2.3在线学习和离线学习**

所谓在线学习，就是指必须本人在场，并且一定是本人一边行动边一学习。而离线学习是你可以选择自己行动，也可以选择看着别人行动，通过看别人行动来学习别人的行为准则，离线学习 同样是从过往的经验中学习，但是这些过往的经历没必要是自己的经历，任何人的经历都能被学习。

**3强化学习应用于AI游戏中**

虽然在传统的机器学习分类中没有提到强化学习，但它实际上可以帮助解决深度学习中智能控制和分析与预测领域的许多实际问题，例如通过用户反馈优化人与计算机之间的多轮对话策略。 ，以及自然语言的产生。 然而，目前，在诸如Dota这样的“复杂”游戏中训练高级人类AI非常困难，但是这里讨论了设计算法的想法和注意事项。

**3.1“类人”定义**

人工智能的行为具有动态性和记忆性，可以根据新的环境信息进行调整。类似于《王者荣耀》游戏的Dota，推测其内置的人机战AI并不是强化学习（“推定”并不意味着一定），而是可以通过大量的逻辑积累，执行逻辑是固定的而不是动态的，玩家和开发者定制的AI也可以用Dota bot脚本来完成。人工智能实现最终目标的手段是政策，而不是通过使用计算机来超越人们的“手速”优势来赢得胜利。虽然计算机具有极快的速度控制能力，但这并不一定代表智能，因此代理必须以“每分钟动作”的方式与游戏进行交互，这受到人类灵活性的限制[4]。

**3.2Dota这类游戏的复杂性**

如果只考虑一个AI（Agent）进行强化学习（实际情况是需要协作和团队智慧的多代理系统），则需要明确定义：环境，代理状态。 RL的传统简单游戏，例如“Cart Pole”和“Flappy Bird”，在时空抽象中很简单，并且它们的环境完全可以观察到目标。对于像Dota这样的游戏，一方只能根据友方单位的视觉看到地图，因此环境是部分可观察的，这需要一些动作来探索阴影地图。对于状态的抽象描述，基本思想是参考二维矩阵描述地图（方便作为DQN的输入）[5]。

Dota游戏的复杂性在于无法在单个空间坐标中描述状态。状态的所有潜在“特征”至少包括每个单位的位置和攻击，地图的视觉范围（白天和黑夜不同），当前携带物品和每个单位的使用范围，技能和铸造范围每个单位，每单位的蓝色血液等，所有这些都会影响动作决定。为了决定，这些“特征”需要分成未使用的层。每层都类似于DeepMind团队用于星际争霸的低分辨率RGB图像数据图和迷你图。与RL的传统简单游戏不同，例如Flappy Bird，它的动作从左到右或从上到下移动。像王者荣耀这样的游戏的复杂性在于有许多动作，例如基本动作：移动类中有八个移动方向，以及常见的攻击，使用技能，买卖物品，使用物品等等。似乎基本动作的数量不是很多，但简单来说，Flappy Bird的基本动作只有两个：向上或向下，但是多步计划环境将基本动作序列视为一个成员。真正考虑的行动，例如当n = 20时，基本动作为up，down两种移动，并且还有2 ^ 20 = 1 E6个动作。数量n。至于为什么我们应该使用一系列基本动作作为一个实际行动，考虑到n长度n是1还是大于1，完成目标所需的相对实际步骤是n的倒数，我们是否需要考虑n是否行动顺序取决于实验的实际情况和问题。例如，真正的玩家对王者荣耀的操作将是一系列基本动作，例如不去友方单位和攻击，不买卖物品，所以如果实际动作使用n长动作序列，在这种情况下在大量案例中，有一些有针对性的解决方案：使用Wolpertinger框架，即首先将状态与连续动作空间对应。对于属于连续空间的原型行为。然后将其映射到原始动作空间。实际上，它是找到K最接近的原始动作，框架或actor-critic框架，更多的是这种离散到连续，连续到离散的变换，它将能够处理大规模的离散问题。另一方面，我们需要考虑：基于多类分类的思想，基于纠错输出码解决由多类分类引起的规模问题（ECOC是多类分类与二元分类器相结合的优化）当类对数很大时，将学习复杂度从O（A2）降低到O（A log（A）和O（log（A））。

在Dota中，最终目标必须是我们击败敌人基地的胜利，但如果我们直接设定目标，奖励将被推迟很多。因此，在稀疏反馈的情况下，代理不能很好地学习，因此提出了一种分层-DQD（h-DQN）框架，它也可以用于Dota游戏中。例如，为了达到击败对手的基础以获胜的最终目标（最终目标），它可以列出由时间尺度产生的一些内在目标。例如，在大约五分钟内，该团将摧毁另一侧，并在大约十分钟内，它将推倒另一侧的几个防御塔等。

**4强化学习缺点**

研究人员在最近半年开始了对DRL的反思。由于发表的文献中往往不提供重要参数设置和工程解决方案的细节，很多算法都难以复现。2017年9月，著名RL专家Doina Precup和Joelle Pineau所领导的的研究组发表了论文Deep Reinforcement Learning that Matters，直指当前DRL领域论文数量多却水分大、实验难以复现等问题。该文在学术界和工业界引发热烈反响。很多人对此表示认同，并对DRL的实际能力产生强烈怀疑。

其实，这并非Precup& Pineau研究组第一次对DRL发难。早在2个月前，该研究组就通过充足的实验对造成DRL算法难以复现的多个要素加以研究，并将研究成果撰写成文Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control。同年8月，他们在ICML 2017上作了题为“Reproducibility of Policy Gradient Methods for Continuous Control”的报告，通过实例详细展示了在复现多个基于策略梯度的算法的过程中，由于种种不确定性因素导致的复现困难。12月，在万众瞩目的NIPS 2017 DRL专题研讨会上，Joelle Pineau受邀作了题为“Reproducibility of DRL and Beyond”的报告。报告中，Pineau先介绍了当前科研领域的“可复现性危机” ：在《自然》杂志的一项调查中，90%的被访者认为“可复现性”问题是科研领域存在的危机，其中，52%的被访者认为这个问题很严重。在另一项调查中，不同领域的研究者几乎都有很高的比例无法复现他人甚至自己过去的实验。可见“可复现性危机”有多么严峻！Pineau针对机器学习领域发起的一项调研显示，同样有90%的研究者认识到了这个危机。

随后，针对DRL领域，Pineau展示了该研究组对当前不同DRL算法的大量可复现性实验。实验结果表明，不同DRL算法在不同任务、不同超参数、不同随机种子下的效果大相径庭。在报告后半段，Pineau呼吁学界关注“可复现性危机”这一问题，并根据她的调研结果，提出了12条检验算法“可复现性”的准则，宣布计划在ICLR 2018开始举办“可复现实验挑战赛”（“可复现危机”在其他机器学习领域也受到了关注，ICML 2017已经举办了Reproducibility in Machine Learning Workshop，并将在今年继续举办第二届），旨在鼓励研究者做出真正扎实的工作，抑制机器学习领域的泡沫。Pineau & Precup研究组的这一系列研究获得了广泛关注。

**5总结**

虽然完全类人在Dota游戏中中的实施难度大，而现实世界中更难发挥人工智能的作用。但人们必须像RL中的agent一样继续探索，以实现最终的目标。

**参考文献**

[1] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540 (2015): 529.

[2] Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." Nature 529.7587 (2016): 484-489.

[3] Silver, David, et al. "Mastering the game of go without human knowledge." Nature 550.7676 (2017): 354.

[4] Levine, Sergey, et al. "End-to-end training of deep visuomotor policies." arXiv preprint arXiv:1504.00702, 2015.

[5] Mao, Hongzi, et al. "Resource management with deep reinforcement learning." Proceedings of the 15th ACM Workshop on Hot Topics in Networks. ACM, 2016.

[6] https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

[7] Jaques, Natasha, et al. "Tuning recurrent neural networks with reinforcement

learning." (2017).