ISSN: 2661-3719(Print); 2661-3727(Online)

人工智能中深度强化学习算法的探索与优化

侯馨玉
皖江工学院　安徽省马鞍山市　243000

摘　要：深度强化学习作为人工智能领域的关键分支，把深度学习感知能力和强化学习决策能力结合在一起，致力于解决复杂环境下的序列决策问题。本文主要关注于该领域的核心探索策略和算法优化技术，系统地论述了基础理论、主流方法以及稳定性的提高途径。文章首先解释了深度强化学习的基本架构和核心概念，然后深入分析各种探索策略的原理及其优势，仔细探讨了价值函数和策略梯度的优化手段，解析了经验重放、目标网络等关键机制对训练稳定性的影响。最后本文分析了性能评价指标体系以及基准测试的方法，给算法性能提供客观衡量的理论基础。

关键词：深度强化学习；探索和利用；算法优化；价值函数；策略梯度；训练稳定；性能评估

引言

随着时代的发展，人工智能 技术更加强大，各行各业都在探索与人工智能接轨的应用。人工智能的其中一个重要目标就是让智能体能够自主做出决策、学会在复杂环境中行动。深度强化学习将深度学习强大的表达能力与强化学习序列决策框架融合，成为这一目标的重要路径。深度强化学习算法训练过程中会遇到探索效率低、优化目标不稳定、收敛性难以保证等很多问题。这些挑战限制了算法在更大范围的应用效果及可靠性。从而对深度强化学习中的探索方式加以仔细的研究分析，并开展对算法优化和稳定化技术的研究，它具有理论价值，也有实际意义。本文围绕探索与优化两个核心议题，展开系统性梳理论述，为后续研究提供清晰理论框架与方法论基础。

1 深度强化学习基础

1.1 强化学习核心概念

强化学习框架的核心就是智能体和环境的交互。智能体用执行动作去影响环境，环境再反馈给智能体新的状态和奖励信号。奖励信号是一个标量值，用来衡量动作的好坏。

智能体要学到策略，使得长短期累计奖励最大。策略给出了在某状态下选择动作的规则。价值函数是用来预测未来累积奖励，用来评价状态或者状态 – 动作对的长期价值。马尔可夫决策过程给强化学习提供了一个形式化、数学化的模型。该模型认为环境是马尔科夫性的，也就是说下一个状态和奖励只跟当前的状态和采取的动作有关，跟之前的状态无关。这个假设成为很多强化学习算法理论分析的基础。

1.2 深度学习技术基础

在人工智能领域，学习算法无疑是核心组成部分，其本质上是统计学的应用。它是一种在分类、预测、 随机分布等常见问题上，融合和借鉴统计学理论来进行数据归纳和分析，深度分析数据的内涵、数据背后蕴含常见算法。从而更加精妙地解决问题的方法。深度学习为强化学习提供了解决高维状态空间的强大工具。深度神经网络有强大的函数近似能力，可以将高维度的原始输入数据映射成低维度的特征表示，或者近似价值函数、策略函数。卷积神经网络经常用来处理有空间结构的状态输入，比如图象信息。它的局部连接、权值共享特性可以有效抽取空间特征。循环神经网络适合处理带有时间序列相关性状态信息。神经网络训练通常用梯度下降算法来优化。反向传播算法可以用来有效计算损失函数关于网络参数的导数。优化器的选择，例如 Adam 或者 RMSProp，对于训练的效率以及最终的效果都有很大的影响。

2 深度强化学习中的探索策略

2.1 探索与利用平衡问题

探索和利用之间的平衡是强化学习中一个根本性难题。利用是智能体根据现有的知识选取预期会带来最大回报的动作。探索是指智能体选择当前情况下非最优动作以获得更多信息。过度使用会导致智能体陷入局部最优，不能发现可能带来更大长期收益的更佳策略。过度探索会使得学习效率降低，智能体会浪费很多时间去试那些已经知道的差动作。所以设计有效的探索策略十分重要。该问题的复杂性在于权衡是动态变化的。学习初期智能体要多探索以快速掌握环

境。随着学习的深入，智能体应该逐步增加利用的比例来稳定策略提高性能。自动化这个过程的研究重点。

2.2 随机探索策略方法

随机探索是最直接最基本的探索方法。它的基本想法是在决策时加入一些随机因素，以让智能体可以接触到非最优的动作。ε – 贪心策略是最著名的一种。ε – 贪心策略中，智能体以概率选择当前认为最优的动作，同时以概率选择另一个动作。概率会随着时间而衰减，实现探索和利用之间的自然转变。该策略容易实现，但是探索效率比较低，因为选择随机动作时并没有考虑到动作的潜在价值。高斯噪声也是一种常见的随机探索方法，在连续动作空间里用的比较多。通过给确定性策略的输出动作加上均值为 0 的高斯噪声实现探索。噪声的幅度可以调节来控制探索的程度。

3 深度强化学习算法的优化研究

3.1 价值函数优化方法

价值函数优化的目标就是能够准确地估计出一个状态或者一个状态 – 动作对的长期价值。深度强化学习的意思就是训练一个深度神经网络去逼近真实的价值函数。优化过程的核心就是最小化时序差分误差。双 Q 学习是针对深度 Q 网络中价值过高估计问题重要优化。它通过解耦动作的选择与价值评估从而减少估计误差。即用一个网络选出最优的行动，用另外一个目标网络估算此行动的价值，这样就能得到更为保守和精确的价值估计。优先级经验回放是均匀采样的优化。按照时序差分误差大小给经验回放缓冲区中的样本分配不同的采样优先级。误差越大的样本，认为它对学习价值函数的贡献更大，所以被采样的概率更大，提高数据利用率。

3.2 策略梯度优化技术

策略梯度方法通过直接对期望累计奖励的表达式做梯度上升来优化策略参数。它的优点是能自然处理连续动作空间和随机策略。关键在于怎样有效估计策略梯度。信赖域策略优化通过约束每次策略更新步长的方法保证训练稳定。它通过优化替换损失函数，保证新策略和旧策略的性能不相距太大，避免了训练过程中的大起大落和性能崩溃。极大地增加了策略梯度方法的鲁棒性。近端策略优化算法是另外一种被广泛应用的策略优化方法。用裁减概率比的方式来间接完成信赖域约束，实现简单于信赖域策略优化，而且实验结果表现优异，是目前最主流的策略优化算法之一。

4 深度强化学习的稳定性提升研究

4.1 经验回放机制改进

经验回放存储并利用以往的经验，打散了经验的顺序关系，使训练过程更接近监督学习所需要的独立同分布假设。这可是稳定深度 Q 网络训练的重头发明。而标准经验回放缓使用先进先出的队列结构，均匀随机采样。这可能导致一些重要的、信息量大的经验很快就会被掩盖，不能得到充分的学习。缓冲区的大小也需要谨慎设置，缓冲区太小会减少经验多样性，太大又会造成学习拖延。为了克服均匀采样的不足，优先级经验回放被提出。它会根据每个样本的时序差分误差给它们分别一个优先级，误差大的优先级就高。采样概率和优先级成正比。同时为了抵消优先级采样带来的偏差，在计算梯度更新时引入了重要性采样权重来保证收敛的正确性。

4.2 目标网络优化策略

在深度 Q 学习中，目标值是由当前网络本身来计算出来的，这就造成了移动目标问题，也就是优化的目标会随着要优化的参数而变化，就好像在追逐一个移动的目标一样，很容易造成训练的不稳定和发散。目标网络用提供一段时期内的稳定的目标值的方式来解决这个问题。它就是一个和现在的网络结构一样，但是更新速度慢一点的副本。计算目标 Q 值时用目标网络的参数而不是用当前网络的参数，使得目标值在短时间内保持稳定，从而大大提高优化的稳定性。目标网络的更新策略主要分为两种，一种是硬更新，另一种是软更新。硬更新是每隔一定的步数就把当前网络的参数全部复制给目标网络。软更新就是在每一次训练之后，使目标网络的参数逐渐接近当前网络参数。软更新能提供更平滑、更稳定的改变目标。

4.3 梯度裁剪与归一化方法

梯度爆炸问题在深度神经网络训练中非常普遍，特别是在深度强化学习中。梯度更新太剧烈会毁掉有价值的策略，引起训练出错。梯度裁剪就是用一个阈值去限定更新梯度向量的范数，不能超过这个阈值。当梯度的模超过预设的阈值时，梯度裁剪就将整梯度向量按比例缩小，使它的模等于阈值。这是一种简单有效的正则化方法，能防止参数更新步幅过大，并保证训练过程平稳运行，在 actor–critic 模型里，策略网络更易受到保护。梯度归一化与裁剪不同，裁剪是直接限制梯度的大小，梯度归一化则是通过调整优化器的内部

状态来适应梯度的变化。对网络输入（如状态）或网络输出（如奖励）进行归一化操作，使二者都为零均值、单位方差，这也会有效改善优化 landscape，使得梯度更稳定从而加速收敛。

5 深度强化学习的性能评估

5.1 评估指标与测试环境

评价深度强化学习算法性能需要一套综合的指标体系。学习曲线是最核心的评价工具，展示的是智能体累积奖励随训练步数或者环境交互次数变化的趋势。其收敛速度、最终性能和稳定性是重要的评判标准。不仅仅是最后的性能，数据效率也非常重要。衡量智能体达到某种性能水平所要的环境交互量或者样本数量。高数据效率意味着算法可以更快的从经验中学习。因为现实世界中的交互成本很高，所以这对实际应用来说非常关键。测试环境必须和训练环境分开，在评估的时候固定智能体策略参数，使得可以公平的评价其学到的策略。为了保证结果的可靠性，要使用不同的随机数生成器，在各个随机数生成器下运行实验并报告出各组实验的均值和标准差，来评价算法的鲁棒性和可重复性。

5.2 基准测试与对比分析

基准测试的环境为不同算法性能对比提供了一个统一的平台。环境一般包含不同的挑战维度，例如部分可观性、延迟奖励、高维动作空间、多任务学习等等，从而全方位检测出算法的普适性和强大程度。做算法比较时要严格控制实验条件一致。包括网络模型结构、计算资源、训练轮数以及评估频率等等都要尽可能的一致。不公平的设计，容易得出有误导性的结论。严谨的对比分析应该从算法的核心思想来比较，而不是工程实现的细小区别。对比分析不能仅仅对性能高低进行简单的排序，还应深入分析不同算法在不同类型环境下的表现差异。一种算法在稀疏奖励环境里表现出色，另一种在需要细致操作的领域里有其长处。分析可以为某一问题选择适合的算法。

5.3 实际应用中的性能因素分析

在将深度强化学习算法从实验室环境迁移到实际应用的时候，就会出现一些新的性能指标。安全问题要放在首位。探索过程中的不安全行为在模拟器中可以接受，在现实生活中会造成严重的后果。所以需要研究出约束强化学习等方式来。算法的可解释性直接关系到其可信度。实际应用中，用户要理解智能体为什么做出某一个决定。黑箱模型不易诊断故障，也得不到信任。因此提高模型决策过程的透明度是一个很重要的研究方向。计算效率以及实时性成为又一个棘手的问题。许多算法训练需要大量交互数据和巨大计算开销。在实际部署时，特别是在边缘设备上，必须要对模型做压缩、加速等操作，以满足实时决策所要求的低延迟。要在性能和效率之间寻到一个新的平衡。

结语

深度强化学习算法探索优化，内容丰富，挑战性大。本文对从基础框架、探索策略、优化方法、稳定性技术到性能评估的整个链条的关键技术进行了系统梳理。探索与利用的均衡是智能学习的主要矛盾，价值函数优化和策略梯度优化则是提升算法性能的两大支撑。经验回放、目标网络等稳定性技术为复杂模型训练提供必要保障。虽然取得了一定的成绩，但是在该领域依然还存在很多的问题。如何实现更高效、更有针对性的探索，设计更稳定高效、更具可行性的优化算法，建立更全面、严格、严谨的评价体系，都是未来需要持续进行的研究方向。这些问题的解决会把深度强化学习推进到成熟、实用的新阶段，最终在更广泛的现实世界中实现其赋能价值。

参考文献：

[1] 魏子杨 . 机器学习算法在人工智能中的应用 [J]. 华东科技 ,2023,(05):107-109.

[2] 巫小勇 . 数学建模在人工智能与机器学习中的应用研究 [J]. 科技资讯 ,2025,23(06):46-48.

[3] 黄国盛 . 机器学习算法在人工智能中的应用 [J]. 集成电路应用 ,2022,39(09):192-193.

[4] 冯蓉 . 机器学习算法在数据挖掘中的应用 [J]. 中国高新科技 ,2022,(20):30-32.

[5] 余涛 , 贾如春 . 基于机器学习算法人工智能技术的发展与应用 [J]. 数学学习与研究 ,2019,(13):149.

作者简介： 侯馨玉 , 女 , 汉族 , 安徽省滁州市，本科 , (在校学生) 无职称 ，研究方向：人工智能