深度强化学习与移动通信资源管理:算法、进展与展望

孙恩昌[1,2,3], 袁永仪[1,2], 吴 兵[1,2], 屈晗星[1,2], 张延华[1,2]
(1. 北京工业大学信息学部, 北京 100124; 2. 先进信息网络北京实验室, 北京 100124;
3. 北京工业大学北京-都柏林国际学院, 北京 100124)

摘 要: 深度强化学习(deep reinforcement learning, DRL)将深度学习从高维数据提取低维特征的能力与强化学习的决策能力相结合,是移动通信资源管理与优化的高效算法之一. 在引入DRL相关算法概念与原理的基础上,重点对DRL在网络切片、云计算、雾计算、移动边缘计算等通信技术与场景中的资源管理与优化效果进行综述与分析,结合DRL在移动通信资源管理的算法原理与研究进展,论述了DRL面临的问题与挑战,并提出相应解决思路. 最后,展望了DRL在移动通信资源管理领域的发展趋势和主要研究方向.

关键词: 深度强化学习(deep reinforcement learning, DRL); 通信资源管理; 网络切片; 云计算; 雾计算; 移动边缘计算
中图分类号: TN 929.5 **文献标志码**: A **文章编号**: 0254-0037(2023)01-0071-18
doi: 10.11936/bjutxb2021040026

Deep Reinforcement Learning and Mobile Communication Resource Management: Algorithms, Progress, and Prospects

SUN Enchang[1,2,3], YUAN Yongyi[1,2], WU Bing[1,2], QU Hanxing[1,2], ZHANG Yanhua[1,2]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Beijing Laboratory of Advanced Information Networks, Beijing 100124, China;
3. Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China)

Abstract: As one of the highly-efficient algorithms for resource management and optimization in mobile communications, deep reinforcement learning (DRL) integrates the ability of deep learning to extract low dimensional features from high dimensional data with the decision-making ability of reinforcement learning. First, the concepts and principles of DRL algorithms were introduced. Then, the resource management and optimization effect of DRL in different scenarios were summarized and analyzed. The technologies and scenarios included network slicing, cloud computing, fog computing, and mobile edge computing. Furthermore, based on the key research progress of DRL in mobile communication resource management, the open issues and challenges of DRL were discussed, and possible solutions were proposed. Finally, development trends and key research directions in the field of mobile communication resoure management were prospected.

Key words: deep reinforcement learning (DRL); communication resource management; network slicing; cloud computing; fog computing; mobile edge computing

随着通信设备的不断增加和用户业务需求的多样化,传统的通信方案已然落后于日益革新的应用

收稿日期: 2021-04-26; 修回日期: 2021-10-20
基金项目: 国家自然科学基金资助项目(61671029);中国国家留学基金高等学校骨干教师研修项目(2018-10038);北京市博士后工作经费资助项目(ZZ2019-73)
作者简介: 孙恩昌(1977—), 男, 副教授, 主要从事未来网络、信号处理与智能优化算法方面的研究, E-mail: ecsun@bjut.edu.cn

场景. 有限的通信资源与激增的通信需求和多维的网络资源与细致精准的资源分配,这 2 组互相对立统一的关系迫使人们对移动通信资源管理进行更为深入的研究与探索. 在此背景下,深度强化学习(deep reinforcement learning, DRL)通过随时与环境进行动态交互而拥有学习与决策能力,在通信资源管理中发挥越来越重要的作用.

在移动通信领域,DRL 主要具有如下优势:首先,DRL 具有模拟人脑的学习能力,能快速适应当前通信网络的异构性与动态性;其次,DRL 在动态网络场景下具有决策能力,能实现对多维通信网络资源的精准分配;最后,DRL 能直接对原始数据进行特征提取,实现端到端的学习,从而输出结果. 基于此,DRL 可以将复杂的移动通信资源管理问题抽象为具体的数学模型,从而设计出高性能的资源优化算法,使其有针对性地满足不同的用户需求成为可能. DRL 在移动通信的功率分配、用户调度和无线资源管理等领域发挥重要作用,同时也为蜂窝网、物联网(Internet of things, IoT)、无线接入网(radio access network, RAN)等不同网络技术发展带来新的契机.

目前,机器学习在移动通信领域的调查研究[1-2]主要侧重于对深度学习(deep learning, DL)和强化学习(reinforcement learning, RL)在移动与无线网络中应用的总结与讨论[3-7],而对于 DRL 与移动通信资源管理与优化的综述研究则缺乏较为深入的系统分析. 例如:文献[5]主要对蜂窝网和 IoT 中基于机器学习的资源管理方案进行研究;文献[6]则主要讨论了基于人工神经网络的极大似然估计在未来无线网络中的应用. 另外,针对 DRL 的探讨一般集中在游戏和机器人等应用[8],而对于 DRL 与移动通信相关的研究多集中在动态网络访问、网络接入、数据速率控制等问题[9-12],鲜有系统讨论 DRL 在不同通信场景中资源管理性能与优化方面的研究报道. 其中:文献[11]针对 DL、DRL 与联邦学习在无线通信中的无线传输、频谱管理、网络接入等应用进行综述;文献[12]主要从频谱、功率和网络资源 3 个角度对基于 DRL 的资源管理进行介绍,均未针对不同通信场景下各类基于 DRL 的资源管理方案进行详细的梳理、归类与总结.

本文针对 DRL 与移动通信的资源管理、分配与调度进行详细分析. 首先,介绍 DRL 的概念、基本原理与算法改进;然后,对 DRL 在网络切片、云计算、雾计算、移动边缘计算(mobile edge computing, MEC)、IoT 等场景下资源管理与分配问题的研究进展进行综述、归纳与总结,同时,对 DRL 结合网络功能虚拟化(network functions virtualization, NFV)、车辆雾计算(vehicle fog computing, VFC)、终端直通(device-to-device, D2D)等通信技术的资源管理方案进行分析对比,讨论存在的问题与挑战,并提出相关的解决思路;最后,对 DRL 在移动通信资源管理的主要研究方向进行展望.

1 DRL 概述

RL 作为机器学习的重要分支,通过将智能体与环境进行交互来学习最优策略. 然而,受维度爆炸的限制,RL 算法的扩展性较差,无法有效应对状态空间和动作空间较大且连续的情况. 基于此,DRL 充分利用 DL 从高维数据提取低维特征的能力,并将 RL 的决策能力与 DL 的感知能力进行融合,获得了更为优越的性能. 目前,DRL 广泛应用于无人驾驶[13]、机器人控制[14]、自然语言处理[15]等领域. 根据神经网络是否对值函数和策略函数进行逼近,DRL 算法分为 3 类[16]:基于值函数的 DRL 算法,如经典的深度 Q 网络(deep Q-network, DQN);基于策略梯度的 DRL 算法,如近端策略优化(proximal policy optimization, PPO)[17],解决了策略算法中步长难以确定的问题;基于值函数与策略梯度的 DRL 算法,如常用于解决移动通信问题的动作评论(actor-critic, AC)[18]和结合了 AC 与 DQN 的深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法,解决了连续动作空间问题.

作为 RL 与 DRL 算法的基础,马尔可夫决策过程(Markov decision process, MDP)[19]是实现随机性策略和奖励的数学模型,它将 RL 的交互过程以概率论的形式表示. 因此,随机环境下的决策问题可以建模为 MDP,从而构建 DRL 算法架构. 机器学习中 DL、RL 和 DRL 的理论关系与常用算法分类如图 1 所示. 下面在介绍 DL 和阐述 MDP 与 RL 基本原理的基础上,以 DQN 为例分析 DRL 的原理与特点,然后,从 DQN 与 AC 算法出发,探讨不同 DRL 的适用性与优缺点.

1.1 DL

DL 源于感知器和玻尔兹曼机技术[20],可分为 2 类:起源于感知器的 DL 根据期望输出训练网络,称为有监督式学习;起源于玻尔兹曼机的 DL 则根据特定样本数据训练网络,称为无监督式学习. DL 的基本思想是利用深度神经网络(deep neural

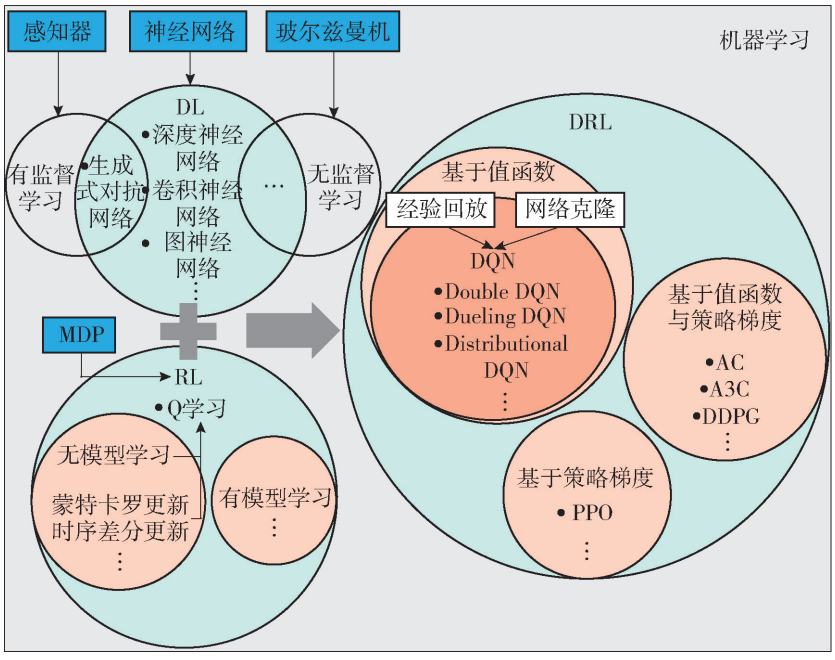

图 1　DL、RL 和 DRL 关系示意图

Fig. 1　Diagram of the relationship between DL, RL, and DRL

networks, DNN)结构拟合输入数据与目标数据的关系. DNN 又称前馈神经网络,由输入层、多个隐藏层与输出层组成. 1986 年,Rumelhart 等[21]合著的论文提出通过反向传播来训练 DNN,为 DL 的发展奠定了基础. 2006 年,随着计算机运算能力的提高,Hinton 等[22]进一步提出了利用受限玻尔兹曼机编码的深度置信网(deep belief networks, DBN). 同年,Hinton 等[23]提出预先训练多层神经网络来重建高维输入向量,可以将高维数据转换为低维数据,作为 DL 的开篇,较为系统地将神经网络引入无监督学习领域.

常用于 RL 的神经网络有 DNN、卷积神经网络(convolutional neural networks, CNN)[24]、图神经网络(graph neural networks, GNN)[25]和生成式对抗网络(generative adversarial networks, GAN)[26]等. 其中,GAN 受博弈论中零和博弈的启发,同时训练生成器神经网络和判别器神经网络 2 个模型,前者用于生成内容数据,后者用于判别前者生成的内容数据.

由于能够对高度抽象的数据进行表征学习,DL 适用于具有多个节点和动态链路的复杂无线网络,在移动通信的访问控制、网络安全、资源分配等领域中发挥重要作用[3]. DL 算法的学习效果依赖于样本数量,但在实际应用中,可用的数据集并非总是完美的,同时还存在样本数据不足或冗余的情况,这都对 DL 的训练效率产生影响[27]. 同时,DNN 的非线性特征令 DL 模型的可解释性较低[28],人们很难理解 DL 是如何通过提取特征进行学习并输出正确结果的,这不利于对神经网络的优化. 此外,由于利用有限的训练样本构建预测模型时存在不确定性[29],DL 缺乏一定的决策能力,在实际应用中具有局限性.

1.2　RL

RL 是机器学习领域的重要分支[30],其基本原理是通过实现系统状态与动作之间的映射,将智能体与环境进行交互,以选取合适的动作当作最优策略来最大化长期奖励,目前,已成功运用于机器人控制[31-32]和无人驾驶[33]等领域. RL 中智能体与环境的交互遵循 MDP. MDP 是针对随机序列决策的研究理论,具体步骤是智能体根据观察到的系统状态选择并执行一个可用动作,获得与所选动作和当前状态相关的奖励,然后转移到下一状态继续照此循环. 其中,奖励的设计通常取决于对应的 RL 问题. 值得注意的是,MDP 假定系统可以直接观测到当前的状态,但是在实际操作中,智能体在每个时间点只能得到一个包含所有可能状态的概率分布,并不能完全观测到系统的状态. 因此,运用部分可观测马尔可夫决策过程(partially observable Markov decision process, POMDP)作为通用化的 MDP[34],能有效模拟现实世界的连续过程.

假设在环境 E 中,智能体感知到的环境状态集合为状态空间 S,智能体能够采取的动作集合为动作空间 A. 利用 MDP 可将 RL 任务描述为[35]:若在状态$s \in S$时采取动作 $a \in A$,环境则按照状态转移函数 P 转移到下一状态 s',同时根据潜在的奖励函数 R 反馈给智能体一个奖励. 然后,智能体通过在环境中不断尝试,得到一个最大化长期奖励的最优策略 π,即可获得在状态 s 下要执行的动作 $a=\pi(x)$,其中 x 为系统的当前状态.

根据智能体在学习之前能否对环境建模,RL 一般分为有模型学习和无模型学习. 有模型学习指对应任务的 MDP 和 $E=\langle S,A,P,R\rangle$ 均已知,即智能体在学习之前已经能模拟出与环境相同或相近的状况. 此时,对于任意状态的 s、s'和 a,在状态 s 下执行动作 a 并转移到状态 s'的概率 $P^a_{s\to s'}$及其带来的奖励 $R^a_{s\to s'}$都是已知的. 然而,在对 RL 任务进行实际操作时,环境中的转移概率和奖励函数通常难以预知,智能体无法在完成学习之前对环境进行建模,因此,这种模式属于无模型学习.

RL 算法的学习目标是找到使长期累积奖励最大化的策略 π,而长期累积奖励通常有 2 种表达方式:一种由状态值函数 $V^\pi(x)$ 表示,即从状态 s 出发,使用策略 π 获得的累积奖励;另一种由状态动作值函数 $Q^\pi(s,a)$ 表示,即在状态 s 下,执行动作 a 后再使用策略 π 获得的累积奖励. 在有模型学习中,状态动作值函数 Q 利用策略迭代算法估计状态值函数 V,再将 V 转化为 Q 来获得最终策略. 但在无模型学习中,从 V 转化到 Q 的过程非常困难,因此,直接对状态动作值函数 Q 进行估计. 值函数的更新主要分为蒙特卡罗(Monte Carlo, MC)和时序差分 2 种方法. 其中,MC 指对多次得到的奖励取平均值,作为期望累积奖励的近似.

Q 学习是无模型和时序差分更新算法[36],属于 RL 中基于值函数的算法. Q 学习将状态与动作组成一张 Q 表用于存储 Q 值,并根据 Q 表更新其 Q 值来实现奖励最大化,学习并获取最优策略. 由于 Q 表的存储能力有限,Q 学习主要用于处理规模较小的 RL 问题.

1.3 DRL

如前所述,RL 通过求解 MDP 问题能使智能体根据环境给出的奖惩进行学习以获得最优策略,有效解决序列决策问题,但 RL 存在状态空间与动作空间爆炸的缺陷,更适合解决规模较小的离散空间问题. 同时,由于无法处理从未出现过的状态,RL 并不具备预测能力. 为此,DRL 融合了 DL 的感知预测能力与 RL 的决策能力,实现了从感知到动作的端到端学习,即将输入的文本、图像、音视频等原始数据,通过 DRL 架构中的 DNN 进行处理,无须人工干预,直接输出结果.

DRL 的出现提高了 RL 技术的实用性,不仅有效解决了现实场景中的复杂问题,也为移动通信的资源管理方案提供了新思路. 然而,在 DRL 框架下,探索-利用困境[37]仍未完全得到解决. 一方面,需要智能体探索新的策略以获得更高的回报;另一方面,智能体应该充分利用现有的知识,避免对策略与状态的不必要探索. 虽然探索增加了智能体对动态环境的适应性与灵活性,但同时也降低了智能体的学习能力. 因此,合理把握探索与利用之间的平衡,仍是 DRL 研究的重要课题. 此外,DRL 算法仍存在过拟合、陷入局部最优、收敛性不佳等问题,下面将以 DQN 和 AC 算法为例对 DRL 的改进方法进行探讨.

1) DQN

DQN 算法是 Mnih 等[38]在 DRL 领域完成的开创性工作,最早用于处理基于视觉感知的控制任务. DQN 的基本原理是通过卷积神经网络对状态动作值函数 Q 进行逼近,同时利用 Q 学习方法更新 Q 值. 在使用非线性函数逼近的过程中,策略对 Q 值的变化非常敏感,这使得 RL 算法的稳定性较差,采用经验回放[39]和网络克隆的方法可以解决. 经验回放是指智能体将以往的经验数据存储到数据集,再从中选择小批量数据来更新 Q 值,有效克服经验数据的非平稳分布问题[38],从而减轻连续样本的相关性,提高数据利用率;网络克隆是指智能体根据目标网络选择动作,并每隔某个时间段将当前网络克隆到目标网络,减轻因为 Q 值频繁变化而引起的震荡,提高算法稳定性[40].

DQN 网络的结构如图 2 所示,已知当前网络的权重 θ 和目标网络的权重 $\hat{\theta}=\theta$,智能体在 t 时刻观察到的状态为 s_t,以一定概率或在满足 $a_t=\underset{a}{\operatorname{argmax}}\, Q(s_t,a;\theta)$ 的条件下选择并执行动作 a_t,得到相应的奖励 $R(s_t,a_t)$,之后转移到下一状态 $s_{t+1}=s'_t$,继续循环. 智能体将 t 时刻的经验数据 $e_t=\langle s_t,a_t,s'_t,R(s_t,a_t)\rangle$ 存入经验存储单元,并从中对小批量的数据进行采样,同时更新 Q 值,即

$$Q^+(s_t,a_t)=R(s_t,a_t)+\gamma \max_{a'} Q^+(s'_t,a') \quad (1)$$

式中 γ 是反映当前奖励对未来奖励重要性的权重.

若此刻时间结束，则 $Q^+(s_t,a_t)=R(s_t,a_t)$. 在以上过程中，智能体使用梯度更新方法对 θ 进行更新，同时，通过每隔某个时间段使 $\hat{\theta}=\theta$，将当前网络克隆到目标网络.

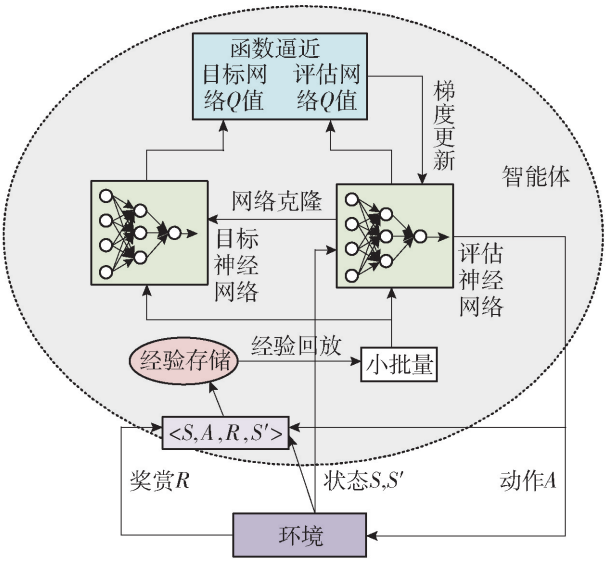

图 2　DQN 的一般结构[41]

Fig. 2　General structure of DQN[41]

2）DQN 及其改进方法

DQN 能有效解决动作离散的 RL 问题，拥有较好的泛化能力和发展潜力，但由于直接选取当前 Q 值最大的动作来更新目标函数，DQN 算法存在对 Q 值的过高估计. 为弥补这一缺陷，同时进一步提高 DQN 的算法效率与性能，出现了优化训练算法、设计网络架构、引入新机制等改进的方法.

典型的是双深度 Q 网络（double deep Q-network，Double-DQN），主要对 DQN 的训练算法进行改进，用不同的值函数实现对动作的选择和评估，在一定程度上降低了随机环境中 Q 学习算法对动作的过高估计[42]，但它所需运算量过多且仍存在估计偏差问题. 针对 DRL 累积奖励过高估计的问题，在训练算法中引入惩罚变量增加随机性，能有效降低 Q 值的估计[43]，这将有助于提高 DQN 与 Double-DQN 的采样效率和算法性能. 同时，竞争深度 Q 网络（dueling deep Q-network，Dueling-DQN）对 DQN 的网络架构进行优化[44]，与 DQN 中单个的深层网络输出层不同，它通过合并 2 个单独的流生成输出，其中一个流计算状态值函数，另一个流计算状态作用值函数，显著提高了算法效率. 但由于学习过程中产生过多簇且簇的数目未知或难以预先估计，Dueling-DQN 应用泛化程度较低.

分布式深度 Q 网络[45]（distributional deep Q-network，Distributional-DQN）通过开发大量的计算资源，构建了随着计算复杂度和存储复杂度增加而提升性能的可扩展结构，是颇有前景的 DRL 算法. Distributional-DQN 运用贝尔曼方程的分布模拟来描述 Q 值的分布，提高了算法的学习稳定性，从而可更为有效地解决动态系统或动态环境的策略问题. 此外，为了进一步提高 DQN 算法在多智能体环境下的收敛性能，可以引入新的学习机制来解决多目标决策问题，并构建相应的架构和网络参数优化机制，防止算法模型陷入局部最优.

3）AC 算法及其改进方法

除 DQN 外，AC 算法是处理通信资源管理问题的另一常用方法，通常由 Q 学习与策略梯度方法结合而成，其单步更新的特点能显著提升学习效率[46]. AC 算法架构由动作（actor）网络与评价（critic）网络两部分组成，其基本原理是将策略梯度中的智能体替换为 actor，利用奖励直接决定行为选择概率，同时，在此基础上创建 critic 网络来计算 Q 值. 然而，单个智能体采集的样本数据不稳定，导致在线学习的更新序列呈现高相关性. 为避免这一缺陷影响学习算法的收敛，DQN 采取了经验回放机制，但这种基于经验回放的 DRL 在训练过程中需要较大的存储空间来存放历史样本，并且它仅适合解决离策略学习问题，而离策略学习只能利用旧策略生成的数据进行参数更新.

异步优势动作评价（asynchronous advantage actor-critic，A3C）算法采用另一种方法替代了经验回放机制. A3C 构造多个训练环境同时进行采样，并利用多个并行的 AC 进行学习来获取分布更为均匀的样本，提升了收敛性能和训练速度[47]. 虽然这种方法同样能够实现稳定的在线学习，但并不代表经验回放机制无作用. 在异构 RL 架构中引入回放机制来对样本数据重复利用，能有效减少学习所需的经验数量，从而提升数据效率，节约算法的运行成本. 同时，文献[48]证明了将 AC 架构与优先经验回放技术相结合的方法是可行的. 因此，在提升 A3C 算法的训练速度和学习效果上，合理应用经验回放机制是值得考虑的办法. 此外，在优化算法收敛性能方面，借鉴 Dueling-DQN 的竞争架构优化 A3C 的神经网络结构也是值得尝试的改进方法.

如表 1 所示，针对上文提到的各类 DRL 算法，本文归纳了其特点与应用. 可见，Distributional-DQN 与 A3C 算法在收敛性与可靠性方面具有优势，能有

效解决多任务协作问题,进而在移动通信的复杂资源管理中发挥重要作用. 除上述对算法架构的改进之外,文献[49]通过使2个智能体进行信息交互实现了对动作的共同决策. 因此,模拟人类之间的协作学习模式,实现智能体之间的信息共享,也是提升DRL算法性能的有效方案之一.

表1 常用DRL算法所解决的问题及意义和优缺点

Table 1 Solved problems and significance, advantages and disadvantages of common DRL algorithms

算法	解决问题及意义	优势	缺陷
DQN	利用经验回放和网络克隆方法进行学习并稳定训练过程	能够处理动作离散问题;算法稳定性好;泛化能力强	算法效率与收敛性能较差;存在过估计问题
PPO	解决策略算法中步长难以确定的问题	实现简单;样本复杂度较低;调参难度较低	算法无法并行运行
Double-DQN	解决值函数的过高估计问题	提升了学习效果	计算量大,算法效率较低;仍存在估计偏差问题
Dueling-DQN	提出新的神经网络结构	提高了算法运行速度和效率	算法复杂度高;应用泛化程度较低
Distributional-DQN	改进神经网络结构以开发计算资源,扩展DQN算法	具有扩展结构;算法性能与稳定性高;节约算法运行时间	算法复杂度较高
AC	结合值函数与策略梯度方法	提升了学习效率;能够处理连续或高维动作空间问题	样本数据不稳定;更新序列相关性较高;收敛较困难
DDPG	结合AC与DQN,解决连续动作空间问题	减少了数据间的相关性;增加了算法的稳定性	算法收敛所需的运算时间长
A3C	实现多个智能体并行学习	能处理高维连续动作空间问题;收敛性佳;训练速度快	算法的数据效率较低

2 DRL与网络切片

网络切片通过使运营商以不同的网络性能满足不同的用户需求,成功适应移动网络生态系统的转换[50-53]. 网络切片是对通信资源进行优化与管理的技术,本节在介绍DRL与网络切片研究进展的基础上,对DRL应用于RAN、核心网、车载通信等场景的切片资源分配进行研究,分析其特点与技术难点,进而探讨可能的发展方向.

2.1 DRL与RAN、核心网的切片资源分配

网络切片需要将资源管理与每个切片的用户活动保持一致,才能满足移动通信对网络性能和成本效益的要求[51]. 网络切片的资源管理可分为RAN和核心网两部分[41]. 由于频谱资源有限,RAN的资源管理主要是通过将无线资源分配到切片来获得最优的频谱效率,实现高速率和低时延的网络接入;核心网则大多利用光传输技术,优化工作通常为设计通用或专用的虚拟化网络功能(virtual network functions, VNFs),从而高效地转发来自特定切片的数据包. 可见,网络切片的资源管理方案需要考虑多个变量作为优化目标. 由于能直接从高维感知输入中学习成功策略,DRL在处理这种大型状态与动作空间问题时具备优势.

目前,DRL在RAN、核心网的切片资源中取得较大进展. 在基于值函数的DRL算法方面,除文献[54]利用DQN实现智能电网的RAN切片分配外,文献[55]使用分布式学习方法改进DQN的经验回放算法,由一个智能体代表一个切片,在满足每个切片需求的前提下,为其分配最少的无线资源块,以达到节约RAN资源的目的. 同时,为了在动态的移动通信场景下最大限度地利用网络资源,文献[56]将动作分支结构[57]引入竞争Double-DQN算法框架,有效解决大型状态空间和多维离散动作空间的问题,实现了智能的核心网切片重新配置. 在基于值函数与策略梯度的DRL算法方面,文献[58]将不同服务需求和资源看作状态和动作,同时,考虑用户移动性,增加了感知环境的难度,在优势动作评价(advantage actor-critic, A2C)算法中引入长短期记忆(long short-term memory, LSTM)网络架构,实现切片间的智能资源

管理，提高了频谱资源利用率. 未来可利用 DRL 算法制定联合资源管理方案，进一步实现切片间的资源分配和切片内的用户调度.

移动通信网络结构的多样性与复杂性给网络切片资源与业务服务的分配带来挑战. 因此，需要根据用户请求与服务需求进行网络切片资源分配的研究[10, 56, 59]. 由于来自不同网络切片用户的各类服务请求随机且不可预测，请求分配的难点在于适应其不确定性以及处理用户产生的大量动作数据. 文献[10]基于 DQN 提出智能的切片分配方案，基本实现了根据用户请求提供个性化服务的功能，但此模型仅针对单个网络实体，并不适合解决大型且异构网络中的复杂资源管理问题，从而无法应对多样化的通信需求. 由此可见，集中式 DRL 算法在当前分布式通信网络场景中具有一定的局限性. 与之不同，文献[59]提出基于 GAN 的分布式深度 Q 网络（generative adversarial network-powered distributional deep Q network，GAN-DDQN）算法，实现了需求感知的网络切片资源分配. 该方案将传统 DRL 改进为分布式架构，缓解了随机噪声对 DRL 算法动作值估计的负面影响，从而有效解决如文献[41]中DQN 方法的动作值估计偏差问题. 同时，进一步提出竞争 GAN-DDQN，利用 Dueling-DQN 架构和离散规范化优势函数算法[60]导出特殊解，将状态值分布和动作优势函数从动作值分布中分离出来，更为高效地学习动作值分布. 由于能在切片服务需求不确定的情况下寻找最优的资源分配策略，竞争 GAN-DDQN 方法在时常变换业务需求的大型通信网络场景中具有发展潜力.

总的来说，Distributional-DRL 算法将训练任务分解为拥有不同作用的子任务，分别用于计算能力中等或有限的单个设备，从而实现对每个切片的动态高效资源分配. 因此，将 Distributional-DRL 架构运用于网络切片资源分配方案值得进一步探索.

2.2　DRL 与其他网络场景的切片资源分配

如何改进 DRL 算法以便对网络切片资源进行智能高效的管理与分配，从而更好地适应动态且异构的网络环境，是 DRL 与网络切片研究的关键任务. 因此，除上述研究进展外，本文将针对其他不同场景下基于 DRL 的切片资源管理方案继续探讨.

除上述 RAN 与核心网场景外，DRL 与网络切片结合的资源管理方案还可应用于车载通信. 针对多租户网络中的车对车通信，文献[61]基于 D2D 技术提出了一种可扩展的切片间资源分配与 D2D 资源聚合方案，使用 DQN 进行虚拟资源分配，然后，将自定义物理资源分配建模为凸优化问题，并利用交替方向乘子法（alternating direction method of multipliers，ADMM）进行求解，实现了高带宽与低时延的车对车通信. 此方案的实现证明了 DRL 算法与 ADMM 等凸优化方法的适配性，为基于 DRL 的资源管理方案提供了新的改进思路.

将内容缓存到离最终用户更近的地方能提高通信的数据传输速率和服务质量（quality of service，QoS），但在供应商共享公共基础设施的情况下，管理有限的缓存资源则更具挑战性. 根据前文提到的各类方法可以看出，DRL 技术在这种复杂场景中具有独特优势. 以文献[62]为例，作者首先利用分布式 ADMM 方法[63]求解缓存内容放置问题，然后利用 DQN 算法实现虚拟缓存切片和自定义的内容放置. 目前，较为成熟的方案大多假设所有虚拟切片都具有相同的目标函数[64]，而此方案支持自定义虚拟切片，每个运营商都可以根据自己独特的 QoS 目标优化自身的缓存布置策略. 同时，不同切片的虚拟缓存资源是独立的，一个切片需求突然激增不会影响其他切片的性能. 可见，针对运营商的差异化服务，基于 DRL 的资源管理方案能合理平衡 QoS 和资源利用率之间的关系，并在需要时引入权重进行调整，具有较好的灵活性与可靠性. 同时，在算法框架中合理利用凸优化方法，能够有效提高 DRL 算法的自主性和环境适应性，为 DRL 方法的改进提供了新的发展空间.

本文总结了上述方案的应用场景、解决的问题、所用算法与方法以及性能表现，如表 2 所示. DRL 凭借其强大的策略选择能力，为网络切片提供了动态的资源管理方案，在网络切片的资源利用率、通信吞吐量、用户个性化 QoS 性能等方面有所提升. 总的来看，DRL 在网络切片的应用已取得阶段性进展，有效解决了需求感知切片资源管理、智能切片资源分配、切片用户请求分配、虚拟缓存切片分配等问题，适用于 RAN、核心网、车载通信等场景. 同时，利用凸优化等数学方法与 DRL 相互配合，从而提升网络切片资源管理方案的综合性能，仍值得进一步研究. 此外，如何利用分布式 DRL 架构在保证收敛性和节约系统能耗的前提下解决更为复杂的网络切片资源管理问题，同样值得深入探索.

表 2 基于 DRL 的网络切片资源管理方案比较

Table 2 Comparison of network slice resource management schemes based on DRL

应用场景	解决问题	所用算法与方法	性能表现与进步	文献
RAN	智能电网切片分配	DQN	降低运行成本;提高系统效益	[54]
	网络切片资源智能分配	分布式 DQN	在不受切片数量影响下,提高切片需求的满意度	[55]
	需求感知切片资源管理	A2C + LSTM	适应用户移动性;提高资源利用率与系统实用性	[58]
	需求感知切片资源分配	竞争 GAN-DDQN	提升算法性能表现;提高算法可扩展性	[59]
核心网络	智能网络切片重新配置	竞争 Double-DQN	减少长期资源损耗;提高资源效率;提高算法动作空间利用率	[56]
5G 网络	网络切片优化	DQN	不考虑系统状态的情况下,提高服务移动用户请求的效率	[10]
车载通信	虚拟切片资源动态管理	DQN + ADMM	提高多个切片的资源利用率;增加系统吞吐量;提升 QoS	[61]
移动边缘网	虚拟缓存切片与自定义的内容放置	DQN + 分布式 ADMM	提高资源利用率;提升环境适配性;降低算法复杂度	[62]

3 DRL 与云、雾、移动边缘计算

本质上,云计算是提供快速安全云服务与数据存储的云网络,其中的每个用户都能随时获取“云”上庞大的计算、存储、应用程序等资源. 然而,随着依赖云计算的通信设备越来越多,从云端到终端的数据传输实时性与可靠性显著降低,在终端和数据中心之间添加网络边缘层,将并不需要放到“云”的数据在这一层直接处理和存储是有效的解决办法. 因此,雾计算和边缘计算应运而生,已经并将继续在移动网络中发挥作用. 但同样地,无论是雾计算还是边缘计算都仍离不开云端. 在云环境中,中央处理器、内存、网络、存储等可用资源都需要根据需求和使用情况进行有效分配,从而减少云数据中心负载和提高资源利用率,因此,如何合理分配与管理资源成为关键任务[65]. 本节在总结 DRL 与云、雾、移动边缘计算资源管理研究进展的基础上,探讨基于 DRL 的资源管理方案,并分析其优势与缺陷.

3.1 DRL 与云计算

RAN 利用无线通信技术在交换节点到用户终端之间添加无线接入,弥补了有线通信模式受传输带宽影响的缺陷,但传统的 RAN 已无法适应通信业务需求的多样化,引入云、雾、边缘计算技术进行改进是有效的解决办法. 例如,云计算与 RAN 结合而成的云无线接入网(cloud radio access network, C-RAN)将计算资源放置在中央无线网络云中,对系统数据进行集中管理,能节约运营成本和系统能耗. 基于 DRL 的资源管理框架在 C-RAN 的功率分配和云资源管理中取得良好性能[66-67],证明了 DRL 算法在提高通信能效和网络动态性方面的优势.

云计算大多通过对多台服务器实体进行虚拟化并构成一个资源池,实现共同计算与资源共享. NFV 将 C-RAN 的部分网络功能进行虚拟化,显著提高了网络资源利用率和通信部署灵活性. 在 NFV 系统中,服务器可直接布置相应 VNFs 执行指令,实现丰富的网络服务[68]. 然而,由于每个虚拟设备都集中部署在同一数据中心,用户距离此中心越远,网络业务数据传输时经过的转发设备就越多,直接导致通信时延增加. 因此,合理编排 VNFs 来实现虚拟资源管理是一大挑战.

在 DRL 框架中加入辅助手段优化算法的学习与训练过程,是改进 DRL 处理虚拟资源管理问题的常用方法. 针对将 IoT 与云计算集成的通信场景,文献[69]利用自然梯度下降法对 DRL 算法的策略进行优化,通过选取新的替代损失函数来减小新旧策略的差异,并在每个时间段都采用适当的学习步骤来提升网络的吞吐量性能,实现了基于 VNFs 的自适应 IoT 资源调度. 文献[70]对 DRL 的训练过程进行优化,通过运用基于优化模型的启发式求解方法指导智能体学习最优策略,提高算法的收敛速度与效率,实现最优 VNFs 编排以最大化网络效用. 此外,为提高可用硬件的利用率,文献[71]利用 DDPG 算法提出策略与动作值函数的联合优化方案,将随机资源优化问题建模为参数化动作的 MDP,有效解决基于值函数的 DRL 方法[72]由于状态与动作空间过大而难以获得最优解的问题.

合理部署 NFV 以实现资源管理的另一常用方案是 DRL 借助区块链、博弈论等关键技术与数学方法弥补自身不足. 例如，文献[73]在通过引入区块链技术实现可信任资源共享的基础上，将服务功能链的调度优化问题建模为 MDP，并利用 A3C 算法对动态分层服务功能链进行编排. 此方案通过用户 QoS 满意度确定服务功能链的状态，智能体根据该状态选择路由和决定 VNFs 布局，从而在满足 QoS 的前提下节约了网络成本. A3C 通过并行训练多个智能体并整合所有经验数据，解决交互过程中使用过多资源而产生巨大计算量的问题，体现了多智能体 DRL 算法的优势. 文献[74]提出非合作混合策略博弈方法，用户根据收入和 QoS 产生的激励来竞争 VNFs 服务链提供的服务. 然而，此方案对算法的性能与运算效率要求较高且需要手动配置最佳 VNFs 资源的大小和位置. 针对此类缺陷，DRL 凭借其自主学习能力和对环境的高适应性可以解决. 因此，后续研究可尝试将 DRL 与博弈论融合，例如，将文献[73]中可信智能的资源共享模型引入文献[74]的激励式 VNFs 供应方案中，获取二者的性能优势，设计出功能更加完备的 NFV 资源管理方案.

总体来看，DRL 与云计算资源管理的最新研究成果较为丰富，并且大多围绕着计算、通信、存储、服务等网络资源的高效利用、资源管理策略优化、安全可信的资源共享等主题. 除上述研究进展外，基于 DRL 的云计算资源管理方案还应用于边缘云的资源调度[75]、边缘微云的多资源公平分配[76]、大数据任务的智能调度[77]等场景. DRL 技术的蓬勃发展与广泛运用为云计算及其延伸领域的资源管理提供了新的思路，如何在云网络中深度融合 DRL，实现高实时性、低成本、保障隐私安全的资源管理方案，是值得思考的问题.

3.2　DRL 与雾计算

云计算将所有数据上传到云端进行保存，在保证灵活性的同时，也容易产生网络延迟或中断. 对此，人们提出了更贴近地面用户的雾计算[78-79]. 雾计算将不需要放在云中的数据与应用程序集中在网络边缘层，减轻了云端的压力. 雾计算是基于“云”概念的延伸，其基本原理与云计算相似，是分布式的云计算.

雾计算能在不依赖云服务器的情况下提高车联网的计算能力，这种技术称为 VFC，是一种利用多个车载设备或接近用户的边缘设备进行通信和计算的协作架构. 因此，如何在车联网中实现基于 VFC 的资源管理与分配来激励车辆共享其空闲的通信与网络资源，是一项关键任务. 目前，利用 DRL 实现 VFC 资源管理的研究主要集中在计算卸载决策[80-82]. 例如，文献[80]利用基于最大熵框架的软动作评论(soft actor-critic, SAC)算法，通过将策略的熵引入奖励中构建基于非策略 DRL 算法的计算任务卸载方案. 该方案设计了动态定价机制来激励车辆共享其空闲的计算资源，有效解决了任务优先级、车辆服务可用性与计算资源共享的联合资源管理问题.

雾计算应用于 RAN 中形成雾无线接入网(fog radio access network, F-RAN)，能弥补 C-RAN 在实时性方面的缺陷，提高频谱效率与能量利用率[83]. DRL 算法通过学习用户请求的不确定性和动态性提高缓存性能，有助于 RAN 的内容访问和功能检索[84-85]，同时也在 F-RAN 的缓存中发挥作用[86-88]. 解决 F-RAN 中缓存策略优化问题的方法之一是设计新的雾接入点(fog access points, F-APs)，从而充分利用本地的缓存功能. 文献[86]利用 DRL 智能地将 F-APs 中有限的缓存空间分配给不同的编码文件，在提高缓存资源利用率的同时满足用户需求. 但由于 F-APs 的缓存空间过小且编码缓存策略依赖于用户的实际传输性能，此方案更适合静态环境. 与之不同，文献[87]基于 DRL 算法提出动态的缓存资源管理方案，利用迁移学习加速 DRL 的训练过程，实现 F-RAN 的通信模式选择与边缘缓存服务器状态控制，减少了网络系统的长期功耗. 同时，文献[88]针对 F-RAN 的延迟优化问题，利用 DRL 实现缓存内容布置与功率分配的联合智能决策. 受此启发，文献[87]可进一步优化 DRL 算法，从而实现功率控制、子信道分配与前向传输资源分配的联合通信资源管理方案，提高通信系统的实用性.

此外，DRL 还用于 F-RAN 的计算卸载[89-90]，例如，文献[89]通过将模式选择、计算卸载决策、计算资源分配和功率分配的联合资源管理这一非凸问题转化为 MDP，从而利用 DQN 算法解决，实现了在计算资源和前端传输容量的约束下，F-RAN 系统的延迟最小化. 综合本小节内容可知，为适应动态且异构的网络环境，针对移动通信资源管理的研究需要并行分析多个问题，所建立的问题模型大多是非凸且复杂的，因此，如何利用 DRL 的优势来简化并解决这些复杂问题，进而实现通信资源的联合管理，是值得探索的研究方向.

3.3 DRL 与移动边缘计算

MEC 是将云计算从核心网络内部迁移到移动接入网络边缘的平台,它以就近原则调用计算资源来满足终端的业务需求,主要用于减少传输时延与网络拥塞[91]. MEC 推进了雾计算中“局域网处理能力”的理念,可以直接在移动网络内的设备上处理数据,因而拥有更快的网络服务响应,在移动边缘服务器部署、协作缓存、多层干扰消除等应用中取得了较为优秀的成绩[92-95].

DRL 在 MEC 的计算卸载中具有较强的实用性与可操作性[96-99]. 文献[96]与文献[97]均采用集中式 DRL 架构,例如文献[96]提出节能的计算卸载方法,将卸载请求分为延迟容忍和非延迟容忍 2 层,利用 DRL 选择适当的层来执行计算卸载任务,在保证延迟性能的同时提高能效. 然而,在处理通信设备较多的问题时,受状态空间和动作空间的限制,集中式 DRL 算法在复杂度和运算时间方面的性能较差. 因此,文献[98]结合拍卖机制提出了半分布式的 DRL 联合资源管理方案,实现基于 MEC 的计算卸载与多用户调度. 具体地,基站根据 IoT 设备进行分布式计算后提交的报价,利用基于值函数的 DRL 算法集中做出资源管理决策,在提高系统性能的同时减少了通信开销. 这种将基站与通信设备进行智能协作的方式符合未来无线网络的设计原则,有望解决 Distributional-DRL 方法带来的高能耗问题.

此外,文献[100]将基于区块链的信任管理方法与基于 DRL 算法的智能合约机制相结合,提出安全智能的车载云网络计算卸载调度方案,在可扩展性、鲁棒性和环境适应性方面具备独特优势. 该方案将资源请求定义为智能体,在车载云网络的计算卸载中有多个资源请求,因此,DRL 具有多个智能体. 其中,该多智能体 DRL 算法具有集中训练与分布执行的功能,可以不断更新智能合约使其适应动态变化的车载场景. 目前,智能合约与人工智能算法的深度集成实现了具有分布式协作框架的学习市场,在公平性、透明性、安全性、去中心化、通用性等方面取得创新性成果[101]. 因此,将 DRL 算法与智能合约进行协作,实现安全、可靠、智能的资源管理与调度,是非常有前景的研究方向.

DRL 还应用于 MEC 的其他资源管理方案,文献[102]基于多任务 DRL 算法提出协作移动边缘计算(collaborative mobile edge computing, CoMEC)的智能资源分配方法,通过拆分 DNN 的最后一层来构造动作维度更高的子神经网络,自动学习网络环境并生成资源分配策略,在平均时延和能耗方面优于贪心搜索方法[103]和 DQN 方法[104]. 但受限于传统 DQN 的算法架构,该方案在处理大型动作空间问题时存在局限,未来可尝试使用 Dueling-DQN 与 Distributional-DQN 等方法改进神经网络结构,进一步提升系统性能.

如表 3 所示,本文总结了上述方案的应用场景、解决的问题、所用算法与方法以及性能表现与进步. 可以看出,DRL 通过与博弈论、凸优化、区块链等方法与技术进行融合,实现了延迟小、成本低和安全可靠的云计算资源管理方案. 同时,基于 DRL 的雾计算资源管理方案在资源利用率、系统成本和算法训练速度等方面取得创新性进展,基于 DRL 的 MEC 资源管理方案在延迟、功耗和算法鲁棒性方面表现优异. 这些成果均体现了 DRL 的可扩展性、稳定性与灵活性. 综合来看,DRL 在云计算、雾计算和 MEC 中得到广泛应用,适用于解决通信、计算与缓存资源的联合管理问题. 同时,将 DRL 与博弈论和区块链进行协作与配合,实现功能更为完备的 MEC 网络资源管理方案,可能成为未来网络的重要发展方向. 此外,如何在动态环境下调整训练结构和网络参数来优化 DRL 的网络架构与训练过程,进而提高算法在收敛性、稳定性与复杂度方面的性能,是解决高维且动态资源管理问题的关键任务,亟待进一步研究.

4 DRL 与其他网络技术

除上述通信与网络场景外,基于 DRL 的资源管理方法还应用于认知物联网[105]、超密集网络[106]、移动社交网络[107]等复杂移动网络. 同时,DRL 在 D2D 通信的功率分配[108-111]中也发挥重要作用,例如,文献[108-109]利用集中式 DRL 实现功率分配策略优化,具有优于其他常规方法的性能. 然而,集中式 DRL 方法虽然存在简单直观和易于实现等优点,但实际通信中的功率分配却受信道状态与同频干扰等多种因素影响,其单一的算法架构导致运算时间与运算量骤增. 因此,Distributional-DRL 算法以及多智能体 DRL 成为新的研究热点[110-111]. 例如,文献[110]利用 Double-DQN 算法有效解决了信道选择与功率控制的联合优化问题,D2D 对只需要根据局部信息和已有的非局部信息即可自主地优化信道选择与发射功率策略,体现了 Distributional-DRL 算法的可探索性和高适用性.

表 3　基于 DRL 的云、雾、移动边缘计算资源管理方案比较

Table 3　Comparison of cloud, fog and mobile edge computing resource management schemes based on DRL

应用场景		解决问题	所用算法与方法	性能表现与进步	文献
云计算	C-RAN	功率分配	Double-DQN	节省能量；减少功耗	[66]
		云资源高效分配	DNN + DQN + 凸优化	减少功耗；满足用户需求；适应高动态场景	[67]
	NFV	VNFs 智能配置	DRL + 自然梯度下降法	提升 QoS；减少网络拥塞，增加吞吐量	[69]
	IoT	可信智能的服务功能链编排	A3C + 区块链	节约成本；提升 QoS；减少运行时间	[73]
		智能大数据任务调度	DRL + LSTM	提高大数据分析的性能；节约内存使用成本	[77]
		VNFs 编排与流量调度	DDPG	提升网络效用；提高算法收敛速度	[70]
		VNFs 管理与编排	DDPG	节约成本；减小延迟；提升 QoS	[71]
	RAN	多租户跨切片资源编排	随机博弈 + Double-DQN	增加供应商的长期收益；提高算法收敛性	[72]
	移动网络	边缘云资源调度	协作 DNN + DRL	减小延迟；节省能量；增加吞吐量；提高资源利用率	[75]
	边缘微云	多资源公平分配	竞争 Double-DQN	使资源利用更均衡；提升 QoS；提高算法收敛性	[76]
雾计算	VFC	优先级感知计算任务卸载	SAC	提高任务完成率；减小延迟；提高算法鲁棒性与样本利用率	[80]
		部分计算卸载	A3C	节约成本；减少算法运行时间；降低算法复杂度	[81]
		节能计算卸载	最大流算法 + Double-DQN	实现路边单元间的负载均衡；降低能耗；减小延迟	[82]
		缓存资源分配	DQN	满足内容缓存多样性；提升 QoS；提高算法收敛性	[86]
		模式选择与资源管理	DQN + 迁移学习	减少系统长期功耗；提高算法学习速率；加快训练过程	[87]
	F-RAN	自主缓存与功率分配	DQN	提高算法收敛性能；减小系统延迟；增加吞吐量	[88]
		计算卸载与资源分配	DQN	减小系统延迟；增加吞吐量；提升 QoS；减少功耗	[89]
		计算卸载策略优化	Dueling-DQN	提升用户终端效用；减小延迟；降低算法复杂度	[90]
移动边缘计算	IoT	计算卸载	DQN	减少系统功耗与延迟；提高 QoS；获得更多奖励	[96]
		计算卸载与多用户调度	拍卖机制 + DRL	优化延迟和功耗的平均加权和；解决算法维度爆炸问题	[98]
		多任务计算卸载	Distribution-DRL	减少能耗；减小延迟；降低算法复杂度	[99]
	车辆边缘计算网络	计算卸载与资源分配	DQN	提升网络长期效用；减小延迟	[97]
	车载云网络	计算卸载	AC + 区块链智能合约	提高环境适应性；增加抵抗恶意攻击的能力；提高算法收敛性与鲁棒性	[100]
	CoMEC	边缘服务器选择与带宽分配	DQN + MC 树搜索	降低服务延迟；减少功耗；提升算法解决复杂动作空间问题的能力	[102]
	5G 网络	多用户计算卸载与资源分配	DQN	节约系统成本与开销；减少能耗；减小延迟	[104]

5 问题与展望

通过对 DRL 在移动通信资源管理中存在的问题与挑战及相应解决思路进行分析讨论,本文进一步展望了 DRL 与移动通信资源管理的未来研究方向.

5.1 问题与挑战

如前所述,虽然基于 DRL 的移动通信资源管理方案取得良好性能,但还有如下挑战与不足.

1) 针对动态的网络结构,如何合理设计 DRL 的状态与动作空间. 目前,常用的方法是利用分布式架构,将大规模的复杂问题拆分为许多小规模问题,同时,令多个智能体之间共享其状态空间[112]. 然而,Distributional-DRL 算法存在局部最优问题. Distributional-DRL 算法的基本原理是将学习任务分解为相对简单的子任务,从问题的局部展开求解,以减少算法复杂度与计算量,进行高效的资源分配. 但局部求解方法导致算法时常收敛到局部而非全局最优,直接影响了 Distributional-DRL 的收敛性能.

2) 针对多维网络资源和多样性通信需求,如何设计有效的 DRL 奖励函数. 对此,已有的研究大多利用多智能体的 DRL 架构分别训练多个智能体,使其拥有不同的功能,从而满足不同的通信需求. 不同于单智能体方法通过与环境进行交互来实现训练效果,多智能体 DRL 算法还能通过使一组智能体进行彼此之间的交互来学习最优策略[113]. 因此,基于多智能体 DRL 的资源管理方案通过并行解决多个问题,能同时满足不同用户与通信设备的个性化需求. 然而,由于需要同时考虑多个智能体,单一智能体的奖励和状态转变不仅取决于自己的动作,还取决于其他智能体的动作. 这意味着单个智能体的变化会影响整个训练环境的复杂度与稳定性,因此,奖励函数的设置将至关重要. 此外,多个智能体之间的频繁交互还使 DRL 的训练过程变得尤为复杂,从而导致动作空间随智能体的数量呈指数增长,进一步影响算法的复杂度.

3) 在实际移动通信场景中,真实数据样本通常是有限的,如何使 DRL 从少量的数据样本中取得良好的学习效果. DRL 是数据驱动的机器学习算法,它需要大量的训练样本来进行“试错”学习,因此,样本数据将直接影响 DRL 算法的训练过程. 随着接入设备的增加和用户需求的多样化,移动通信网络变得更加异构且动态,将智能体与环境进行频繁交互以获得大量样本数据则更为困难. 例如,在不完全感知或部分可观测情况下,MDP 模型不再有效,状态信息的数量不足以支撑 DRL 对最佳动作的决策. 同时,DRL 存在稀疏奖励问题[114],即由于完成训练目标的次数太少或者完成训练目标所需的步数太长,奖励空间中的负奖励样本远远多于正奖励样本,导致智能体在训练算法中难以获得奖励,从而学习缓慢甚至无法学习. 因此,如果交互样本不能获得奖励,那么该样本对 DRL 训练的贡献便很小,进而降低样本的利用效率.

5.2 研究展望

针对上述 DRL 在移动通信资源管理方案中存在的问题,本文尝试提出相关解决思路,并对可能的研究方向进行展望.

1) 针对 Distributional-DRL 算法存在局部最优的问题,可以适当增加算法的随机性来扩大对解的搜索范围,同时找到局部最优与全局最优的平衡点,在保证算法收敛速度与复杂度性能的前提下无限接近全局最优. 具体地,局部优化分别对每个子任务进行求解,其优势在于增强局部搜索能力并加快收敛速度,因此,适用于智能体的训练过程;全局优化则利用智能体之间的信息交互,协同实现全局优化的目标,因此,适用于寻找最优策略.

2) 针对多智能体 DRL 的奖励设置问题,可以通过模仿人类社会个体之间的信息共享和团结合作以及调节多个智能体之间的关系形成交流、共享与协作机制. 同时,将智能体按照不同的功能种类加以区别,分成不同类型的智能体组合,分别设置相应的奖励函数. 已有的研究多采取集中训练、分布式执行的方法,这意味着多个智能体拥有共同的奖励函数,缺乏灵活性与可扩展性. 例如,文献[115]针对 D2D 通信场景,提出共享邻近智能体的历史数据来进行训练. 一方面,集中训练能简化智能体的学习过程,但每个智能体只能执行不同的任务而非自主决策;另一方面,如果针对不同智能体分别设置激励机制,DRL 算法的复杂度将增加,从而占用更多运算资源,影响算法效率. 因此,如何设计智能体的激励机制以及多个智能体间的关系,从而在满足多样化需求的同时提升训练效果,实现智能且高效的移动通信资源管理、分配与调度,仍值得继续探索.

3) 针对 DRL 算法的少样本学习问题,一种改进方法是挖掘除状态信息外的其他可用信息. 除了让智能体常规地通过提取样本特征进行学习以外,还可以引入专家知识作为经验数据,辅助或指导 DRL 的训练过程. 例如,文献[116]利用模仿学习

在 RL 框架中灌输专家知识，智能体先尝试专家策略而非盲目行动，显著增加了 DRL 的学习速度. 另一种改进方法是针对已有的样本进行充分学习和利用. 例如，引入额外的存储器来存储近期的最优数据，令智能体在它和原始数据之间进行采样，然后，快速抓取价值较高的样本进行学习，有助于提高算法的样本利用率和训练效率. 此外，文献[117]利用迁移学习对元学习中 DNN 的权重进行转移，提出了元迁移学习，证明了大规模预训练的 DL 可以提供一个良好的“知识库”来进行有效的少样本学习. 其中，元学习可以利用大量类似的少样本任务，学习如何使智能体适应只有少量标记样本的新任务. 在此基础上，未来可将该架构引入 DRL，以弥补 DRL 在少样本学习方面的缺陷.

最后，通过对 DRL 与移动通信资源管理的进展与现状进行分析，本文认为 DRL 与通信资源管理的深度结合还需要考虑如下研究方向：

1）如何提升 DRL 在处理多任务训练问题和多智能体问题时的性能表现，从而更好地适应通信设备的激增与用户需求的多样化，是未来的重点研究方向. 针对多任务训练问题，DRL 可以结合迁移学习的能力，即将源任务中学习到的经验知识（如网络参数和策略知识）重新用于目标任务的迁移能力[118]，实现多任务间的知识共享，提高 DRL 算法在解决复杂资源管理问题时的学习效率与训练效果. 针对多智能体问题，已经有研究利用平均场博弈引导 DRL 的学习方向，解决了协作 MEC 的多任务分配问题[119]. 在此背景下，如何合理运用博弈论等理论与方法将 DRL 智能体之间的相互关系抽象为具体的数学公式，从而构建高效且智慧的移动通信资源管理方案，是值得进一步探索的研究思路.

2）如何保护用户数据的隐私与安全，成为移动通信资源管理研究的重要发展趋势. 随着区块链技术的发展与逐渐成熟，利用其安全性、透明性和分散性，构建可靠平台来提供可信的服务管理已被普遍接受. 因此，可以将 DRL 的学习与决策能力结合区块链的不可篡改与可追溯特性，实现保障用户隐私与数据安全的自治事务管理. 此外，由于能使多个智能体或任务在保证隐私安全以及合乎法律法规的基础上进行机器学习建模，联邦学习框架也可以实现安全智慧的移动通信资源管理.

3）DRL 与 MEC 的深度融合，将是 DRL 与移动通信资源管理中值得开展的研究方向. 随着多智能体 DRL 架构的广泛运用，基于 DRL 算法设计智能的 MEC 资源管理方案，实现通信服务和网络应用本地化、近距离和分布式的高效部署，成为当前研究热点. 另一方面，由于通信设备的激增、随时间变化的网络环境以及网络设备资源的异构性，利用 MEC 技术实现边缘设备与边缘服务器之间稳定、可靠与实时的交互是一项挑战. 因此，如何将多智能体 DRL 与 MEC 进行协同合作，从而为通信用户提供低延迟与高可靠性的智能服务将成为未来的研究重点.

参考文献：

[1] COTE D. Using machine learning in communication networks [Invited] [J]. IEEE/OSA Journal of Optical Communications and Networking, 2018, 10(10): 100-109.

[2] SUN Y, LIU J, WANG J, et al. When machine learning meets privacy in 6G: a survey [J]. IEEE Communications Surveys & Tutorials, 2020, 22(4): 2694-2724.

[3] MAO Q, HU F, HAO Q. Deep learning for intelligent wireless networks: a comprehensive survey [J]. IEEE Communications Surveys & Tutorials, 2018, 20(4): 2595-2621.

[4] YAU K L A, KOMISARCZUK P, TEAL P. Reinforcement learning for context awareness and intelligence in wireless networks: review, new features and open issues [J]. Journal of Network and Computer Applications, 2012, 35(1): 253-267.

[5] HUSSAIN F, HASSAN S A, HUSSAIN R, et al. Machine learning for resource management in cellular and IoT networks: potentials, current solutions, and open challenges [J]. IEEE Communications Surveys & Tutorials, 2020, 22(2): 1251-1275.

[6] CHEN M, CHALLITA U, SAAD W, et al. Artificial neural networks-based machine learning for wireless networks: a tutorial [J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3039-3071.

[7] ZHANG C, PATRAS P, HADDADI H. Deep learning in mobile and wireless networking: a survey [J]. IEEE Communications Surveys & Tutorials, 2019, 21(3): 2224-2287.

[8] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey [J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[9] LUONG N C, HOANG D T, GONG S, et al. Applications of deep reinforcement learning in communications and networking: a survey [J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133-3174.

[10] XIONG Z, ZHANG Y, NIYATO D, et al. Deep reinforcement learning for mobile 5G and beyond: fundamentals, applications, and challenges [J]. IEEE Vehicular Technology Magazine, 2019, 14(2): 44-52.

[11] 梁应敞, 谭俊杰, DUSIT N. 智能无线通信技术研究概况 [J]. 通信学报, 2020, 41(7): 1-17.
LIANG Y C, TAN J J, DUSIT N. Overview on intelligent wireless communication technology [J]. Journal on Communications, 2020, 41(7): 1-17. (in Chinese)

[12] 谭俊杰, 梁应敞. 面向智能通信的深度强化学习方法 [J]. 电子科技大学学报, 2020, 49(2): 169-181.
TAN J J, LIANG Y C. Deep reinforcement learning for intelligent communications [J]. Journal of University of Electronic Science and Technology of China, 2020, 49(2): 169-181. (in Chinese)

[13] PIAO C, LIU C H. Energy-efficient mobile crowdsensing by unmanned vehicles: a sequential deep reinforcement learning approach [J]. IEEE Internet of Things Journal, 2020, 7(7): 6312-6324.

[14] XIE J, SHAO Z, LI Y, et al. Deep reinforcement learning with optimized reward functions for robotic trajectory planning [J]. IEEE Access, 2019, 7: 105669-105679.

[15] CHEN L, CHEN Z, TAN B, et al. Agentgraph: toward universal dialogue management with structured deep reinforcement learning [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(9): 1378-1391.

[16] KIRAN R B, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: a survey [J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 4909-4926.

[17] ZHANG Z, LUO X, LIU T, et al. Proximal policy optimization with mixed distributed training [C]// 2019 IEEE 31st International Conference on Tools with Artificial Intelligence. Piscataway: IEEE, 2019: 1452-1456.

[18] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation [C] // 12th International Conference on Neural Information Processing Systems. New York: ACM, 1999: 1057-1063.

[19] BELLMAN R. A Markovian decision process [J]. Journal of Mathematics and Mechanics, 1957, 6(5): 679-684.

[20] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521: 436-444.

[21] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323: 533-536.

[22] HINTON G E, OSINDERO S, TEH Y. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.

[23] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[24] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[25] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model [J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[26] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM, 2014: 2672-2680.

[27] FADLULLAH Z M, TANG F, MAO B, et al. State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems [J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2432-2455.

[28] KOH P W, LIANG P. Understanding black-box predictions via influence functions [C] // 34th International Conference on Machine Learning. New York: ACM, 2017: 1885-1894.

[29] KENDALL A, GAL Y. What uncertainties do we need in bayesian deep learning for computer vision? [C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5580-5590.

[30] SUTTON R S, BACH F. Reinforcement learning: an introduction [M]. Cambridge, U. K.: Cambridge Univ. Press, 1998: 1-13.

[31] MODARES H, RANATUNGA I, LEWIS F L, et al. Optimized assistive human-robot interaction using reinforcement learning [J]. IEEE Transactions on Cybernetics, 2016, 46(3): 655-667.

[32] CHIANG H L, HSU J, FISER M, et al. RL-RRT: kinodynamic motion planning via learning reachability estimators from RL policies [J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4298-4305.

[33] SADEGHIANPOURHAMAMI N, DELEU J, DEVELDER C. Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning [J]. IEEE Transactions on Smart Grid, 2020, 11(1):

203-214.

[34] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains [J]. Artificial Intelligence, 1998, 101(1/2): 99-134.

[35] 周志华. 机器学习 [M]. 北京：清华大学出版社, 2016：371-393.

[36] JANG B, KIM M, HARERIMANA G, et al. Q-learning algorithms: a comprehensive classification and applications [J]. IEEE Access, 2019: 133653-133667.

[37] YOGESWARAN M, PONNAMBALAM S G. Reinforcement learning: exploration-exploitation dilemma in multi-agent foraging task [J]. Opsearch, 2012, 49(3): 223-236.

[38] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning [C] // Proceedings of the Workshops at the 26th Neural Information Processing Systems. New York: ACM, 2013: 201-220.

[39] LIN L. Self-improving reactive agents based on reinforcement learning, planning and teaching [J]. Machine Langnage, 1992, 8(3/4): 293-321.

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

[41] LI R, ZHAO Z, SUN Q, et al. Deep reinforcement learning for resource management in network slicing [J]. IEEE Access, 2018, 6: 74429-74441.

[42] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning [C] // Thirtieth AAAI Conference on Artificial Intelligence. New York: ACM, 2016: 2094-2100.

[43] 刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述 [J]. 计算机学报, 2019, 42(6): 1406-1438.
LIU J W, GAO F, LUO X L. Survey of deep reinforcement learning based on value function and policy gradient [J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438. (in Chinese)

[44] WANG Z Y, FREITAS D, LANCTOT M. Dueling network architectures for deep reinforcement learning [C] // 33rd International Conference on Machine Learning. New York: ACM, 2016: 1995-2003.

[45] BELLEMARE M G, DABNEY W, MUNOS R. A distributional perspective on reinforcement learning [C] // 34th International Conference on Machine Learning. New York: ACM, 2017: 449-458.

[46] PETERS J, SCHAAL S. Natural actor-critic [J]. Neurocomputing, 2008, 71(7/8/9): 1180-1190.

[47] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C] // 33rd International Conference on Machine Learning. New York: ACM, 2016: 1928-1937.

[48] TASFI N, CAPRETZ M. Noisy importance sampling actor-critic: an off-policy actor-critic with experience replay [C] // 2020 International Joint Conference on Neural Networks. Piscataway: IEEE, 2020: 1-8.

[49] ZHANG X, LI Z, JIANG J. Emotion attention-aware collaborative deep reinforcement learning for image cropping [J]. IEEE Transactions on Multimedia, 2021, 23: 2545-2560.

[50] LI X, SAMAKA M, CHAN H A, et al. Network slicing for 5G: challenges and opportunities [J]. IEEE Internet Computing, 2017, 21(5): 20-27.

[51] KATSALIS K, NIKAEIN N, SCHILLER E, et al. Network slices toward 5G communications: slicing the LTE network [J]. IEEE Communications Magazine, 2017, 55(8): 146-154.

[52] ZHANG H, LIU N, CHU X, et al. Network slicing based 5G and future mobile networks: mobility, resource management, and challenges [J]. IEEE Communications Magazine, 2017, 55(8): 138-145.

[53] ZHOU X, LI R, CHEN T, et al. Network slicing as a service: enabling enterprises' own software-defined cellular networks [J]. IEEE Communications Magazine, 2016, 54(7): 146-153.

[54] MENG S, WANG Z, DING H, et al. RAN slice strategy based on deep reinforcement learning for smart grid [C] // 2019 Computing, Communications and IoT Applications. Piscataway: IEEE, 2019: 6-11.

[55] ABIKO Y, SAITO T, IKEDA D, et al. Flexible resource block allocation to multiple slices for radio access network slicing using deep reinforcement learning [J]. IEEE Access, 2020, 8: 68183-68198.

[56] WEI F, FENG G, SUN Y, et al. Network slice reconfiguration by exploiting deep reinforcement learning with large action space [J]. IEEE Transactions on Network and Service Management, 2020, 17(4): 2197-2211.

[57] TAVAKOLI A, PARDO F, KORMUSHEV P. Action branching architectures for deep reinforcement learning [C] // Thirty-Second AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2018: 4131-4138.

[58] LI R, WANG C, ZHAO Z, et al. The LSTM-based advantage actor-critic learning for resource management in network slicing with user mobility [J]. IEEE

Communications Letters, 2020, 24(9): 2005-2009.

[59] HUA Y, LI R, ZHAO Z, et al. GAN-powered deep distributional reinforcement learning for resource management in network slicing [J]. IEEE Journal on Selected Areas in Communications, 2020, 38(2): 334-349.

[60] QI C, HUA Y, LI R, et al. Deep reinforcement learning with discrete normalized advantage functions for resource management in network slicing [J]. IEEE Communications Letters, 2019, 23(8): 1337-1341.

[61] SUN G, BOATENG G O, AYEPAH-MENSAH D, et al. Autonomous resource slicing for virtualized vehicular networks with D2D communications based on deep reinforcement learning [J]. IEEE Systems Journal, 2020, 14(4): 4694-4705.

[62] SUN G, AL-WARD H, BOATENG G O, et al. Autonomous cache resource slicing and content placement at virtualized mobile edge network [J]. IEEE Access, 2019, 7: 84727-84743.

[63] LEINONEN M, CODREANU M, JUNTTI M. Distributed joint resource and routing optimization in wireless sensor networks via alternating direction method of multipliers [J]. IEEE Transactions on Wireless Communications, 2013, 12(11): 5454-5467.

[64] LIANG C, YU F, YAO H, et al. Virtual resource allocation in information-centric wireless networks with virtualization [J]. IEEE Transactions on Vehicular Technology, 2016, 65(12): 9902-9914.

[65] EALIYAS A, JENO-LOVESUM S P. Resource allocation and scheduling methods in cloud: a survey [C]//2018 Second International Conference on Computing Methodologies and Communication. Piscataway: IEEE, 2018: 601-604.

[66] IQBAL A, THAM M, CHANG Y. Double deep Q-network for power allocation in cloud radio access network [C] // 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology. Piscataway: IEEE, 2020: 272-277.

[67] XU Z, WANG Y, TANG J, et al. A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs [C]//2017 IEEE International Conference on Communications. Piscataway: IEEE, 2017: 1-6.

[68] MIJUMBI J, SERRAT J, GORRICHO J, et al. Network function virtualization: state-of-the-art and research challenges [J]. IEEE Communications Surveys & Tutorials, 2016, 18(1): 236-262.

[69] HE B, WANG J, QI Q, et al. Towards intelligent provisioning of virtualized network functions in cloud of things: a deep reinforcement learning based approach [J]. IEEE Transactions on Cloud Computing, 2022, 10(2): 1262-1274.

[70] GU L, ZENG D, LI W, et al. Intelligent VNF orchestration and flow scheduling via model-assisted deep reinforcement learning [J]. IEEE Journal on Selected Areas in Communications, 2020, 38(2): 279-291.

[71] PUJOL-ROIG J S, GUTIERREZ-ESTEVEZ D M, GUNDUZ D. Management and orchestration of virtual network functions via deep reinforcement learning [J]. IEEE Journal on Selected Areas in Communications, 2020, 38(2): 304-317.

[72] CHEN X, ZHAO Z, WU C, et al. Multi-tenant cross-slice resource orchestration: a deep reinforcement learning approach [J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2377-2392.

[73] GUO S, DAI Y, XU S, et al. Trusted cloud-edge network resource management: DRL-driven service function chain orchestration for IoT [J]. IEEE Internet of Things Journal, 2020, 7(7): 6010-6022.

[74] CHEN X, ZHU Z, GUO J, et al. Leveraging mixed-strategy gaming to realize incentive-driven VNF service chain provisioning in broker-based elastic optical inter-datacenter networks [J]. IEEE/OSA Journal of Optical Communications and Networking, 2018, 10(2): A232-A240.

[75] HUANG Y, QIAO X, REN P, et al. A lightweight collaborative deep neural network for the mobile Web in edge cloud [J]. IEEE Transactions on Mobile Computing, 2022, 21(7): 2289-2305.

[76] GUO T, ZHANG H, HUANG H, et al. Multi-resource fair allocation for composited services in edge micro-clouds [C] // 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. Piscataway: IEEE, 2019: 405-412.

[77] RJOUB G, BENTAHAR J, ABDEL O, et al. Deep smart scheduling: a deep learning approach for automated big data scheduling over the cloud [C] // 2019 7th International Conference on Future Internet of Things and Cloud. Piscataway: IEEE, 2019: 189-196.

[78] BONOMI F, MILITO R, ZHU J, et al. Fog computing and its role in the Internet of things [C]//Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing. New York: ACM, 2012: 13-16.

[79] PENG M, YAN S, ZHANG K, et al. Fog-computing-

based radio access networks: issues and challenges [J]. IEEE Network, 2015, 30(4): 46-53.

[80] SHI J, DU J, WANG J, et al. Priority-aware task offloading in vehicular fog computing based on deep reinforcement learning [J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 16067-16081.

[81] WANG J, LV T, HUANG P, et al. Mobility-aware partial computation offloading in vehicular networks: a deep reinforcement learning based scheme [J]. China Communications, 2020, 17(10): 31-49.

[82] NING Z, DONG P, WANG X, et al. Deep reinforcement learning for intelligent Internet of vehicles: an energy-efficient computational offloading scheme [J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(4): 1060-1072.

[83] PENG M, ZHANG K. Recent advances in fog radio access networks: performance analysis and radio resource allocation [J]. IEEE Access, 2016, 4: 5003-5009.

[84] WANG W, LAN R, GU J X, et al. Edge caching at base stations with device-to-device offloading [J]. IEEE Access, 2017, 5: 6399-6410.

[85] ZHAO Z, PENG M, DING Z, et al. Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks [J]. IEEE Journal on Selected Areas in Communication, 2016, 34(5): 1207-1221.

[86] ZHOU Y, PENG M, YAN S, et al. Deep reinforcement learning based coded caching scheme in fog radio access networks [C]//2018 IEEE/CIC International Conference on Communications in China. Piscataway: IEEE, 2018: 309-313.

[87] SUN Y, PEMG M, MAO S. Deep reinforcement learning-based mode selection and resource management for green fog radio access networks [J]. IEEE Internet of Things Journal, 2019, 6(2): 1960-1971.

[88] RAHMAN G M S, PENG M, YAN S, et al. Learning based joint cache and power allocation in fog radio access networks [J]. IEEE Transactions on Vehicular Technology, 2020, 69(4): 4401-4411.

[89] RAHMAN G, DANG T, AHMED M. Deep reinforcement learning based computation offloading and resource allocation for low-latency fog radio access networks [J]. Intelligent and Converged Networks, 2020, 1(3): 243-257.

[90] JIANG F, MA R, SUN C, et al. Dueling deep Q-network learning based computing offloading scheme for F-RAN [C]//2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. Piscataway: IEEE, 2020: 1-6.

[91] ABBAS N, ZHANG Y, TAHERKORDI A, et al. Mobile edge computing: a survey [J]. IEEE Internet of Things Journal, 2018, 5(1): 450-465.

[92] RODRIGUES T, SUTP K, NISHIYAMA H, et al. Machine learning meets computation and communication control in evolving edge and cloud: challenges and future perspective [J]. IEEE Communications Surveys & Tutorials, 2020, 22(1): 38-67.

[93] TRAN T X, HAJISAMI A, PANDEY P, et al. Collaborative mobile edge computing in 5G networks: new paradigms scenarios and challenges [J]. IEEE Communications Magazine, 2017, 55(4): 54-61.

[94] MAO Y, YOU C, ZHANG J, et al. A survey on mobile edge computing: the communication perspective [J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2322-2358.

[95] MEHRABI M, SALAH H, FITZEK F H P. A survey on mobility management for MEC-enabled systems [C]// 2019 IEEE 2nd 5G World Forum. Piscataway: IEEE, 2019: 259-263.

[96] KHAN I, TAO X, RAHMAN G M S, et al. Advanced energy-efficient computation offloading using deep reinforcement learning in MTC edge computing [J]. IEEE Access, 2020, 8: 82867-82875.

[97] LIU Y, YU H, XIE S, et al. Deep reinforcement learning for offloading and resource allocation in vehicle edge computing and networks [J]. IEEE Transactions on Vehicular Technology, 2019, 68(11): 11158-11168.

[98] LEI L, XU J, XIONG X, et al. Multiuser resource control with deep reinforcement learning in IoT edge computing [J]. IEEE Internet of Things Journal, 2019, 6(6): 10119-10133.

[99] QIAN L, WU Y, JIANG F, et al. NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial Internet of things [J]. IEEE Transactions on Industrial Informatics, 2021, 17(8): 5688-5698.

[100] XU S, GUO C, HU R, et al. Blockchain inspired secure computation offloading in a vehicular cloud network [J/OL]. IEEE Internet of Things Journal [2021-08-30]. https://ieeexplore.ieee.org/document/9336659.

[101] OUYANG L, YUAN Y, WANG F. Learning markets: an AI collaboration framework based on blockchain and smart contracts [J/OL]. IEEE Internet of Things Journal [2021-08-30]. https://ieeexplore.ieee.org/document/9234516.

[102] CHEN J, CHEN S, WANG Q, et al. IRAF: a deep reinforcement learning approach for collaborative mobile edge computing IoT networks [J]. IEEE Internet of Things Journal, 2019, 6(4): 7011-7024.

[103] LI J, GAO H, LV T, et al. Deep reinforcement learning based computation offloading and resource allocation for MEC [C]//2018 IEEE Wireless Communications and Networking Conference. Piscataway: IEEE, 2018: 1-6.

[104] MIN M, XIAO L, CHEN Y, et al. Learning-based computation offloading for IoT devices with energy harvesting [J]. IEEE Transactions on Vehicular Technology, 2019, 68(2): 1930-1941.

[105] YANG H, ZHONG W, CHEB C, et al. Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive Internet of things [J]. IEEE Internet of Things Journal, 2020, 7(6): 5677-5689.

[106] WEI Y, YU F, SONG M, et al. User scheduling and resource allocation in HetNets with hybrid energy supply: an actor-critic reinforcement learning approach [J]. IEEE Transactions on Wireless Communications, 2018, 17(1): 680-692.

[107] CHEN X, PROULX B, GONG X, et al. Exploiting social ties for cooperative D2D communications: a mobile social networking case [J]. IEEE/ACM Transactions on Networking, 2015, 23(5): 1471-1484.

[108] ZHANG H, CHONG S, ZHANG X, et al. A deep reinforcement learning based D2D relay selection and power level allocation in mmWave vehicular networks [J]. IEEE Wireless Communications Letters, 2020, 9(3): 416-419.

[109] BI Z, ZHOU W. Deep reinforcement learning based power allocation for D2D network [C]//2020 IEEE 91st Vehicular Technology Conference. Piscataway: IEEE, 2020: 1-5.

[110] TAN J, LIANG Y, ZHANG L, et al. Deep reinforcement learning for joint channel selection and power control in D2D networks [J]. IEEE Transactions on Wireless Communications, 2021, 20(2): 1363-1378.

[111] NGUYEN K K, DUONG T Q, VIEN N A, et al. Non-cooperative energy efficient power allocation game in D2D communication: a multi-agent deep reinforcement learning approach [J]. IEEE Access, 2019, 7: 100480-100490.

[112] ZHAO N, LIANG Y, NIYATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks [J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141-5152.

[113] FERIANI A, HOSSAIN E. Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: a tutorial [J]. IEEE Communications Surveys & Tutorials, 2021, 23(2): 1226-1252.

[114] 杨惟轶, 白辰甲, 蔡超, 等. 深度强化学习中稀疏奖励问题研究综述 [J]. 计算机科学, 2020, 47(3): 1-13.
YANG W Y, BAI C J, CAI C, et al. Survey on sparse reward in deep reinforcement learning [J]. Computer Science, 2020, 47(3): 1-13. (in Chinese)

[115] LI Z, GUO C. Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications [J]. IEEE Transactions on Vehicular Technology, 2020, 69(2): 1828-1840.

[116] GUO W, TIAN W, YE Y, et al. Cloud resource scheduling with deep reinforcement learning and imitation learning [J]. IEEE Internet of Things Journal, 2021, 8(5): 3576-3586.

[117] SUN Q, LIU Y, CHEN Z, et al. Meta-transfer learning through hard tasks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 3: 1443-1456.

[118] TEH Y W, BAPST V, CZARNECKI W M, et al. Distral: robust multitask reinforcement learning [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 4499-4509.

[119] SHI D, GAO H, WANG L, et al. Mean field game guided deep reinforcement learning for task placement in cooperative multiaccess edge computing [J]. IEEE Internet of Things Journal, 2020, 7(10): 9330-9340.

(责任编辑 梁 洁)