计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800095

移动边缘计算中基于深度强化学习的任务卸载研究进展

梁俊斌[1,2]　张海涵[1,2]　蒋　婵[3]　王天舒[4]
1 广西大学计算机与电子信息学院　南宁 530004
2 广西多媒体通信与网络技术重点实验室　南宁 530004
3 广西大学行健文理学院　南宁 530004
4 东软集团(南宁)有限公司　南宁 530007
(liangjb2002@163.com)

摘　要　移动边缘计算是近年出现的一种新型网络计算模式,它允许将具有较强计算能力和存储性能的服务器节点放置在更加靠近移动设备的网络边缘(如基站附近),让移动设备可以近距离地卸载任务到边缘设备进行处理,从而解决了传统网络由于移动设备的计算和存储能力弱且能量较有限,从而不得不耗费大量时间、能量且不安全地将任务卸载到远方的云平台进行处理的弊端。但是,如何让仅掌握局部有限信息(如邻居数量)的设备根据任务的大小和数量选择卸载任务到本地,还是在无线信道随时间变化的动态网络中选择延迟、能耗均最优的移动边缘计算服务器进行全部或部分的任务卸载,是一个多目标规划问题,求解难度较高。传统的优化技术(如凸优化等)很难获得较好的结果。而深度强化学习是一种将深度学习与强化学习相结合的新型人工智能算法技术,能够对复杂的协作、博弈等问题作出更准确的决策,在工业、农业、商业等多个领域具有广阔的应用前景。近年来,利用深度强化学习来优化移动边缘计算网络中的任务卸载成为一种新的研究趋势。最近三年来,一些研究者对其进行了初步的探索,并达到了比以往单独使用深度学习或强化学习更低的延迟和能耗,但是仍存在很多不足之处。为了进一步推进该领域的研究,文中对近年来国内外的相关工作进行了详细地分析、对比和总结,归纳了它们的优缺点,并对未来可能深入研究的方向进行了讨论。

关键词:移动边缘计算;深度强化学习;任务卸载;卸载决策;深度学习;强化学习
中图法分类号　TP393

Research Progress of Task Offloading Based on Deep Reinforcement Learning in Mobile Edge Computing

LIANG Jun-bin[1,2],ZHANG Hai-han[1,2],JIANG Chan[3] and WANG Tian-shu[4]
1 School of Computer,Electronics and Information,Guangxi University,Nanning 530004,China
2 Guangxi Key Laboratory of Multimedia Communication and Network Technology,Nanning 530004,China
3 XingJian College of Science and Liberal Arts of Guangxi University,Nanning 530004,China
4 Neusoft Group (Nanning) Co.,Ltd,Nanning 530007,China

Abstract　Mobile edge computing is a new type of network computing mode that has emerged in recent years. It allows server nodes with strong computing power and storage performance to be placed closer to the edge of the network of mobile devices (such as near base stations),allowing mobile devices to offload tasks to edge devices for processing closely,thereby alleviates the disadvantages of traditional networks that have to spend a lot of time,energy and unsafely offload tasks to remote cloud platforms for processing due to weak computing and storage capabilities of mobile devices and limited energy. However,how to make a device that only has limited local information (such as the number of neighbors) chooses to offload tasks to the local site according to the size and number of tasks,or chooses the mobile edge computing server with the optimal delay and energy consumption in a dynamic network where the wireless channel changes with time,to perform all or part of the task offloading,is a multi-objective programming problem and has a high degree of difficulty in solving. It is difficult to obtain better results with traditional optimization techniques(such as convex optimization). Deep reinforcement learning is a new type of artificial intelligence algorithm tech-

到稿日期:2020-08-16　返修日期:2020-12-02　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61562005);广西重点研发计划项目(桂科 AB19259006);广西自然科学基金(2019GXNSFAA185042,2018GXNSFBA281169)
This work was supported by the National Natural Science Foundation of China(61562005),Guangxi Key Research and Development Plan Project (AB19259006) and Natural Science Foundation of Guangxi(2019GXNSFAA185042,2018GXNSFBA281169).
通信作者:张海涵(1146795832@qq.com)

nology that combines deep learning and reinforcement learning. It can make more accurate decision-making results for complex collaboration, game and other issues. It has broad application prospects in many fields such as industry, agriculture and commerce. In recent years, It has become a new research trend to use deep reinforcement learning method to optimize task offloading in mobile edge computing networks. In the past three years, some researchers have conducted preliminary explorations on it, and achieved lower latency and energy consumption than using deep learning or reinforcement learning alone in the past, but there are still many shortcomings. In order to further advance the research in this field, this paper analyzes, compares and summarizes the domestic and foreign related work in recent years, summarizes their advantages and disadvantages, and discusses the possible future in research directions.

Keywords Mobile edge computing, Deep reinforcement learning, Task offloading, Offloading decision, Deep learning, Reinforcement learning

1 引言

移动边缘计算是近年来出现的一种新型网络数据处理模式,它将服务器放置在移动网络的边缘(简称边缘服务器),如基站,使得移动设备(简称设备)可以把计算密集型任务卸载到这些服务器进行处理,从而扩展移动设备的计算和存储能力,降低移动设备的能耗,进而延长移动设备电池的使用时间[1-4]。此外,相比传统的云计算模式,这些任务不需要传输到远方的云平台,因此任务处理的延迟得到了大幅降低[5-7]。

但是,在移动边缘计算中存在多个设备和边缘服务器,而设备仅掌握通信半径内的局部边缘服务器的信息,如何在全局范围内选择合适的边缘服务器进行任务卸载,从而使得能耗、延迟均最优,是一个 PSPACE-Hard 问题[8],比 NP-Hard 问题更难以在多项式时间内得到解决。此外,任务到达的不确定性和信道状态的动态性也给任务卸载带来很大的挑战[9]。面对这一多目标优化问题,传统的优化技术(如凸优化等)很难获得较好的结果。

深度强化学习是将深度学习与强化学习相结合的一种人工智能算法,被广泛应用于各种复杂决策求解领域[10-12],如组合优化、多方博弈等,具有重要的研究价值。其中,强化学习是一种机器学习方法,其思想是在一个交互环境中利用创建的软件智能体(Agent)不断与环境互动来进行测试,通过环境反馈的奖励或惩罚信息,逐步逼近最优的结果。深度学习也是一种机器学习技术,它能够通过建立多层人工神经网络,对原始数据进行自动特征提取。深度强化学习就是利用深度学习技术来扩展传统强化学习的方法,从而获得更好的求解多目标优化问题的能力。此外,深度强化学习面对以上问题还有两点优势:1)与许多一次性优化的方法[13-17]相比,深度强化学习可以随环境变化调整策略;2)其在学习过程中不需要了解关于网络状态随时间变化规律的相关先验知识。

因此,目前已有一些研究提出利用深度强化学习方法来解决移动边缘计算中的任务卸载问题,并取得了一定的进展(如文献[18-20]),而且获得了比使用深度学习更优的延迟与能耗。虽然此前已有使用深度学习或强化学习来解决该问题的方案,但将两者结合进行移动边缘计算网络的任务卸载优化是近三年才开始兴起的。该领域的研究尚处于起步阶段,仍存在很多亟待解决的问题。为了推动和深化该领域的研究,我们对当前的研究进展进行详细的分析和对比,总结出各个工作的优缺点,并对存在的问题和下一步需要开展的研究进行了讨论和展望。

2 国内外研究现状

与已有的优化方案相比,使用深度强化学习进行任务卸载的优化可以获得更优的时延、能耗和任务完成率。根据不同的决策流程,可将近三年来的研究工作分为两类:基于价值的深度强化学习方案和基于策略的深度强化学习方案。其中,基于价值的方案通过评估候选的多个动作的价值来选出要执行的动作。动作的价值是指动作所能带来的预期收益,这个收益包括系统通过执行所选动作得到的环境的即时奖惩和后续得到的奖惩。而基于策略的方案直接学习由状态到动作的映射,在选择动作时不需要评估候选动作的价值。此外,基于策略的方案可以对连续型动作进行决策,如卸载任务时的传输功率。而基于价值的方案只能将传输功率离散化为几个级别,进行较为粗颗粒的决策。

2.1 基于价值的深度强化学习方案

基于价值的深度强化学习算法寻求学习最优行为,它利用各种行为来探索环境,从而找到全局最优的结果。根据探索过程中执行和改进策略方案的不同,可以将目前存在的工作分为两种类型:脱离策略方案和依存策略方案。

(1)脱离策略方案是指算法由两种策略组成:一种称为目标策略,主要用于学习最优行为;另一种称为行为策略,用于探索不同的行为。算法通过行为策略与环境互动产生用于学习的数据,从而改进目标策略,由此得到更优的结果。在算法执行过程中,由于行为策略的结果会偏离目标策略的结果,脱离策略方案由此而得名。

(2)依存策略方案是指算法中行为策略与目标策略相一致。算法通过行为策略与环境互动产生用于学习的数据,从而改进行为策略,然后通过这个策略与环境互动,再改进,如此反复迭代得到最优结果。在算法执行过程中,由于行为策略的结果与目标策略的结果相一致,依存策略方案由此而得名。

脱离策略方案更能保证行为策略的探索性,使产生的数据覆盖范围更全面。此外,它运用起来更加灵活,比如可以使用多个行为策略同时为目标策略提供学习数据。但是,行为策略偏离目标策略而产生的行为可能带来更低的奖励或更高的惩罚。

依存策略方案由于行为策略与目标策略相一致,在学习中可能带来更高的奖励或更低的惩罚。但是,其行为策略的

探索性受目标策略的影响更大，更易陷入局部最优。

2.1.1 脱离策略方案

根据优化方案思考角度的不同，可以将以下工作分成3类：优化价值函数学习的方案、优化神经网络的方案和优化移动边缘卸载可靠性的方案。

(1)优化价值函数学习的方案

2018年，Li等[18]针对多个移动设备向一台移动边缘计算服务器卸载计算任务的问题提出了卸载方案，并证明了此问题是一个由背包问题扩展而来的非凸问题，属于NP-Hard的范围。他们设计了一种基于价值的深度强化学习方案。该方案首先以一定概率进行探索，一定概率根据价值函数来决定是否卸载任务到服务器，然后使用一个多层人工神经网络来近似这个价值函数，并通过反复的试错过程来优化这个价值函数。同时，算法通过不断训练这个网络，使整个系统的总成本最小化。其中，总成本包括所有用户设备向边缘服务器卸载任务的时延总和及能耗总和。仿真实验结果表明，该方案与传统强化学习算法相比可以达到更优的时延和能耗。但是，通过该方案学得的价值函数准确度还不够高。在基于价值的方案中，算法根据价值函数作决策。因此其学得的价值函数误差越小，就越有助于进行决策。文献[13,21-23]分别从优化学习价值函数的角度改进了方案。2019年，Huang等[21]以最小化能耗成本和时延为目标，改进了决策方案。文献[18]中的方案在更新神经网络参数时，使用同一个网络与得到的奖惩值来计算损失函数的值，以更新这个网络的参数。然而，每次迭代后又用更新过的网络来计算新的损失函数值，这使得损失函数的值变得不稳定。而文献[21]中使用了另一个参数相对固定的神经网络(目标网络)来计算损失函数。算法每迭代一定的次数之后，目标网络的参数才会更新一次。因此，其损失函数相对稳定，从而提高了该方案的收敛速度。此外，该方案的决策方式也与文献[18]不同，它是分别决定每个任务是否卸载，而不是决定是否卸载一个时间周期内到来的所有任务。仿真实验结果表明，该方案在性能上优于文献[24]中提出的卸载方案MUMTO。

以上方案在对目标策略进行更新时根据状态选取值最大的动作，然后使用这个值与奖惩值来更新价值函数，从而改进策略。但是，这种取最大值的操作可能会导致对动作价值的估计过高。Yao等[22]提出了另一种改进方案。该方案首先将选取最大值动作与估算动作价值两个操作解耦，分别使用两个神经网络进行操作[25]。然后，算法每迭代一定次数后，将这两个网络的值同步。此外，该方案考虑到真实环境中不同任务的重要程度不同，比较重要的任务如果处理失败则可能造成更大的损失。为此该方案将任务的重要程度分为5个优先级。在学习过程中，优先级更高的任务被处理后所获得的奖励也越大。但是，以上的价值函数都是表示某一状态下执行某一动作的值，而价值函数还可以进一步解耦，从而得到更接近真实值的结果。He等[23]提出了一种新方案，将价值函数分解为某一状态的价值(状态价值)和这一状态下某一动作价值与所有动作平均价值的差(动作优势)。算法通过这种方式可以得到当前的奖惩受动作还是受状态影响大，从而作出更优的决策。以上成果的优缺点如表1所列。

表1 优化价值函数学习的方案优缺点对比

Table 1 Comparison of advantages and disadvantages of schemes for optimizing value function learning

方案	优点	缺点
Li等[18]	结构相对简单，占用计算资源与存储资源相对较少	收敛速度慢，对价值函数估计准确度较低
Huang等[21]	收敛速度得到提升	可能对动作价值估计过高
Yao等[22]	防止动作价值的过高估计	无法分辨状态与动作分别对当前的奖惩有多大影响
He等[23]	将价值函数解耦为状态与动作两个价值函数，对价值估计更准确	引入更多的参数和计算量

(2)优化神经网络的方案

以上方案都是从优化价值函数学习方式的角度改进方案。而深度强化学习由于引入了多层人工神经网络，会带来参数过多、收敛慢等问题。文献[26-28]从优化深度神经网络的角度改进了方案。2019年，Min等[26]针对深度强化学习中深度神经网络会带来参数过多、训练速度慢的问题，提出了卸载方案。他们使用卷积神经网络来计算价值函数，通过卷积运算减少了参数数量和算法的时间复杂度。其次，该方案使用迁移学习[29]，利用类似环境中的卸载经验来初始化卷积神经网络的参数，从而减少了程序在初始阶段的随机探索卸载过程，加快了学习速度。此外，他们还引入了能量收集技术，通过捕获周围的可再生能源(如太阳能辐射、风力发电、人体运动)为移动设备提供能源，来延长电池寿命。这样可以缓解移动设备能源不足问题，从而提高任务卸载的性能。在与文献[18]中方案的实验对照中，该方案的能源消耗、计算延迟和任务丢失率分别降低了58.3%，26.7%，55.5%。

然而，深度强化学习利用人工神经网络来近似求解价值函数会遇到以下问题。由于状态空间过大，神经网络不能观测到所有的状态，需要这个网络在未观测过的数据上也能表现良好，即泛化能力强。如果泛化能力不足，所训练的网络可能对已经观测过的状态表现良好，而遇到新的状态就表现很差，即过拟合问题。Ning等[27]针对过拟合问题改进了卸载方案。该方案在每轮神经网络训练中随机隐藏一部分神经元，这样简化了神经网络的结构[30]，不但缓解了过拟合问题，并且减少了总的计算量。

在卸载过程中，一些任务要在其前置任务都执行结束后才能开始执行。然而，以上方案没有考虑到任务之间的这种依赖性，并且，传统的神经网络不能很好地提取先后到来的任务之间相互的作用关系。2020年，Lu等[28]针对这个问题提出了考虑任务间依赖性的卸载方案。该方案首先将任务与任务之间的依赖关系表示为有向无环图，并且还考虑到一些任务必须在本地执行，比如设备I/O任务不能被卸载。该方案使用一种带有门控的人工神经网络[31]，使网络具有学习输入数据的长期依赖关系的能力，从而可以考虑历史数据来对当前状态进行估计。最后，该方案在边缘计算仿真平台iFogSim[32]上对多种算法生成的卸载策略进行仿真。通过比较能耗、成本、负载均衡、延迟、平均执行时间等因素，表明该方案优于目前已有的深度强化学习方法DQN[10]。以上方案的分析对比如表2所列。

表 2 优化神经网络的方案优缺点对比

Table 2 Comparison of advantages and disadvantages of schemes for optimizing neural networks

方案	优点	缺点
Min 等[26]	使用卷积网络和迁移学习加快学习速度	可能会有过拟合问题
Ning 等[27]	使用 Dropout 方法防止过拟合	没有考虑任务之间的依赖性
Lu 等[28]	可以考虑历史数据来对当前状态进行估计	网络更加复杂,引入更多的参数

(3)优化移动边缘卸载可靠性的方案

移动边缘网络中如何进行安全且可靠的卸载也是一个重要问题。针对这个问题,文献[33-34]分别提出了优化任务卸载可靠性的方案。Huang 等[33]考虑到安全是移动边缘计算的关键问题之一,被卸载到边缘服务器的任务很容易受到外部的恶意攻击。例如,从移动设备转移到边缘服务器的计算任务可能会被恶意的窃听者故意偷听[35]。他们提出了保证任务信息保密性和完整性的卸载策略。首先,该算法使用加密的方式保护任务信息不被窃取,并使用哈希函数验证任务信息的完整性,防止任务信息被篡改。然后,算法将任务的保密性等级和完整性等级设置成不同级别,不同级别的任务使用不同的加密算法和哈希函数,从而造成不同的开销,最终影响卸载决策。最后,算法将任务与边缘服务器信息作为输入通过价值函数作出卸载决策。该方案在一定的时延和能耗约束下能保证安全地卸载任务。但是,该工作没有考虑到移动设备的移动性。2020 年,Zhang 等[34]面向具有高速移动性和时变拓扑的车联网分别提出了卸载方案。该方案的应用场景是拥有多个 MEC 服务器的车载网络环境,车辆可以通过以下几种传输方式进行任务卸载:1)通过基站卸载;2)通过路边单元卸载;3)通过其他车辆接力再通过路边单元卸载。首先,该方案通过一个控制中心,将任务信息和服务器状态等信息作为输入传输到深度神经网络,输出每个卸载动作的近似价值,通过训练得到近似的价值函数。然后,算法根据价值函数的大小判断任务应该卸载到哪个服务器。最后,在卸载过程中,算法采用冗余的方法,通过多条路径传输多个任务副本到目标服务器,避免因某条路径传输失败而导致卸载失败。此方案提高了移动边缘卸载中任务卸载的可靠性。以上成果的优缺点如表 3 所列。

表 3 优化移动边缘卸载可靠性的方案优缺点对比

Table 3 Comparison of advantages and disadvantages of schemes for optimizing mobile edge offloading reliability

方案	优点	缺点
Huang 等[33]	提高了卸载的安全性,防止任务信息被窃取或篡改	为了保证安全卸载,会使用额外的计算资源,没有考虑设备的移动性
Zhang 等[34]	一定程度上保证了可靠的卸载	不能确保可靠的卸载,并且会占用更多的网络资源

2.1.2 依存策略方案

2019 年,Chen 等[20]提出了一种基于函数分解的同策略方案。该方案首先运用函数分解技术将价值函数表示为多个更简单的价值函数的累加和;然后分别用多个深度神经网络来近似这些价值函数;最后,通过迭代训练这些深度神经网络得到多个价值函数的近似值,根据它们的累加和来进行决策。另外,该方案在迭代过程中,行为策略与目标策略都是以一种概率探索的方式产生动作。这种同策略方案对学习过程中智能代理的表现好坏更加敏感。仿真实验结果表明,本文提出的同策略方案在学习过程中相比异策略的方案有更好的性能。

Maurice 等[36]针对多址接入的移动边缘计算卸载问题,设计了基于备选动作的卸载算法。与文献[20]中方案不同的是,该算法中的价值函数已经确定,而不是通过训练人工神经网络得到的。该算法首先通过向神经网络输入信道状态信息,从而输出一组备选动作;然后,通过已给出的价值函数来比较备选动作的价值,选择其中最优的动作并执行;最后,将此动作与信道状态信息加入待训练的数据集,并通过此数据集训练人工神经网络。实验结果表明该方案可以得到近似最优的结果。但是,该方案需要收集全部移动设备的信道状态信息来进行决策,而这些信息是全局且时变的,收集需要耗费大量的时间和能量。

2020 年,Alfakih 等[37]设计了 3 层架构的方案,计算任务不仅可以在终端和边缘服务器上计算,还可以卸载到云上。这样可以利用云中心的计算资源处理计算密集型且时延要求不高的计算任务,从而缓解移动边缘服务器的计算压力。该方案还考虑到移动设备的移动性,将边缘服务器按照距离分为不同区域。移动设备如果从一个区域移动到另一个区域,卸载任务的处理结果可以通过一个总的控制器传输到相应区域的边缘计算服务器,然后发送到相应的移动设备。以上方案的分析对比如表 4 所列。

表 4 依存策略方案优缺点对比

Table 4 Comparison of advantages and disadvantages of on-policy schemes

方案	优点	缺点
Chen 等[20]	使用函数分解方式使价值函数的估计更准确	需要训练更多的网络
Maurice 等[36]	使用备选动作提高高价值动作出现的可能性	需要获得全局信息,更加耗费资源
Alfakih 等[37]	考虑利用云资源进行卸载	网络结构更加复杂,更难以管理

2.2 基于策略的深度强化学习方案

基于策略的深度强化学习算法学习的目标策略是从环境的感知状态到采取各行动的概率的映射。因为目标策略根据状态映射的动作概率随机选择动作,所以称这种策略为随机性策略。相反,如果目标策略是根据状态直接映射到确定的某个动作,则称为确定性策略。根据目标策略类型的不同,可以将目前存在的工作分为两种类型:随机性策略方案和确定性策略方案。

随机性策略方案由于其根据概率随机选择动作,所以天然具有探索性,不需要配合探索性的方法进行学习。但是,此类方案由于要学习逼近所有动作的概率,所以可能需要相对更多的样本。

确定性策略方案最终学到的目标策略不具有探索性,因此在其学习过程中要用其他方法来探索环境。但是,其只需要学习每个状态下预期收益最大的动作,更加简单直接,学习过程只需较少的样本。

2.2.1 随机性策略方案

随机性策略方案可以使用一个线程来学习策略，也可以使用多个线程并行地学习一个策略。

根据是否可以并行学习策略的标准，可以将目前存在的工作分为两种类型：非并行学习方案和并行学习方案。

(1)非并行学习方案

2019 年，Zhang 等[38]针对单个移动设备向单个移动边缘服务器卸载任务的问题，提出了基于策略的深度强化学习方案。该方案首先将策略表示为从环境的感知状态到采取行动的概率的映射；然后通过一个人工神经网络表示这个策略；最后，通过反复的试错学习调整神经网络，调整各状态下采取各行动的概率，从而学习最优策略。基于策略的方案所学的策略是在一个状态下根据概率采取动作的随机策略，基于价值的方案则不同，它是在一个状态下根据价值函数采取值最大的行动。仿真实验表明，该方案提出的策略能在一定程度上降低移动设备计算任务的时延和能耗。但是，每次更新策略时对网络做多大幅度的调整是一个问题。幅度太大可能会导致不能收敛的问题，而幅度太小又可能导致收敛慢的问题。

针对上述问题，Liu 等[39]提出了改进的方案。该方案首先使用一个人工神经网络(行为网络)来学习策略作出决策，使用另一个人工神经网络(评判网络)来对这个决策打分[40]；然后通过这个分数对行为网络做出调整，从而控制调整的幅度大小；最后，通过反复迭代学习最优策略。其中，评判网络近似计算一个动作价值函数。与基于价值的深度强化学习方案不同的是，算法不通过这个网络作决策，而是通过它来辅助调整评判网络。仿真实验结果表明，该方案的收敛速度和性能优于基于价值的深度强化学习基线方案，但是该方案未将行为策略和目标策略分离，是一个依存策略方案，可能会陷入局部最优。

2020 年，Zhan 等[41]为了解决这一问题，提出了一种脱离策略的方案。该方案首先使用两个人工神经网络分别近似行为策略和目标策略；然后由行为策略产生学习数据，训练目标策略的神经网络；最后将经过训练的目标策略的参数赋值给行为策略。该方案经过反复迭代学习目标策略[42]。此外，该方案还使用卷积神经网络通过卷积运算减少参数数量并降低算法的时间复杂度。以上方案的分析对比如表 5 所列。

表 5 非并行学习方案优缺点对比

Table 5 Comparison of advantages and disadvantages of non-parallel learning schemes

方案	优点	缺点
Zhang 等[38]	结构相对简单，占用计算资源较少	难以找到合适的人工神经网络更新幅度，收敛较慢
Liu 等[39]	收敛速度有所提升	使用依存策略，可能导致局部最优
Zhan 等[41]	使用脱离策略，更不容易陷入局部最优	引入了更多的人工神经网络和更多的参数

(2)并行学习方案

并行学习方案可以在一台计算机利用多线程并行来加速学习。但是，多线程的合作与数据的同步也可能会对计算资源造成额外的开销。该类方案也可以使用多台计算机进行分布式学习，以扩展一台计算机相对有限的计算能力。在分布式学习过程中，指定一台计算机将用来学习策略的人工神经网络及其参数同步给其他计算机；然后，每台计算机分别与环境互动，从而计算其神经网络的参数需要调整的部分和幅度；最后，所有计算机定时将参数调整信息汇总合并到指定的计算机来更新其神经网络参数。算法通过重复以上步骤来学习目标策略。

2019 年，Zhang 等[43]提出了并行学习的卸载方案。该方案首先通过多层人工神经网络(主网络)来表示一个全局性的策略；然后将主网络的参数和结构分别发送给多个智能代理；最后通过迭代更新主网络的参数来学习策略。在每次迭代中，首先，这些智能代理分别与环境互动，通过试错的方式学习策略，更新各自的神经网络；然后，智能代理将更新的网络参数发送给主网络进行更新；最后，主网络将已更新的网络参数发送给智能代理，使智能代理的网络参数与主网络的参数保持同步。该方案使用并行的方式提升了学习速度。

在此基础上，Feng 等[44]考虑到卸载过程中的安全问题，提出了基于信任值的卸载方案，并且该方案引入了区块链技术[45]来保证卸载的可靠性和不可逆性。在该方案中，移动设备可以通过中继节点将任务卸载到移动边缘服务器。其中，每个中继节点有一个信任值属性，代表其卸载任务时的可靠程度。为了确保通信安全，该方案选择具有高信任值的中继节点将任务传输到服务器。如果计算任务由信任值较低的中继节点传输，则中继节点可能会采取丢弃中继数据包等恶意行为。其中，中继节点信任值通过直接信任值与间接信任值共同计算得出。直接信任值通过节点的成功通信次数和失败通信次数等信息进行计算，而间接信任值通过引入区块链技术给出。该方案通过基于信任值选取可靠的节点使卸载更加安全可靠。但是，其引入区块链技术也使系统更加复杂，带来了额外的开销。以上成果的优缺点如表 6 所列。

表 6 并行学习方案优缺点对比表

Table 6 Comparison of advantages and disadvantages of parallel learning schemes

方案	优点	缺点
Zhang 等[43]	使用并行的方式加速学习	没有考虑卸载的可靠性
Feng 等[44]	使用区块链等技术使卸载更加安全可靠	为了保证安全可靠的卸载，带来额外的开销

2.2.2 确定性策略方案

虽然基于价值的深度强化学习方案也能实现确定性策略，但是，面对连续型动作，如计算功率或传输功率，这种方案不能进行精细的控制[46]，而是通过将该动作离散化为不同级别再进行决策。针对这一问题，2018 年 Chen 等[47]提出了改进方案对卸载过程中的计算功率和传输功率进行控制。该方案策略的表示方式与以上基于策略的方案不同。在以上基于策略的方案中，策略网络表示为从环境的感知状态到采取行动的概率的映射，而该方案是在得到行动的概率分布后选取其中概率最大的动作来执行[48]。此外，在执行之前，为了保持策略的探索性，需要将选取的动作增加一个随机噪声。仿真实验结果表明，该方案的性能优于基于价值的基线方法。但是，该方案存在连续型动作空间，需要较大数量的样本来探索最优策略。

2019年,Qiu等[49]针对这一问题提出了基于遗传算法[50]的确定性策略方案。该方案使用文献[39]中提到的行为网络来学习策略作出决策,并通过评判网络来对这个决策打分。在行为网络作出决策后,使用一种以概率探索的方式来探索最优策略。其探索方式如下:首先,判断更新评判网络时的误差是否大于预设的阈值。如果大于阈值,算法判断评判网络不能很好地评判动作,于是随机生成一个动作。如果小于阈值,则按以下步骤操作:首先,使用遗传算法生成动作集合;然后,使用评判网络给这些动作打分;最后,选取其中分数相对较高的一部分动作生成下一代动作集合。如此迭代,直到满足预设的迭代次数,在最后一代动作集合中选取评分最高的动作输出。此外,该方案改进了以上方案在训练人工神网络时对训练样本使用效率较低的问题。以上方案将训练数据存储在内存中,每次训练神经网络时随机选取一小批数据,而该方案将训练数据按照对神经网络更新的幅度来排序。其中,对神经网络更新幅度越大的数据越容易被选中进行网络的更新[51],从而提高了训练数据的使用效率。然而,该方案的评判网络对动作的价值估计还不够准确。

2020年,He等[52]针对这一问题提出了改进方案。该方案采用与文献[23]中对价值函数的相同处理方法,将评判网络分解为两个网络,分别表示某一状态的价值(状态价值)和这一状态下某一动作价值与所有动作平均价值的差(动作优势)[53]。算法通过这种方式可以得到当前的奖惩受动作还是受状态影响大,从而给策略打出更合理的分数。仿真实验结果表明,该方法的性能优于未将评判网络分解的方法。但是,该方案没有考虑到移动设备在作决策时可能会互相影响的问题。在移动边缘计算网络中,无线网络的信道数量有限。因此,如果多个移动设备在同一时间段内选择同一信道,同时将计算任务卸载到服务器,那么它们之间可能会产生严重的相互干扰,从而导致网络拥塞,降低网络资源的利用率。针对这一问题,Ren等[54]提出了基于软件定义网络的解决方案。该方案引入软件定义网络,通过区域内的总控制器来观察全局状态,并根据全局状态合理地安排卸载方案,防止移动设备抢占个别网络资源而导致网络拥塞。此外,该方案还根据移动设备的位置变化将任务迁移到合适的服务器进行运算。但是,该方案需要获取全局状态信息,相对比较耗时。以上成果的优缺点如表7所列。

表7 确定性策略方案优缺点对比

Table 7 Comparison of advantages and disadvantages of deterministic policy schemes

方案	优点	缺点
Chen等[47]	可以对连续型动作进行更为精细的控制	需要较多训练样本,训练样本利用率较低
Qiu等[49]	生成更多有价值的训练样本,并且提高了训练样本的利用率	评判网络对动作的评估还不够准确
He等[52]	进一步提高了评判网络对动作评估的准确性	需要训练更多的人工神经网络
Ren等[54]	缓解网络拥塞问题	需要掌握全局信息,比较耗时

3 存在的问题以及未来的研究方向

移动边缘计算中任务卸载决策问题的难点在于移动设备需要在网络拓扑和通信质量随时变化的网络环境中进行决策,而移动设备仅掌握局部信息,这为决策增加了难度。深度强化学习的方法可以根据网络环境的变化动态调整学习策略,并且不需要了解网络状态随时间变化规律的先验知识,而是仅从有限的已知状态中提取有用的信息,从而进行决策。近三年来,该方向的研究虽然已经取得了一些进展,但仍然存在一些问题,主要表现在以下几个方面。

(1)移动设备的移动性。移动设备的移动性导致网络拓扑随时变化,从而可能导致移动设备向服务器卸载任务时传输失败或服务器将结果回传给移动设备时传输失败,并且移动设备的位置变化也可能导致最优的卸载策略发生改变。

(2)可靠性与安全性。被卸载到边缘服务器的任务很容易受到外部的恶意攻击。攻击者可能会故意窃取任务信息或者篡改其数据。

(3)更多样的计算资源和能源。移动设备可以利用更多可用资源来卸载任务,如云中心或其他移动设备。这样做虽然一定程度上缓解了移动边缘服务器的负担,但是也会使网络结构变得更加复杂。此外,移动设备还可以利用能量收集技术来缓解自身能源不足的问题,但是会增加问题的复杂程度。

(4)任务的依赖性与多样性。一些任务要在其前置任务都执行结束后才能开始执行,这就要求算法在进行决策时要考虑任务的先后顺序。而不同任务对时延的要求也各不相同,这要求决策对不同的任务要进行区别处理。

(5)移动设备间互相影响。在移动边缘计算网络中,无线网络的信道数量是有限的。因此,如果多个移动设备在同一时间段内选择同一信道,并同时将计算任务卸载到服务器,那么它们之间可能会产生严重的相互干扰,导致数据传输时间延长,任务卸载效率降低。

(6)深度强化学习方法的优化。主要从两方面考虑其优化方向:一方面,如何加快策略的学习速度;另一方面,如何使学习到的策略更接近最优结果。

针对以上存在的问题,未来的研究方向可以从如下方面考虑。

(1)引入能量收集技术的任务卸载方案,利用能量采集技术来缓解移动设备能量不足的问题。

(2)考虑利用云计算和其他移动设备来分担移动边缘服务器的压力。

(3)设计更加安全可靠的卸载方案。可以运用区块链等技术使卸载更具安全性和可靠性。

(4)考虑多个移动设备的决策之间的相互影响,设计包含多个软件智能体的方案。

(5)使用迁移学习技术,利用相似环境的决策经验来加快深度强化学习方案的学习速度。

(6)使用遗传算法等技术对环境进行更有效的探索。

(7)改进多层人工神经网络结构,从而提高其对有用信息的提取能力。

(8)改进算法对自身决策行为的评估方式,使其评估更加准确,从而做出更优的决策。

结束语 本文综述了近三年来移动边缘计算中基于深度

强化学习的任务卸载研究进展。移动设备要在仅掌握局部信息的情况下，面对拓扑和通信质量随时可能改变的网络环境进行决策，使延迟和能耗达到最优。而深度强化学习的方法可以根据网络环境的变化动态调整学习策略，并且可以在掌握部分环境信息的条件下，从中提取有用信息作出近似最优的决策。本文对比分析了近年来基于深度强化学习方案的优劣。可以看到，现有的研究虽然已经取得一些进展，但仍然存在一些问题。为了推进数据保存工作的进行，提出了几个待解决的问题，并对下一步的研究方案进行了展望。

参 考 文 献

[1] YOU C, HUANG K, CHAE H, et al. Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading[J]. IEEE Transactions on Wireless Communications, 2017, 16(3): 1397-1411.

[2] JEONG S, SIMEONE O, KANG J, et al. Mobile Edge Computing via a UAV-Mounted Cloudlet: Optimization of Bit Allocation and Path Planning[J]. IEEE Transactions on Vehicular Technology, 2018, 67(3): 2049-2063.

[3] SARDELLITTI S, SCUTARI G, BARBAROSSA S, et al. Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2015, 1(2): 89-103.

[4] CHEN Y, ZHANG N, ZHANG Y, et al. TOFFEE: Task Offloading and Frequency Scaling for Energy Efficiency of Mobile Devices in Mobile Edge Computing[J]. IEEE Transactions on Cloud Computing, 2019(99): 1-1.

[5] REN J, YU G, CAI Y, et al. Latency Optimization for Resource Allocation in Mobile-Edge Computation Offloading[J]. IEEE Transactions on Wireless Communications, 2018, 17(8): 5506-5519.

[6] TALEB T, SAMDANIS K, MADA B, et al. On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration[J]. IEEE Communications Surveys and Tutorials, 2017, 19(3): 1657-1681.

[7] 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.

[8] PAPADIMITRIOU C H, TSITSIKLIS J N. The complexity of Markov decision processes[J]. Mathematics of Operations Research, 1987, 12(3): 441-450.

[9] CHEN Y, ZHANG N, ZHANG Y, et al. Energy efficient dynamic offloading in mobile edge computing for internet of things [J/OL]. IEEE Transactions on Cloud Computing, 2019. http://www.semanticscholar.org/paper/Energy-Efficient-Dynamic-Offloading-in-Mobile-Edge-Chen-Zhang/fbe4cb7777cdd2485d1e5fb0072c896b045027fc.

[10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[11] 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.

[12] HE Y, ZHAO N, YIN H, et al. Integrated Networking, Caching, and Computing for Connected Vehicles: A Deep Reinforcement Learning Approach[J]. IEEE Transactions on Vehicular Technology, 2018, 67(1): 44-55.

[13] WANG C, LIANG C, YU F R, et al. Computation Offloading and Resource Allocation in Wireless Cellular Networks With Mobile Edge Computing[J]. IEEE Transactions on Wireless Communications, 2017, 16(8): 4924-4938.

[14] ZHOU Y, YU F R, CHEN J, et al. Resource Allocation for Information-Centric Virtualized Heterogeneous Networks With In-Network Caching and Mobile Edge Computing[J]. IEEE Transactions on Vehicular Technology, 2017, 66(12): 11339-11351.

[15] YOU C, HUANG K, CHAE H, et al. Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading[J]. IEEE Transactions on Wireless Communications, 2017, 16(3): 1397-1411.

[16] LYU X, TIAN H, NI W, et al. Energy-Efficient Admission of Delay-Sensitive Tasks for Mobile Edge Computing[J]. IEEE Transactions on Communications, 2018, 66(6): 2603-2616.

[17] ZHAO P, TIAN H, FAN S, et al. Information Prediction and Dynamic Programming-Based RAN Slicing for Mobile Edge Computing[J]. IEEE Wireless Communications Letters, 2018, 7(4): 614-617.

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

[19] HUANG L, BI S, ZHANG Y A, et al. Deep Reinforcement Learning for Online Computation Offloading in Wireless Powered Mobile-Edge Computing Networks[J/OL]. http://arXiv.org/abs/1808.01977v6.

[20] CHEN X, ZHANG H, WU C, et al. Optimized Computation Offloading Performance in Virtual Edge Computing Systems Via Deep Reinforcement Learning[J]. IEEE Internet of Things Journal, 2019, 6(3): 4005-4018.

[21] HUANG L, FENG X, ZHANG C, et al. Deep reinforcement learning-based joint task offloading and bandwidth allocation for multi-user mobile edge computing[J]. Digital Communications and Networks, 2019, 5(1): 10-17.

[22] YAO P, CHEN X, CHEN Y, et al. Deep reinforcement learning based offloading scheme for mobile edge computing[C]// 2019 IEEE International Conference on Smart Internet of Things (SmartIoT). IEEE, 2019: 417-421.

[23] HE Y, YU F R, ZHAO N, et al. Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach[J]. IEEE Communications Magazine, 2017, 55(12): 31-37.

[24] CHEN M, LIANG B, DONG M, et al. Joint offloading decision and resource allocation for multi-user multi-task mobile cloud [C]// International Conference on Communications. 2016: 1-6.

[25] VAN HASSELT H, GUEZ A, SILVER D, et al. Deep reinforcement learning with double Q-Learning[C]// National Conference On Artificial Intelligence. 2016: 2094-2100.

[26] 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.

[27] NING Z, DONG P, WANG X, et al. Deep reinforcement learning

for vehicular edge computing: An intelligent offloading system [J]. ACM Transactions on Intelligent Systems and Technology (TIST),2019,10(6):1-24.

[28] LU H,GU C,LUO F,et al. Optimization of lightweight task offloading strategy for mobile edge computing based on deep reinforcement learning[J]. Future Generation Computer Systems, 2020,102:847-861.

[29] PAN S J,YANG Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010,22(10): 1345-1359.

[30] SRIVASTAVA N,HINTON G E,KRIZHEVSKY A,et al. Dropout:a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[31] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[32] GUPTA H,DASTJERDI A V,GHOSH S K,et al. iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments [J]. Software-Practice and Experience, 2017, 47(9):1275-1296.

[33] HUANG B,LI Y,LI Z,et al. Security and Cost-Aware Computation Offloading via Deep Reinforcement Learning in Mobile Edge Computing [J]. Wireless Communications and Mobile Computing,2019(2019):1-20.

[34] ZHANG K,ZHU Y,LENG S,et al. Deep Learning Empowered Task Offloading for Mobile Edge Computing in Urban Informatics[J]. IEEE Internet of Things Journal, 2019, 6(5): 7635-7647.

[35] XU X,ZHANG X,GAO H,et al. BeCome: Blockchain-Enabled Computation Offloading for IoT in Mobile Edge Computing[J]. IEEE Transactions on Industrial Informatics,2020,16(6):4187-4195.

[36] MAURICE N,PHAM Q V,HWANG W J. Online Computation Offloading in NOMA-based Multi-Access Edge Computing: A Deep Reinforcement Learning Approach[J]. IEEE Access,2020 (99):1-1.

[37] ALFAKIH T,HASSAN M M,GUMAEI A,et al. Task Offloading and Resource Allocation for Mobile Edge Computing by Deep Reinforcement Learning Based on SARSA[J]. IEEE Access,2020:8:54074-54084.

[38] ZHANG H,WU W,WANG C,et al. Deep Reinforcement Learning-Based Offloading Decision Optimization in Mobile Edge Computing[C]// Wireless Communications and Networking Conference. 2019:1-7.

[39] LIU Y,CUI Q,ZHANG J,et al. An Actor-Critic Deep Reinforcement Learning Based Computation Offloading for Three-Tier Mobile Computing Networks[C]// International Conference on Wireless Communications and Signal Processing. 2019: 1-6.

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

[41] ZHAN W,LUO C,WANG J,et al. Deep Reinforcement Learning-Based Offloading Scheduling for Vehicular Edge Computing [J]. IEEE Internet of Things Journal,2020:7(6):5449-5465.

[42] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal Policy Optimization Algorithms[J]. arXiv:1707.06347,2017.

[43] ZHANG T,CHIANG Y,BORCEA C,et al. Learning-Based Offloading of Tasks with Diverse Delay Sensitivities for Mobile Edge Computing[C]// Global Communications Conference. 2019.

[44] FENG J,YU F R,PEI Q,et al. Cooperative Computation Offloading and Resource Allocation for Blockchain-Enabled Mobile Edge Computing: A Deep Reinforcement Learning Approach [J]. IEEE Internet of Things Journal,2019:1-1.

[45] XIONG Z,ZHANG Y,NIYATO D,et al. When Mobile Blockchain Meets Edge Computing[J]. IEEE Communications Magazine,2018,56(8):33-39.

[46] LILLICRAP T,HUNT J J,PRITZEL A,et al. Continuous control with deep reinforcement learning[C]// International Conference on Learning Representations. 2016.

[47] CHEN Z,WANG X D. Decentralized Computation Offloading for Multi-User Mobile Edge Computing: A Deep Reinforcement Learning Approach[J]. arXiv:1812.07394,2018.

[48] SILVER D,LEVER G,HEESS N,et al. Deterministic Policy Gradient Algorithms[C]// International Conference on Machine Learning. 2014:387-395.

[49] QIU X,LIU L,CHEN W,et al. Online Deep Reinforcement Learning for Computation Offloading in Blockchain-Empowered Mobile Edge Computing[J]. IEEE Transactions on Vehicular Technology,2019,68(8):8050-8062.

[50] SRINIVAS M,PATNAIK L M. Adaptive probabilities of crossover and mutation in genetic algorithms[J]. IEEE Transactions on Systems,Man,and Cybernetics,2002,24(4):656-667.

[51] SCHAUL T,QUAN J,ANTONOGLOU I,et al. Prioritized Experience Replay[J/OL]. http://arXiv:org/bas/1511.05952v4.

[52] HE X M,LU H D,HUANG H W,et al. QoE-Based Cooperative Task Offloading with Deep Reinforcement Learning in Mobile Edge Networks [J]. IEEE Wireless Communications, 2020, 27(3):111-117.

[53] VAN HUYNH N,HOANG D T,NGUYEN D N,et al. Optimal and Fast Real-Time Resource Slicing With Deep Dueling Neural Networks[J]. IEEE Journal on Selected Areas in Communications,2019,37(6):1455-1470.

[54] REN Y L,YU X M,CHEN X Y,et al. Vehicular Network Edge Intelligent Management: A Deep Deterministic Policy Gradient Approach for Service Offloading Decision[C]// IWCMC. 2020: 905-910.

LIANG Jun-bin, born in 1979, Ph. D, professor, Ph. D supervisor. His main research interests include wireless sensor networks, network deployment and optimization.

ZHANG Hai-han, born in 1993, postgraduate. His main research include wireless sensor networks and artificial intelligence.