计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000008

基于强化学习的高能效基站动态调度方法

曾德泽[1]　李跃鹏[1]　赵宇阳[1]　顾　琳[2]
1 中国地质大学(武汉)计算机学院　武汉 430074
2 华中科技大学计算机科学与技术学院　武汉 430074
(deze@cug.edu.cn)

摘　要　随着移动通信技术的升级与移动通信产业的兴起,移动互联网正蓬勃发展。然而,由于移动设备爆发式增长,网络规模不断扩大和用户对服务质量的要求的不断提高,移动互联网络正面临着下一场技术革命。虽然5G技术可以通过密集的网络部署来实现千百倍的网络性能提升,但同信道干扰和高突发性的用户请求等问题使得该方案下需要消耗巨大的能量。为了在5G网络中提供高性能服务,升级改进现有网络管理方案势在必行。针对这些问题,使用带缓存队列的短周期管理框架实现对请求突发场景的敏捷平滑管理,避免由突发性请求导致的服务质量剧烈波动。此外,采用深度强化学习方法对用户分布、通信需求等进行自我学习,从而推测出基站的负载变化规律,进而实现对能量的预调度和预分配,在保证服务质量的同时提高能量的利用率。文中提出的双缓冲DQN算法在收敛速度上比传统DQN算法提高了近20%,且与当前广泛使用的基站常开策略相比,该算法能够节约4.8%的能量消耗。

关键词:移动网络管理;基站休眠;异构网络;双缓冲区;请求突发;深度强化学习
中图法分类号　TP391

Reinforcement Learning Based Dynamic Basestation Orchestration for High Energy Efficiency

ZENG De-ze[1],LI Yue-peng[1],ZHAO Yu-yang[1] and GU Lin[2]
1 School of Computer Science,China University of Geosciences,Wuhan 430074,China
2 School of Computer Science and Technology,Huazhong University of Science and Technology,Wuhan 430074,China

Abstract　The mutual promotion of mobile communication technology and mobile communication industry has achieved unprecedented prosperity in the mobile Internet era. The explosion of mobile devices,expansion of the network scale,improvement of service requirements are driving the next technological revolution in wireless networks. 5G meets the requirements for the thousand-fold improvement of service performance through intensive network deployment,but co-channel interference and bursty request problems make the energy consumption of this solution very huge. In order to support 5G network to provide energy-efficient and high-performance services,it is imperative to upgrade and improve the management scheme of mobile networks. In this article,we use a short-cycle management framework with cache queues to achieve agile and smooth management of request burst scenarios to avoid dramatic fluctuations in service quality due to request bursts. We use deep reinforcement learning to learn the user distribution and communication needs,and infer the load change rules of the base station,and then realize the pre-scheduling and pre-allocation of energy,while ensuring the quality of service and improving the energy efficiency. Compared with the classic DQN algorithm,the two-buffer DQN algorithm proposed in this paper can provide nearly 20% acceleration in convergence. In terms of decision performance,it can save 4.8% energy consumption compared to the currently widely used keep on strategy.

Keywords　Mobile network management,Base station sleep,Heterogeneous network,Double buffer,Request burst,Deep reinforcement learning

1　引言

随着移动互联网时代的到来,现有的通信网络面临着1000倍的移动流量增长、10～100倍的无线设备连接、10～100倍的用户速率需求、更低的延迟及更高的可靠性要求等巨大的挑战。由于用户数量呈指数级增加和无线网络规模的不断扩大,百倍甚至千倍级的服务指标提升迫使传统的无线网络进行一场技术革新。

到稿日期:2020-10-03　返修日期:2021-04-10
基金项目:国家自然基金(61772480,61972171,62073300);之江实验室开放课题项目基金(2021KE0AB02)
This work was supported by the National Natural Science Foundation of China(61772480,61972171,62073300) and Open Research Projects of Zhijiang Lab(2021KE0AB02).
通信作者:顾琳(anheeno@gmail.com)

在现今的移动网络中，能源消耗占据了运营成本的大部分比例。统计结果显示，每年全球的通信产业消耗了 1 500 TWh 的能源，接近全世界发电量的 10%[1]。此外，基站的负载在峰值负载的 10%以下的时间超过 30%，而基站空载时也要消耗 60%～80%的电能，这表明当前基站的能量利用率存在很大的优化空间。所以研究节能高效的基站能量管理方案是减少运营支出、提高利润的重要手段。在基站能效管理问题中，用户请求的突发问题导致基站的能效管理规律难寻，无法使用固定的策略；而基站间的同信道干扰直接导致了发射能量的白白消耗，这些都是提升基站服务能力的主要障碍[2-3]。到了 5G 时代，基站的部署更为密集，用户数量更多，流量更大，其突发性和同信道干扰问题将更加严重。而且不合理的基站管理决策可能还会加重基站的能量消耗问题，因此一种先进的节能的基站能量管理方案显得十分必要。

然而在现有的研究中，部分工作仅考虑了异构蜂窝网络中的同信道干扰问题，而忽略了用户请求的突发性；另一些研究工作虽然考虑了用户请求突发情况，但是决策周期较长、灵活性较差，没有有效地保障用户的服务体验。此外，当前移动网络的管理算法和策略大多是基于既定规则或者数学模型，在真实环境中使用往往存在着很多限制[4-6]。由于物理网络环境非常复杂，因此很难准确建模，且现有的工作中存在大量的假设和选择性的忽略[7-9]。而数据建模中计算得到的数据与观测到的真实数据往往有很大的差异，比如请求量和时延等。再则现有的算法大都是基于固定的规则，无法应对时刻变化的网络状态，也无法充分利用网络中历史数据的特征和规律。

为了使算法能够适应更加复杂的真实环境，需要寻求一种不完全依赖数学模型的方法来适应网络环境和用户需求的变化，从而更好地管理移动网络。近些年来，以强化学习为首的人工智能技术为解决更复杂的控制问题带来了希望，并已经被广泛应用于网络的各种领域。为了应对 5G 网络的高动态性，减少对问题的不必要假设，本文采用深度强化学习对用户分布、通信需求等进行自我学习，并推断出基站下一时刻的负载量的变化，进而为基站的能量进行预调度和预分配，在保证服务质量的同时提高能量的利用率。5G 网络基站的覆盖范围更小，流量突发性更强，使用较短的决策周期便能够对网络进行更精细的控制，在保障用户体验的同时，进一步提高能效。此外，短决策周期还可以通过为各基站设置一个缓存队列来提高对于突发情形的适应性，当无法满足用户的需求时，将请求放入队列，等到下一回合再处理。另外，这种方式可以根据用户请求的爆发程度以及用户请求对时延的敏感程度来调节基站的队列大小，实现能效与服务质量之间的权衡。

本文的主要贡献如下：

(1)针对异构突发蜂窝网络的小基站休眠管理问题进行了全面的建模，使用短周期的敏捷管理方案，并采用 DQN 算法进行基站休眠控制策略学习。

(2)为了让学习模型更好地适应本问题的突发性特征，对经典 DQN 算法进行改进，并使用双缓冲队列优化 DQN 算法的经验回放过程，从而加快算法的收敛速度，提升基站休眠决策质量。

(3)设计仿真实验，并使用 3 种算法与本文算法进行性能对比。实验结果显示，本文算法的性能明显优于其他 3 种算法，能够很好地适应请求突发状况，快速稳定地收敛并进行优秀的休眠决策。

本文第 2 节主要介绍了本文研究相关的工作，即当前研究如何从时间和空间两方面来提高基站能效；第 3 节对 5G 异构网络基站能效提升问题进行了详细建模；第 4 节根据问题要求设计和改进了深度强化学习算法；第 5 节对实验结果进行了全方位的对比分析；最后总结全文并展望未来。

2 相关工作

在网络管理中，由于用户的移动性，移动网络的用户需求在时间和空间上都是不断变化的。无线通信业务在时间和空间上呈现不均匀性，导致基站能量利用效率提升困难。当前的工作主要从时间和空间两个角度寻求节省网络能耗的方法。

2.1 从时间上进行节能管理

无线通信业务在时间上具有不均匀性，即基站覆盖区域内用户的业务需求量会随着时间发生波动，存在峰值和低谷。在网络设计时，为了保障用户体验，需要依据最大请求量设计网络的峰值容量，但实际运营时，用户请求量并不会时刻达到峰值。有统计数据表明，在一天之中有超过 30%的时间，用户请求量不超过基站峰值容量的 1/10。针对这一现象，我们可以通过适当地调整基站的发射功率或者采用适当的基站休眠策略来减少能耗。

文献[10-15]研究了同构网络中的基站高效管理问题，其中 Wu 等[10-11]为了解决用户请求量随着时间的突变导致的基站能效较低问题，采用了一种名为“N-policy”的决策方案来决定基站的开关。其中文献[10]研究了用户请求数、请求突发机率以及基站的发射功率之间的关系，得出了当请求的突发性越高时，“N-policy”决策方案能提供更高的能量效率的结论。文献[11]在前者的基础上增加了对用户体验的考虑，着重研究了基站的发射功率以及用户体验时延之间的权衡。Liu 等[12]为了提高用户请求数随着时间突变的场景的基站能效，使用基于深度强化学习改进的“DeepNap”算法来决策基站的开关。该文献着重考虑了用户请求的突发性，对其强化学习算法的奖励机制和记忆库进行了相应的改进，使得它不仅能够应对间断泊松过程产生的突发用户请求，也能对复杂的真实用户请求数据集做出令人满意的决策，增强了算法的实用性和稳定性。Wang 等[14]面向协同多点下行效能优化，提出了基于 Q 学习的区域呼吸策略。此外，从基站开关的角度而言，文献[15]针对异构超密集网络中宏基站覆盖的小型基站之间的不协调操作导致的功耗和频率复用问题，综合小基站睡眠和频谱分配策略，从而在最大化覆盖概率的同时最小化基站功耗。

Ghadimi 等[13]通过调节基站发射功率来提高基站的能量效率。其主要思路是利用深度强化学习算法学习用户请求的产生规律，进而决策如何调整基站的发射功率，从而在保障用户体验的前提下，显著提高了基站的能量使用效率。而文献[10-12]考虑了用户请求突发的情况，不同于文献[13]及其他

大多数文献将用户的请求发生假设为泊松过程（Poisson Process），其使用间断泊松过程（IPP）来实现有突发的用户请求产生模型。

2.2 从空间上进行节能管理

不同区域的网络请求量是不同的，一般情况下业务量与用户的分布呈正比例关系，这就是无线通信业务在空间上的不均匀。若使用统一的带宽来覆盖所有区域，必将导致高需求区域的服务质量欠佳，低业务量区域的资源严重浪费。为了解决这一问题，业界的通用做法是部署异构蜂窝网络，从空间上提高基站能效。

异构网络主要由具有较高发射功率的宏基站（Macrocell）与规格较小的小基站组成。宏基站负责提供全区域覆盖，小基站主要起着信号盲区覆盖以及局部容量提升的作用，主要部署在宏基站覆盖区域边缘以及用户请求热点区域（hotspot）。部署异构网络并不是没有缺陷的，虽然小基站的发射功率不高，但由于其距离更近且一般与宏基站共用频谱，因此同信道干扰问题在异构网络中不容忽视，小基站密集部署的 5G 网络更是如此。不合理的部署和管理异构网络，有时不仅不能提高网络容量，反而可能白白浪费了资源，甚至可能导致容量下降。加上用户请求在时域上的变化，异构网络中的基站管理问题很有研究的必要。

文献[16-21]研究了异构蜂窝网络中的基站高能效管理问题。Niu 等[16]重点研究了基站休眠问题中的能效与延迟的权衡，它提出了一个称为迟滞休眠的等待期，以在实施休眠策略的同时保障系统的稳定性。迟滞休眠指基站在休眠之前必须持续开启一段时间或完成特定数量的任务。此外文献[17-19]采用博弈论的思想来决策小基站的休眠问题。Feng 等[18]以最大化基站能效为目的，将异构网络中的基站休眠和用户连接问题构建成一个整数规划问题，通过一定的松弛，利用拉格朗日对偶方法将该问题转化为一个标准的线性规划问题。为了解决这一问题，他们采用用户和基站竞价博弈的方法，并证明该博弈收敛于纳什均衡（Nash Equilibrium）。Samarakoon 等[19]将两层异构网络的高能效基站开关决策建模成一个基站间的非合作博弈问题，并采用分布式的学习算法来决策基站开关，使其在满足用户 QoS 需求的前提下，最小化能量消耗。值得一提的是，这个分布式算法还考虑了基站间的负载均衡问题。Chen 等[17]使用强化学习算法学习最佳的基站开关决策，并将异构蜂窝网络的小基站开关决策问题建模成一个离散时间马尔可夫决策过程（DTMDP）。考虑到状态空间的规模，他们对 Q-learning 算法进行了改进，提出了 QC-learning 算法以应对维度灾难问题（Curse of Dimensionality）。并且他们还提出了 QC-learning 算法的分布式版本，在保证决策性能的基础上，加快了收敛的速度。无独有偶，文献[20]也基于 Q-learning 算法来对基站的激活和睡眠进行管理，从而在保障服务质量的同时最小化能量开销。此外，针对现有基于传统 Q-learning 的基站管理算法中智能体过多难以收敛的问题，文献[21]提出了基于分布式 Q-learning 的基站管理策略，通过将每个智能体进行单独训练，可以有效加快训练时的收敛速度。Zhang[22]面向云无线接入网，提出了一个基于深度强化学习的优化框架来解决资源分配问题。

综上所述，现有的研究方案要么没有考虑关闭基站对于服务质量的影响，要么对某些条件进行了简化或假设，导致其在实际场景中的实用性有限。因此，在异构 5G 移动网络中，在保障用户体验的前提下，有效地提高基站能效，并加强设计方案的实用性是一个还需深入研究的问题。

3 系统建模

本节将从异构无线网络场景模型、用户请求到达模型、网络管理架构以及研究目标等方面来描述基站成本问题。

3.1 异构网络模型

如图 1 所示，本文研究的场景是一个由若干宏基站和小基站组成的两层异构蜂窝网络。所有宏基站确保了全区域信号覆盖，且各宏基站的覆盖范围互不重叠。每个宏基站内部署了若干个小基站，它们与对应的宏基站通过逻辑接口相连。服务区域被划分为一系列的网格，网格的覆盖范围很小，是本问题的基本管理单元（而不是单个用户）。我们将网格作为一个整体来考虑，网格内的用户具有相同的信道状态，即网格内所有用户的传输带宽一样。

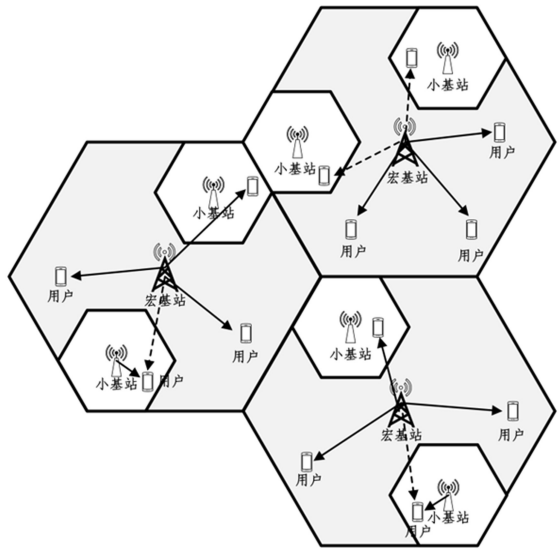


图 1 异构蜂窝网络

Fig. 1 Heterogeneous cellular network

本文重点关注在蜂窝网络的下行数据链路，即基站响应用户请求，将数据下发给对应用户。宏基站始终保持开启，保证全区域、全时间段内用户的请求都会被处理。如果小基站处于激活状态，则在其覆盖范围内的用户都只会连接到小基站上并向其请求数据。未被小基站覆盖或小基站处于休眠状态的用户将连接到对应的宏基站上。为了专注于同信道干扰问题，我们隐去了信道分配方面的内容，假定所有基站共享同一段频谱。激活小基站能够为对应的宏基站分担一部分的用户请求，以降低宏基站用户之间的带宽竞争压力。但是由于共享频谱，激活小基站也会给宏基站和其他小基站带来新的干扰，造成其信号与干扰加噪声比的下降。而且小基站的开启不是瞬时的，需要消耗一定的时间。因此，对于小基站的开启和关闭，都需要进行慎重的考虑。

3.2 用户请求突发到达模型

由于蜂窝网络的用户数量庞大，而且网格的覆盖范围较

小，基站管理研究一般不拘泥于单个用户是否会产生请求，而是将一个网格看成一个整体，并使用网格的用户请求数来描述网络状态。由于真实的网络数据流量具有自相似性和突发性，使得它远比泊松过程复杂。在5G网络中，由于基站的覆盖范围更小、用户请求的资源更大、网络带宽更高等因素，使得网络数据流量的突发性更加明显。间断泊松过程和马尔可夫调制泊松过程(MMPP)由于能够较好地表征网络流量的突发性，且具有易于分析的特点，在各种类型的多媒体流量建模中应用广泛，因此我们使用双状态的 MMPP 模型来描述用户请求的到达。两个状态分别表征用户请求突发和平时的网络请求情况，它们的持续时间均满足指数分布。通过调节参数，我们能够模拟出平均请求量、突发强度以及突发频率不同的网络场景。另外，本文使用指数分布来描述每个用户请求要求的数据量大小。

在现实场景中，小基站一般部署在用户请求较多的热点区域来提升局部区域的服务能力。因此，在本文中，小基站覆盖的网格的用户请求到达率均高于未被小基站覆盖的网格，这使得小基站覆盖区域的突发性更强，增加了本模型的真实性。

3.3 网络管理架构

本文使用一个时隙管理框架，整个时间被切分为一系列的时间片，时间片内网络状态不会发生改变，宏基站作为基站休眠控制器，集中式地决策所属小基站的工作状态。在每个时隙的开始阶段，宏基站通过逻辑接口知晓整个网络状态，包括本时间片内各网格的数据请求量、前一时隙各小基站状态等，然后依据网络状态决策当前时间片小基站的工作状态。

基站休眠决策研究的决策周期(即时间片)长短不一，较长的决策周期的时间跨度从几分钟到数小时不等，而较短的决策周期一般是数秒到几分钟不等。较长的决策周期能够减少小基站的状态变化，但由于其时间片较长，会出现将突发请求挪移拖延到后续请求较少时刻以减少基站开启的现象，容易导致瞬时服务质量骤降，对于网络动态变化的适应性欠佳，而且要预先知道未来数小时的用户请求情况是不现实的。短决策周期能够抓住细粒度的请求量变化，进行敏捷的基站管理。但它难以把握网络状态的长期趋势，容易频繁地改变小基站状态。而且对于短决策周期来说，决策模型的时间复杂度也是一个必须考虑的问题。5G网络基站数目更多、请求突发性更强、服务要求更高，使用较短的决策周期能够对网络进行更精细的控制，在保障用户体验的同时，进一步提高基站能效。而较长的决策周期无法察觉出短期的用户请求突发和QoS变化，无法做出敏捷的基站决策。通过综合考虑5G网络的请求状况和服务要求，以及长短决策周期各自的利弊，本文选用较短的决策周期决策基站开关。

较长的决策周期都存在一个潜在假设，基站能够在当前决策周期内处理完所有的用户请求。但对于短决策周期来说，无法在一个决策周期内处理完所有的用户请求是常有的事情(特别是流量突发较严重的场景或小基站休眠时)。为了解决这一问题，各个基站内都设有一个缓存队列，当用户的需求无法满足时，将请求放入队列，等到下一回合再处理。这一方式可以根据用户请求的爆发程度以及用户请求对时延的敏感程度，调节基站的队列大小，从而实现能效与服务质量之间的权衡。

基站以切分时间片的方式，公平地为辖区内的网格提供服务。也就是说，当基站无法完成全部请求时，会等比例地为所有网格完成部分任务。当小基站状态变化时，缓存队列中积压的用户请求会在对应的宏基站和小基站之间传递。也就是说，当小基站开启时，它会接管对应宏基站在它的覆盖范围内积压的请求，而当小基站关闭时，则会将队列中的用户请求转移给对应宏基站。当基站未能完成全部用户请求而产生积压时，基于积压的程度会对当前回合的基站开关决策的服务质量评价造成一定影响，但在后续回合处理积压请求时，积压的请求和该回合产生的请求没有区别，应同等对待。另外，当用户请求实在过多，超出了基站缓存队列的上限时，会丢弃无力处理的部分请求，并在基站开关决策的服务质量评价中进行体现。

3.4 研究目标

直观上来说，基站开关的时机就是：当基站的系统负载较低时，关闭小基站，节约能量；当基站的负载较重时，酌情开启小基站，保证服务质量。本问题的研究目标是在保证用户的服务质量的前提下，尽量地关闭小基站，降低总体的基站能量消耗，提高系统能效。要实现这一目标，有两个问题需要明确：1)基站的能量消耗计算；2)用户服务质量评价指标。

基站的能量消耗主要由两部分组成：运行能耗(即基站的信号处理单元以及冷却设备等的能量消耗)以及传输能耗(即基站传输用户请求消耗的能量)。运行能耗在基站总能耗中占很大比重，尤其是在宏基站中。而且运行能耗与基站负载近乎无关，基站满载和空载时的运行能耗非常相近。传输能耗与基站的负载相关，基站负载越高，传输能量消耗越大。

在评判移动网络的服务质量时，一般采用用户侧体验到的时延或者吞吐量来衡量。然而在本课题中，用户数量庞大，而且基站请求缓存队列的引入将请求处理情况大大复杂化，时延或吞吐量的描述方式均不适用于本问题。考虑到基站的负载情况可以从侧面反映出用户的服务等待时间，这也是衡量服务质量的重要标准之一，而且在本问题中，用户请求缓存队列的长度以及由于请求过多而丢弃的用户请求数也是服务质量评价的重要指标，因此本问题的服务质量评判原则要综合考虑以上三大因素。

基站的能效管理是一个整体的、长期的过程，本研究的终极目标是找到一个基站开关策略 $\boldsymbol{\Phi}:\boldsymbol{X}\rightarrow\boldsymbol{Y}$，在任意网络状态 $\boldsymbol{x}(t)$ 下，为所有的小基站决策合适的工作状态 $\boldsymbol{y}(t)$，在满足用户 QoS 要求的前提下，最小化网络的总能量开销。我们将这一寻求最优策略的优化问题定义为：

$$\min_{\boldsymbol{\Phi}\in\Omega} Y(\Phi) \tag{1}$$

$$\text{s.t.}\ \Theta(x(t),y(t))\leqslant\Theta^{\text{th}},\forall t\in T \tag{2}$$

其中，$\boldsymbol{\Omega}$ 是所有可选的基站开关策略组成的集合，$f(\Phi)$ 表示在策略 $\boldsymbol{\Phi}$ 下的长期网络能量消耗期望。每一个无线网络都有其服务质量要求，Θ^{th} 表示网络中用户能够忍受的服务质量评判指标的阈值。

4 算法设计

鉴于本问题高维度的复杂性，以及离散的动作空间，本文

选用 DQN 算法来进行最佳的基站开关决策学习。根据问题特点，本节对 DQN 算法进行针对性的改进，从而加快算法的收敛，提高决策算法的性能。

4.1 强化学习框架设计

基于前面对系统模型的详细描述，我们能够看出基站高能效管理问题是一个序贯决策过程：基站休眠控制器根据当前的网络状态（用户请求量、缓存队列长度、小基站开关状态）来决定小基站的开关动作，直接影响信道干扰状况以及网络负载分配，实现对全区域的用户 QoS 以及基站能耗的管控。而当前回合的小基站开关动作以及缓存队列长度又将成为网络状态的一部分并传递到下一回合，进而影响后续的小基站开关决策。因此，任意两个连续的基站开关动作都是相关联的，基站开关决策应考虑其对基站能耗的长期影响。

强化学习是一个应用广泛的解决复杂序贯决策问题的框架，利用奖励函数指导智能体在不断的尝试中学出最佳的策略。它能够从长远的角度考虑小基站开关问题，在一定程度上解决短决策周期带来的对于长期网络状态不明确的问题，有效减少不必要的基站开关操作。而且不同于启发式算法，其一旦训练完成，不仅决策时间很短，且还能够持续学习、持续改进，适应网络状态变化规律的改变。鉴于本问题高维度的复杂性，以及离散的动作空间，对比不同的强化学习算法的特性，本文选用 DQN 算法来进行最佳的基站开关决策学习。

使用强化学习算法进行基站管理时，首先需要构建强化学习算法框架，即强化学习的状态、动作和奖励。本问题的算法框架定义如下：

状态：t 时刻的网络状态 $\boldsymbol{s}(t)$ 是由当前时刻各个小基站下属的各个网格的用户请求情况以及上一时刻的小基站的开关状态组成的向量。

动作：t 时刻的动作 $\boldsymbol{a}(t)$ 表示所有小基站在该时刻的开关状态。

奖励：状态 $\boldsymbol{s}$ 下采取动作 $\boldsymbol{a}$ 所得到的即时奖励，用于判断当前基站开关决策的好坏程度。为了准确地衡量基站开关决策的好坏程度，我们需要综合考虑无线网络传输中的收益和开销。现有的研究工作大多把蜂窝网络总能耗作为开销，用网络传输的总数据量来衡量收益。此外，我们还应考虑放入基站缓存队列中的用户请求以及由于基站负载过大丢弃的请求带来的服务质量下降问题，因此本文的奖励函数如式（3）所示：

$$R(\boldsymbol{s},\boldsymbol{a})=\frac{\sum_{i\in J}U_i(\boldsymbol{s},\boldsymbol{a})\beta_1+P_i(\boldsymbol{s},\boldsymbol{a})\beta_2+Q_i(\boldsymbol{s},\boldsymbol{a})\beta_3}{E(\boldsymbol{s},\boldsymbol{a})} \tag{3}$$

其中，$E(\boldsymbol{s},\boldsymbol{a})$ 和 $U_i(\boldsymbol{s},\boldsymbol{a})$ 分别表示在网络状态为 $\boldsymbol{s}$，基站开关决策为 $\boldsymbol{a}$ 时，蜂窝网络的总能量开销以及基站 $i\in J$ 处理完成的用户请求数据量。$P_i(\boldsymbol{s},\boldsymbol{a})$ 和 $Q_i(\boldsymbol{s},\boldsymbol{a})$ 分别表示在网络状态为 $\boldsymbol{s}$、基站开关决策为 $\boldsymbol{a}$ 时，基站 $i\in J$ 的缓存队列中的用户请求数据量以及负载过大丢弃的请求数据量。β_1，β_2，β_3 分别表示 3 个元素的加权参数，β_1 大于 0，而 β_2 和 β_3 小于 0，可依据网络的不同管理要求进行相应的调节。

4.2 DQN 算法改进

直接使用经典的 DQN 算法无法有效地解决场景中面临的突发性请求的问题，此外在决策能力以及收敛速度方面，传统的 DQN 算法还存在严重问题，所以需要根据问题的特性对其进行一定的改进。

4.2.1 状态空间压缩

本问题的状态是一个由各网格的用户请求数据量以及各小基站的开关状态组成的向量。用户请求数据量本身具有很大的变化范围。此外，合并临近用户组成网格的管理模式虽然降低了本问题的状态空间维度，但用户的聚合增大了状态空间中元素的变化范围，而且任务突发时刻的数据量更是平时的好几倍，这会使得本问题中状态空间中的元素值跨度非常大，从而增大状态空间，进而导致 DQN 算法难以收敛。

深入考虑本问题的场景可以发现，基站的带宽也是比较大的，数据量间的细微差别无法左右小基站的开关。因此，我们可以再次使用聚合的方法，通过对网格的请求数据量划分等级来缩小其变化范围。

这种处理方式的难点在于，如果聚合程度过大，不容易获得状态的特征而导致系统的控制精度降低，反之如果聚合程度过小，又不能有效地缩小状态空间范围。考虑到请求数据量终究与基站的带宽相关，经过仔细的设计，我们使用无干扰情况下的宏基站覆盖范围一半的位置的带宽作为分级单位，并将请求数据量除以该带宽的商保留到小数点后两位作为新的状态值，在不降低算法性能的前提下，加快 DQN 算法的收敛。

4.2.2 经验回放机制的改进

DQN 算法中，经验回放的过程是：先将每一步的状态转移过程以 $(S(t),A(t),R(t),S(t+1))$ 的格式存储在记忆库中，当要更新神经网络参数时就从记忆库中随机抽取一批记忆进行训练，并计算损失度来训练 DQN 网络。记忆库的容量是有限的，当样本数量超过记忆库的容量时，一般用“先进先出”的原则保证其中的记忆都是最近的数据。

从全局来看，请求突发期占全时间区间的比例较小，因此记忆库中的请求突发记忆所占的比例很小。传统 DQN 一般从记忆库随机抽取记忆，这种方式使得训练数据中请求突发记忆出现的几率较小，神经网络对于相关状态的训练也较少，导致决策性能不佳。然而，请求突发期的决策会在很大程度上影响整体的服务质量，因此，我们需要设法增大训练数据中请求突发记忆的比重，从而增加请求突发记忆对神经网络的训练。

从短期来看，本问题采用的决策周期很短，记忆库中的记忆替换很快，请求突发期和常规时期记忆库中的两类记忆占比不同：请求突发期时，记忆库中的请求突发记忆的比例会明显上升。由此导致周期交替时神经网络的损失度变化剧烈，神经网络参数来回变化，DQN 算法不易收敛，而且周期交替时的服务质量也无法保证。为了提高 DQN 算法决策性能的稳定性，加速神经网络的收敛，我们需要对经验回放机制进行一定的改进。

增大记忆库容量的方法虽然能减小周期交替带来的记忆占比剧烈变化，但也只是从数量上改变记忆库中的突发记忆数，并不能改变请求突发记忆所占的比重，也不能增加突发记忆的训练次数。而增大训练数据的大小，等于变相地增加算法训练的次数，没有实质性的意义。

为了解决上述问题，本文提出一种双缓存机制：构建两个记忆库分别存储常规记忆和请求突发记忆。基于网络状况设计一个请求突发阈值，根据网络状态的负载情况与阈值的大小关系，将记忆分别存入常规记忆库和突发记忆库。当抽取记忆组成训练数据时，根据网络特性，定义一个抽取比例 ξ，并分别从两个记忆库中随机抽取相应数量的记忆组成训练数据，训练神经网络。以此增加算法对于请求突发情况的适应性，实现稳定、快速的基站开关决策。

关于记忆抽取比例 ξ，我们需要注意：在训练初期，为了让 DQN 算法能够更快意识到请求突发状况的存在，更好地适应突发情况，抽取记忆进行训练时应该适当调高突发记忆的占比。但这种做法会给予 DQN 算法一个错误的信息——请求突发情况发生概率较高，使得它对环境的认知出现偏差，在真实环境中的适应性降低。因此，训练到一定阶段后，我们需要调节记忆抽取比例，使其接近真实的网络环境。也就是说，ξ 不是一个固定的值，会随着训练的推进不断变化。

用户请求的到达规律是难以预先获取的，因此预先设计记忆抽取比例的目标值是不切实际的。由于两个记忆库为了进行记忆的新老替换，都有一个计数器统计已产生的对应记忆总量，因此随着训练次数的增加，两种记忆更加趋向真实比例。因此，在双缓冲 DQN 算法中，记忆抽取比例会随着训练的推进，逐渐接近两个记忆库计数器的比值。

4.3 算法描述

基于上述考虑，我们设计了双缓冲的改进 DQN 算法，伪代码如算法 1 所示。当系统模型确认后，DQN 算法的状态、动作维度以及奖励的计算也就确定下来。首先根据本问题的状态和动作维度以及场景的复杂度，建立 Q 估计网络和 Q 现实网络，随机初始化神经网络的参数，并保持两者参数一致；初始化两个容量为 M_1 和 M_2 的记忆库，分别存储常规记忆和请求突发记忆。而且由于请求突发记忆的产生速率远不及常规记忆，因此 $M_1 > M_2$。

算法 1 基于双缓冲的改进 DQN 算法

1. 初始化网络超参数
2. 初始化两个记忆库 M_1 和 M_2
3. 初始化 Q 估计网络 $Q(\mathbf{s}_t, \mathbf{a}|\theta)$ 和 Q 现实网络 $Q(\mathbf{s}_t, \mathbf{a}|\theta')$，且 $\theta' = \theta$
4. FOR each episode DO
5. 环境初始化，获得初始状态 S_0
6. FOR each step DO
7. 以 ε 的概率选择 Q 值最大的动作，$\mathbf{a}_t = \max Q(\mathbf{s}_t, \mathbf{a}|\theta)$，否则获取随机动作 $\mathbf{a}_t$
8. 执行动作 $\mathbf{a}_t$，获得即时奖励以及下一状态 S_{t+1}
9. 根据状态 $\mathbf{s}_t$ 的各基站负载与请求突发与之比较，将记忆 $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ 存入对应记忆库
10. 依据双记忆库抽取比例，从两个记忆库中随机取出一批记忆
11. 根据这批数据计算损失度，并利用梯度下降更新神经网络参数 θ
12. 每隔 D 步同步一次两个神经网络的参数，即 $\theta' \leftarrow \theta$
13. END FOR
14. END FOR

在每一步训练中，环境先将状态传给 Q 估计网络，从而获得该状态下所有动作对应的 Q 值。算法第 6－13 行描述了 DQN 的决策过程。基于 ε-greedy 算法，随机产生一个 0 到 1 之间的数，若随机数小于 ε，则这一步执行 Q 值最大的动作，否则随机选择动作执行。

执行基站开关动作后，蜂窝网络会返回一个服务质量评价作为奖励，并进入到下一个状态，生成一条记忆。第 8，9 行分别描述的是双记忆库的存、取过程，在上节已进行了详细说明。

算法 1 的第 10、11 行是 DQN 的 Q 估计网络参数更新的过程。其计算 minibatch 的 loss，利用梯度下降调整神经网络参数 θ。第 12 行是将 Q 现实网络与 Q 估计网络的神经网络参数同步的过程，每隔 D 步同步一次。

5 模拟验证与分析

本节选择了 3 种算法与本文提出的算法在相同条件下进行性能的比较。实验分别比较了算法的收敛速度、决策效果等指标，验证了本文算法的良好性能。

5.1 对比算法

为了检验本文设计的双缓冲 DQN 算法（T-DQN）对于请求突发异构无线网络基站高能效管理问题的决策性能，我们选择了 3 个对比算法，在相同的条件下比较它们的各项性能。

（1）经典 DQN 算法（DQN）：其强化学习框架与双缓冲 DQN 算法一致，并且算法参数以及神经网络结构也相同。

（2）贪心算法（Greedy）：最大化每一步的奖励，从而使整个回合的奖励趋近最大。具体的做法是：对每一个状态，遍历所有的基站开关动作，并执行其中奖励最大的动作。

（3）基站常开策略（Keep_On）：所有小基站时刻保持开启。

5.2 实验参数设置

本问题考虑的服务场景如图 2 所示，整个区域被分为 4 个子区域，每个子区域都有一个宏基站负责全区域的信号覆盖，两个小基站部署在热点区域，负责应对突发请求。全服务区是一个 800 m×800 m 的正方形区域，宏基站的覆盖半径是 282.8 m，小基站的覆盖半径是 50 m，而一个用户网格是一个 50 m×50 m 的正方形区域。一个小基站覆盖 4 个网格，而一个宏基站覆盖 64 个网格。DQN 的状态维度为 257，动作空间大小为 $2^8 = 256$。基站负载突发阈值 $V_{th} = 1.0$，也就是说，当基站的负载超过 1 时，我们认定该网络状态为突发状态。

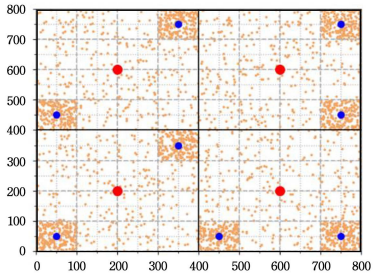


图 2 模拟网络场景图

Fig. 2 Simulated network scenario

考虑到本问题的规模，DQN 算法中的神经网络共有 4 层，其中输入层维度即状态维度为 257、两个中间层分别有 300 个神经元、神经网络的输出为各个动作的 Q 值，所以输出

层的维度为 256。此外，DQN 算法超参数设置如下：学习率 $\alpha=0.005$，衰减率 $\gamma=0.9$，开发探索比 $\varepsilon=0.9$，minibatch 大小 $\eta=200$，训练回合数上限 $T=200\,000$，每回合步数 $U=50$，记忆库总容量 $z=10\,000$，其中突发记忆库容量为 3 000，常规记忆库容量为 7 000，Q 现实网络更新频率 $K_{target}=5\,000$，Q 估计网络训练频率 $K_{eval}=10$。开始训练时每次抽取突发记忆 60 条，常规记忆 140 条。

5.3 算法性能分析

5.3.1 模型训练阶段

深度强化学习算法的训练结束条件大致分为两种：1）达到设置的训练回合数上限；2）达到设置的 loss 值下限。训练回合数上限一方面须保障能够及时结束无法收敛或收敛极慢的训练；另一方面，又须确保算法收敛而不过早地结束训练。因此，训练回合数上限难于准确设置。为了清晰地比较算法的收敛速度，设置 loss 下限的方式更加直观，但在 DQN 算法训练时，loss 是来回波动的，单次的 loss 到达下限并不能说明算法已经收敛了，存在偶然性因素。因此，在本文中设置了两种训练结束条件：

（1）以累计出现 200 次训练的 $loss<0.01$ 作为训练的主动结束条件；

（2）以最大回合数 200 000 作为训练的被动结束条件。

在本实验场景下，两种 DQN 算法的收敛速度对比如表 1 所列。可以看出，双缓冲 DQN 算法相较于传统 DQN 算法而言，能够更快地收敛，且收敛速度提升了约 20%。

表 1 DQN 和 T-DQN 收敛回合数对比

Table 1 Comparison on number of training rounds till convergence of DQN and T-DQN

DQN	T-DQN
71 758	59 039
76 960	57 562

从图 3 可看出，随着训练步数的增加，两种算法的 loss 均在波动中趋近收敛。loss 曲线波动越大，说明算法越不稳定，很明显 T-DQN 算法的 loss 曲线的波动程度小于传统 DQN 算法，其稳定性更好。结合表 1 和图 3 可以得出结论：T-DQN 算法相比传统 DQN 算法能够更快、更稳地收敛。

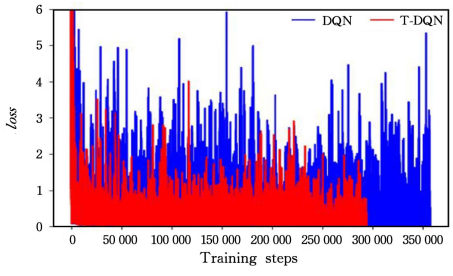


图 3 DQN 和 T-DQN 训练 loss 对比

Fig. 3 Comparison on training loss of DQN and T-DQN

图 4 是两种 DQN 算法在训练过程中每个回合总奖励的变化曲线对比图。由于各个回合的网络状况不同，因此回合奖励并不会特别收敛到一个固定值，而且训练阶段的探索过程也会使奖励上下波动。但随着神经网络决策性能的提升，回合奖励会有明显的增大趋势。仔细对比两条曲线能够看出，T-DQN 算法的增大趋势更加明显。DQN 算法虽然在初始时奖励上限较大，但一直在反复波动，最终被 T-DQN 算法超越。

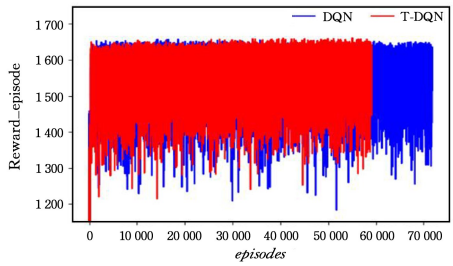


图 4 DQN 和 T-DQN 训练奖励对比

Fig. 4 Comparison on reward of DQN and T-DQN

如 3.4 节所述，本文的研究目标是在保障服务质量的前提下，尽可能地减少能量开销。能量消耗少，并不能说明算法决策效果好，也有可能是盲目关闭小基站，牺牲了用户体验带来的（训练初期的会有该情况发生）。结合能量变化曲线（见图 5）和奖励变化曲线（见图 6），可以得出结论：T-DQN 算法相比于 DQN 算法，训练效率更高，做出的决策也更能满足高能效基站管理的要求。

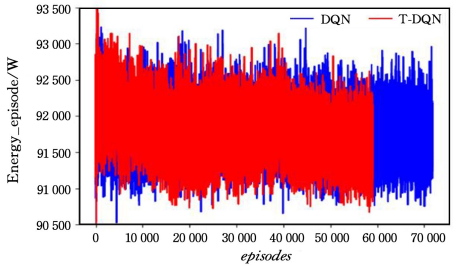


图 5 DQN 和 T-DQN 训练能耗对比

Fig. 5 Comparison on the energy consumption of DQN and T-DQN

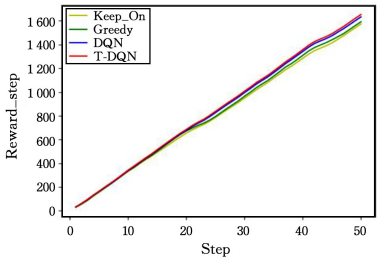


图 6 累计奖励对比

Fig. 6 Comparison on the accumulated reward loss of DQN and T-DQN

5.3.2 结果测试阶段

当两种 DQN 算法结束训练后，把 Keep_ON，Greedy，DQN 和 T-DQN 4 种算法置于相同的网络环境下，进行决策性能比较。与训练阶段不同的是，测试阶段两种 DQN 算法每次的小基站休眠决策都是由训练结束后智能体根据网络实时状况所做的决策。

由于双记忆库的记忆抽取比例 ξ 最终将接近两个记忆库的记忆产生总数比，也就是说，到训练后期，T-DQN 算法与 DQN 算法在抽取记忆组成 minibatch 时，两种记忆的抽取比例是相同的，两种算法本质上是一样的。因此，进行性能测试

时，两者的决策效果也相差无几。

图 7 和图 8 分别是在一个测试回合内 4 种算法的奖励和能量消耗逐步累加变化的对比图。虽然图 7 和图 8 中 4 条曲线的差别都比较小，但我们还是能明显看出 T-DQN 算法的能量消耗最少，决策奖励最高。T-DQN 算法相较于 DQN 算法、Greedy 算法、Keep_On 算法的决策奖励分别能够提升 1.3%，3.8%，5.1%，而在网络能耗方面，相较于其他 3 种算法，它分别能够减少 0.7%，1.8%，4.8%的网络能耗。

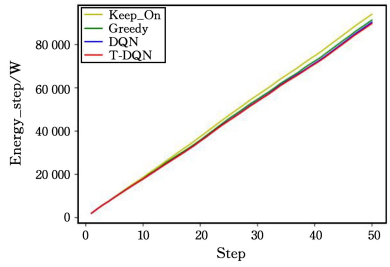


图 7 累计能耗对比

Fig. 7 Comparison on the total energy consumption

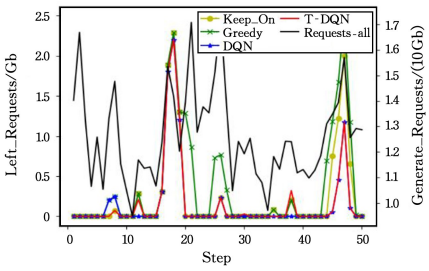


图 8 单步的请求产生量以及请求剩余量

Fig. 8 Number of requests and un-processed requests in each step

在图 8 中，为了更加清晰地比较 4 种算法对于请求突发的适应性，我们将一个测试回合内每一步产生的请求量曲线以及 4 种算法每一步决策未处理完的请求量曲线绘制在一起。在 17—19 步时，4 种算法均剩余了大量的请求未能完成，而且 4 种算法的差别很小，这说明此时整个网络产生了大量的请求。但观察请求产生曲线可以发现，此时产生的请求量并非该回合的峰值，这种情况表明，请求突发的区域不在小基站的覆盖范围，开启小基站作用有限，宏基站的覆盖区域依然会有大量的请求处理不完。与此相对的是，在 21 步和 26 步的用户请求数据量也非常高，但请求剩余量并不多。这说明小基站覆盖区域内发生了请求突发，通过开启小基站，显著增大了网络有效带宽，避免了大量请求剩余，提高了服务质量。

在 45—48 步，T-DQN 算法和 DQN 算法剩余的请求数量低于基站全开的 Keep_On 算法，这说明盲目开启小基站会使基站间的同信道干扰增大，反而可能导致网络的有效带宽变小(尤其会影响靠近小基站的宏基站覆盖区域的用户)。由此可以得出结论：训练完成后的 DQN 和 T-DQN 算法，通过有策略地关闭部分小基站，不但不会降低用户服务质量，反而能够缓解基站间的同信道干扰问题，实现网络的高能效管理。

图 9 和图 10 是 4 种算法在 25 个测试回合中的回合总奖励和回合总能耗变化曲线对比图。从图中可以看出，T-DQN 算法的总奖励大于其他 3 种对比算法，而总能量消耗相对来说低于其他 3 种对比算法。

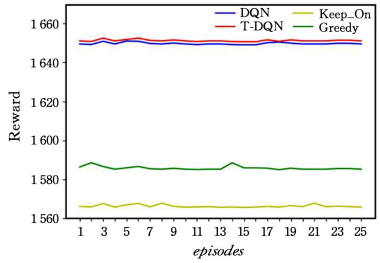


图 9 回合总奖励对比

Fig. 9 Comparison on the accumulated reward in each episode

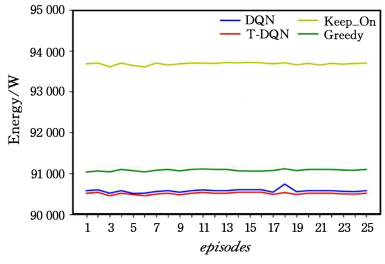


图 10 回合总能耗对比(电子版为彩色)

Fig. 10 Comparison on the energy consumption in each episode

综合考虑算法的训练阶段和测试过程，我们能够得出如下结论：本文提出的 T-DQN 算法不仅收敛迅速、稳定性好，而且其决策效果也明显优于其他 3 种算法，能够很好地适应请求突发的网络场景，针对异构蜂窝网络中的基站能效管理问题做出高效的决策。

结束语 本文围绕异构蜂窝网络中的基站高能效管理问题展开研究，重点研究请求突发场景下的小基站休眠决策问题。与已有研究工作不同，本工作不要求网络系统的先验知识，并且着重考虑了异构网络间的同信道干扰问题以及用户请求突发问题，旨在满足用户服务体验的前提下，实现最小化长远的基站能量消耗的目标，并利用双缓冲队列来改进现有的 DQN 算法，实现灵活、敏捷的高能效基站休眠管理。本文提出的改进算法对于真实 5G 网络的高能效基站开关决策方案的实现具有重要的参考价值，但其实际应用仍存在改进之处。例如，本文将场景简化为由宏基站和小基站构成的两层异构网络，仅考虑了同信道干扰问题及请求突发问题，而在现实网络中，基站开关决策还面临基站类型多样及频谱分配复杂等问题，不同的基站发射功率以及覆盖范围、各基站的信道分配问题等使得该方法在应用于现实网络环境时仍面临着诸多挑战。而且初始双缓冲记忆库抽取比例 ξ 会在很大程度上影响算法的收敛速度和实验结果，ξ 初值的设置仍是一项需要深入研究的工作。

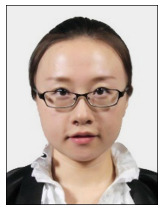
参 考 文 献

[1] OH E, KRISHNAMACHARI B, LIU X, et al. Toward dynamic energy efficient operation of cellular network infrastructure[J]. IEEE Communications Magazine, 2011, 49(6): 56-61.

[2] LÄHDEKORPI P, HRONEC M, JOLMA P, et al. Energy efficiency of 5G mobile networks with base station sleep modes [C]// 2017 IEEE Conference on Standards for Communications

and Networking (CSCN). IEEE,2017:163-168.
[3] ONIRETI O,MOHAMED A,PERVAIZ H,et al. Analytical approach to base station sleep mode power consumption and sleep depth[C]// 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE,2017:1-7.
[4] ONIRETI O,MOHAMED A,PERVAIZ H,et al. A tractable approach to base station sleep mode power consumption and deactivation latency[C]// 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE,2018:123-128.
[5] PERVAIZ H,ONIRETI O,MOHAMED A,et al. Energy-efficient and load-proportional eNodeB for 5G user-centric networks:a multilevel sleep strategy mechanism[J]. IEEE Vehicular Technology Magazine,2018,13(4):51-59.
[6] LI J,WANG H,WANG X,et al. Optimized sleep strategy based on clustering in dense heterogeneous networks[J]. EURASIP Journal on Wireless Communications and Networking, 2018, 2018(1):1-10.
[7] RATHEESH R,VETRIVELAN P. Energy efficiency based on relay station deployment and sleep mode activation of eNBs for 4G LTE-A network[J]. Automatika,2019,60(3):322-331.
[8] KLAPEZ M,GRAZIA C A,CASONI M. Energy Savings of Sleep Modes Enabled by 5G Software-Defined Heterogeneous Networks[C]// 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI). IEEE, 2018:1-6.
[9] JAWAD A M,JAWAD H M,NORDIN R,et al. Wireless power transfer with magnetic resonator coupling and sleep/active strategy for a drone charging station in smart agriculture[J]. IEEE Access,2019,7:139839-139851.
[10] WU J,BAO Y,MIAO G,et al. Base station sleeping and power control for bursty traffic in cellular networks[C]// 2014 IEEE International Conference on Communications Workshops (ICC). IEEE,2014:837-841.
[11] WU J,BAO Y,MIAO G,et al. Base-station sleeping control and power matching for energy-delay tradeoffs with bursty traffic [J]. IEEE Transactions on Vehicular Technology,2015,65(5): 3657-3675.
[12] LIU J,KRISHNAMACHARI B,ZHOU S,et al. Deepnap:Data-driven base station sleeping operations through deep reinforcement learning[J]. IEEE Internet of Things Journal,2018,5(6): 4273-4282.
[13] GHADIMI E,CALABRESE F D,PETERS G,et al. A reinforcement learning approach to power control and rate adaptation in cellular networks[C]// 2017 IEEE International Conference on Communications (ICC). IEEE,2017:1-7.
[14] WANG L,PETERS G,LIANG Y C,et al. Intelligent User-Centric Networks: Learning-Based Downlink CoMP Region Breathing[J]. IEEE Transactions on Vehicular Technology, 2020, 69(5):5583-5597.
[15] LIU Q,SHI J. Base station sleep and spectrum allocation in heterogeneous ultra-dense networks[J]. Wireless Personal Communications,2018,98(4):3611-3627.
[16] NIU Z,GUO X,ZHOU S,et al. Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control[J]. IEEE Journal on Selected Areas in Communications, 2015,33(4):641-650.
[17] CHEN X,WU J,CAI Y,et al. Energy-efficiency oriented traffic offloading in wireless networks:a brief survey and a learning approach for heterogeneous cellular networks[J]. IEEE Journal on Selected Areas in Communications,2015,33(4):627-640.
[18] FENG M,MAO S,JIANG T. Boost:Base station on off switching strategy for energy efficient massive mimo hetnets[C]// IEEE INFOCOM 2016 The 35th Annual IEEE International Conference on Computer Communications. IEEE,2016:1-9.
[19] SAMARAKOON S,BENNIS M,SAAD W,et al. Opportunistic sleep mode strategies in wireless small cell networks[C]// 2014 IEEE International Conference on Communications (ICC). IEEE,2014:2707-2712.
[20] SALEM F E,ALTMAN Z,GATI A,et al. Reinforcement learning approach for advanced sleep modes management in 5G networks[C]// 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE,2018:1-5.
[21] EL-AMINE A,ITURRALDE M,HASSAN H A H,et al. A distributed Q-Learning approach for adaptive sleep modes in 5G networks[C]// 2019 IEEE Wireless Communications and Networking Conference (WCNC). IEEE,2019:1-6.
[22] ZHANG Y T. A Deep Reinforcement Learning based Dynamic C-RAN Resource Allocation Method[J]. Journal of Chinese Computer Systems,2021,42(1):132-136.

ZENG De-ze, born in 1984, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include edge computing and artificial intelligence.



GU Lin, born in 1985, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include edge computing and artificial intelligence.