

基于深度强化学习的异构云无线接入网自适应无线资源分配算法

陈前斌 管令进* 李子煜 王兆堃 杨 恒 唐 伦

(重庆邮电大学通信与信息工程学院 重庆 400065)

(重庆邮电大学移动通信技术重点实验室 重庆 400065)

摘 要: 为了满足无线数据流量大幅增长的需求, 异构云无线接入网(H-CRAN)的资源优化仍然是亟待解决的重要问题。该文在H-CRAN下行链路场景下, 提出一种基于深度强化学习(DRL)的无线资源分配算法。首先, 该算法以队列稳定为约束, 联合优化拥塞控制、用户关联、子载波分配和功率分配, 并建立网络总吞吐量最大化的随机优化模型。其次, 考虑到调度问题的复杂性, DRL算法利用神经网络作为非线性近似函数, 高效地解决维度灾难问题。最后, 针对无线网络环境的复杂性和动态多变性, 引入迁移学习(TL)算法, 利用TL的小样本学习特性, 使得DRL算法在少量样本的情况下也能获得最优的资源分配策略。此外, TL通过迁移DRL模型的权重参数, 进一步地加快了DRL算法的收敛速度。仿真结果表明, 该文所提算法可以有效地增加网络吞吐量, 提高网络的稳定性。

关键词: 异构云无线接入网络; 资源分配; 深度强化学习; 迁移学习

中图分类号: TN929.5

文献标识码: A

文章编号: 1009-5896(2020)06-1468-10

DOI: [10.11999/JEIT190511](https://doi.org/10.11999/JEIT190511)

Deep Reinforcement Learning-based Adaptive Wireless Resource Allocation Algorithm for Heterogeneous Cloud Wireless Access Network

CHEN Qianbin GUANG Lingjin LI Ziyu WANG Zhaokun

YANG Heng TANG Lun

(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

(Key Laboratory of Mobile Communication Technology, Chongqing University of Post and Telecommunications, Chongqing 400065, China)

Abstract: In order to meet the demand of the substantial increase of wireless data traffic, the resource optimization of the Heterogeneous Cloud Radio Access Network (H-CRAN) is still an important problem that needs to be solved urgently. In this paper, under the H-CRAN downlink scenario, a wireless resource allocation algorithm based on Deep Reinforcement Learning (DRL) is proposed. Firstly, a stochastic optimization model for maximizing the total network throughput is established to jointly optimize the congestion control, the user association, subcarrier allocation and the power allocation under the constraint of queue stability. Secondly, considering the complexity of scheduling problem, the DRL algorithm uses neural network as nonlinear approximate function to solve the dimensional disaster problem efficiently. Finally, considering the complexity and dynamic variability of the wireless network environment, the Transfer Learning(TL) algorithm is introduced to make use of the small sample learning characteristics of TL so that the DRL algorithm can obtain the optimal resource allocation strategy in the case of insufficient samples. In addition, TL further accelerates the convergence rate of DRL algorithm by transferring the weight parameters of DRL model. Simulation results show that the proposed algorithm can effectively increase network throughput and improve network stability.

收稿日期: 2019-07-08; 改回日期: 2020-03-09; 网络出版: 2020-04-15

*通信作者: 管令进 1633634329@qq.com

基金项目: 国家自然科学基金(6157073), 重庆市教委科学技术研究项目(KJZD-M201800601)

Foundation Items: The National Natural Science Foundation of China (61571073), The Science and Technology Research Program of Chongqing Municipal Education Commission (KJZD-M201800601)

Key words: Heterogeneous Cloud Radio Access Networks(H-CRAN); Resource allocation; Deep Reinforcement Learning(DRL); Transfer Learning(TL)

1 引言

根据爱立信2018年发布的移动市场报告预测,2024年5G用户将达到15亿,移动数据总流量将增长5倍以上。由于用户的业务需求量急剧增长,资源优化面临着较以往更加严峻的挑战。为了提升网络的服务质量和业务体验,异构云无线接入网(Heterogeneous Cloud Radio Access Networks, H-CRAN)得到了广大学者的关注。H-CRAN将云无线接入网(Cloud-Radio Access Network, C-RAN)和异构网络融合,通过大功率结点(High Power Node, HPN)实现控制平面功能和全网的无缝覆盖,利用远端无线射频单元(Remote Radio Head, RRH)实现热点区域海量业务的大容量传输。深度强化学习(Deep Reinforcement Learning, DRL)是近年来人工智能领域最受关注的方向之一,学术界和企业界纷纷将DRL算法应用于实际,分别在视频游戏^[1]和机器博弈^[2]领域取得丰硕的成果。因此,将DRL算法应用到H-CRAN中具有重要的价值。

文献[3]研究了基于正交频分多址的异构蜂窝网络中功率分配和无线回程带宽分配,设计了近似最优迭代资源分配算法来解决资源分配问题。文献[4]在具有用户协作和服务质量(Quality of Service, QoS)保证的H-CRAN下行链路场景中研究了网络的能效,联合中继选择、功率分配和网络选择,以此最大化具有高QoS移动用户的能效。文献[5]在干扰对齐无线网络中考虑实际的时变信道,通过DRL算法在高速缓存条件下得到最优的用户选择策略,显著地提高了网络的总吞吐量和能效。文献[6]在云无线接入网络中研究网络切片中的虚拟资源分配,通过马尔可夫决策算法进行在线资源分配,进而最大化网络吞吐量。

尽管上述文献在无线资源分配上都取得了较好的研究成果,但仍需要进一步的改进,主要存在3方面的问题:(1)多数工作没有考虑H-CRAN带来的优势,同时忽略了用户队列的拥塞控制;(2)大多数工作仍然采取传统的迭代运算方式进行研究,人工参与因素明显,资源优化效率较低,无法面对复杂的网络环境变化;(3)DRL模型在新的网络中无法进行复用,不仅需要从环境中获取大量的训练样本,还需要重新训练神经网络,耗费了大量的训练时间。针对上述问题,本文提出一种基于深度强化学习的自适应资源分配算法。本文所提算法的主要创新点如下:(1)为了进一步提高资源分配的有

效性,在H-CRAN下行传输场景下,联合队列的拥塞控制、用户关联、子载波分配和功率分配,从而最大化网络的总吞吐量;(2)针对网络中复杂的资源调度问题,本文采用DRL算法,将队列状态信息(Queue State Information, QSI)、信道状态信息(Channel State Information, CSI)和基站的发射功率作为状态空间,执行用户关联、子载波分配和功率分配等动作,根据吞吐量的变化定义相应的奖励函数,使资源分配更加自动化和智能化;(3)为了使模型更好地适应H-CRAN动态变化的环境,提出了一种新颖的迁移深度强化学习(Transfer Learning-Deep Reinforcement Learning, TLDQN)模型,根据TL的小样本学习特性,使DRL模型在少量样本的条件下,也能取得最优资源分配方案。在另一方面,通过迁移神经网络的权重参数,进一步提高了算法的收敛速度。

2 问题描述与系统模型

2.1 5G异构云无线接入网架构

考虑H-CRAN下行传输场景,如图1所示,包含1个HPN和多个RRH的异构无线网络,H-CRAN中数量众多的低功率RRH互相协作,并在集中式基带单元(Base Band Unite, BBU)池中实现信号处理,RRH只需执行射频处理以及简单的基带处理,主要的基带信号处理以及上层协议功能均在BBU池中实现。相比于传统的C-RAN架构,H-CRAN将云计算技术嵌入到异构网络中,从而实现大规模的协同信号处理和组网功能,因此H-CRAN极大地提高了频谱效率。H-CRAN中的BBU池与HPN通过S1和X2接口连接,且通过协作HPN来缓解RRH和BBU池之间的前传容量受限难题。此外,H-CRAN充分利用3G和4G等蜂窝网络的宏基站来实现无缝覆盖,同时实现控制和业务平面功能分离。所有的控制信令和系统广播信息由HPN发送给用户,突发业务以及即时消息等低速率数据信息也由HPN承载,确保业务的无缝覆盖,RRHs主要负责热点区域海量数据业务的高速传输^[7]。

2.2 系统模型

针对传统资源优化和管理方式效率较低的问题,本文设计一种智能资源管理架构。主要包括2个部分:智能资源管理器 and H-CRAN。首先,用户提交各种业务请求,并在用户队列处进行排队;其次,基于CSI, QSI和基站发射功率,控制器执行它的学习算法,从而满足用户的业务请求。监控器的任务是收集系统的信道状态信息、队列状态信

息、基站发射功率和用户的QoS需求来更新控制器。分配器根据控制器的配置参数将应用请求映射到BBU池,进行无线资源的智能化分配。

控制器是智能资源管理架构的关键部分,其核

心是DRL和TL算法,DRL将强化学习(Reinforcement Learning, RL)与深度学习(Deep Learning, DL)联合起来。如图2所示,控制器根据RL与环境进行交互,再通过DL的非线性近似特征,利用存储的样

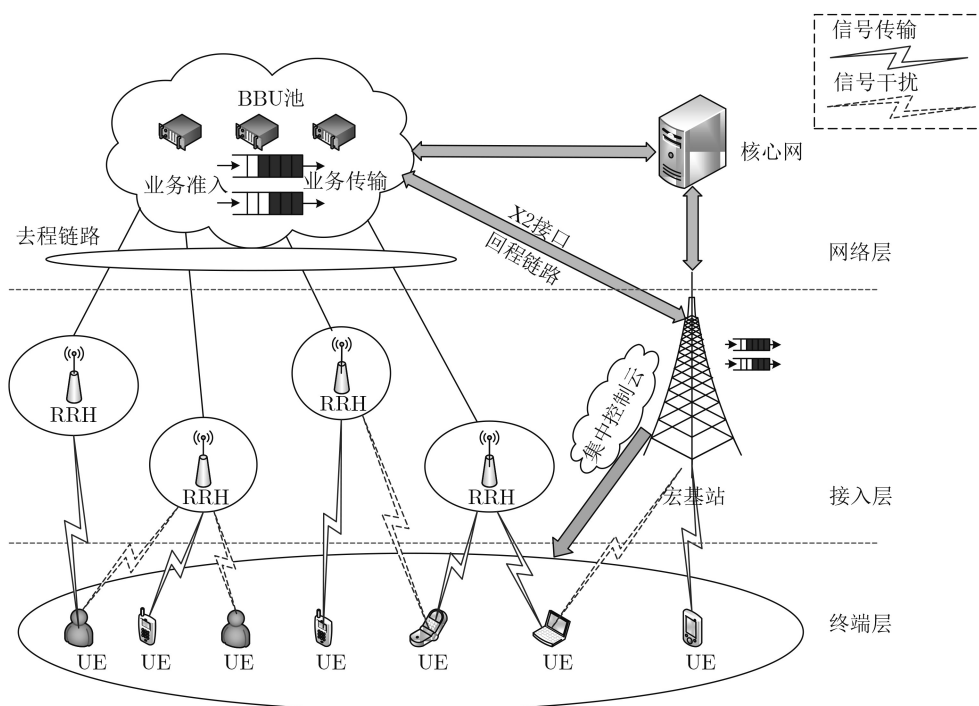
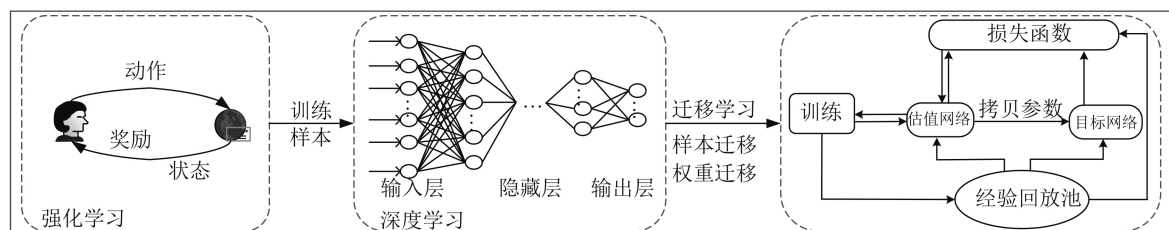


图1 H-CRAN下行传输场景



深度强化学习(DRL)+迁移学习

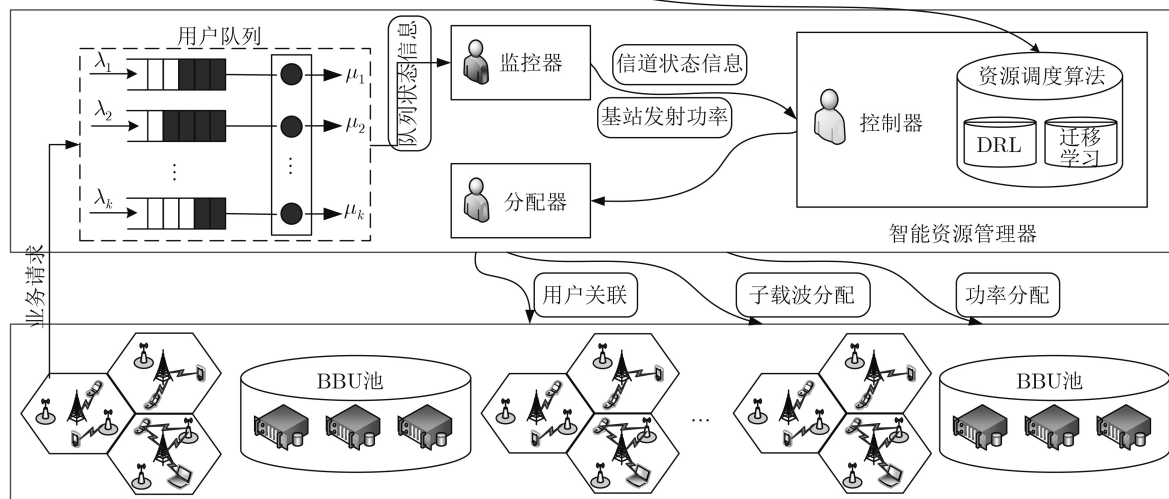


图2 系统架构

本对神经网络的权重参数进行梯度更新。TL算法在DRL的基础上,通过迁移RL与环境进行交互的训练样本,实现TL的小样本学习特征,减少DRL的训练时间;另一方面,通过迁移神经网络的权重参数,进而提高DRL算法的收敛速度。

2.3 物理层模型

基于文献[4]的分析,为了进一步提高资源分配的有效性,研究H-CRAN下行传输场景。定义HPN和RRHs的集合分别为 $\{0, 1, 2, \dots, J\}$,其中0代表HPN, $J = \{1, 2, \dots, J\}$ 表示RRHs的集合。由于HPN主要用于传递信令并保证网络的覆盖,具有低速率的用户设备(HUE设备)通常被HPN服务,反之,具有高QoS的用户设备(RUE设备)由RRHs来提供服务,用 $U_H = \{1, 2, \dots, u_h\}$ 和 $U_R = \{1, 2, \dots, u_r\}$ 分别表示HUE和RUE设备的集合。假设H-CRAN系统的运行时间被划分为多个持续时间为 τ 的时隙,各时隙的索引用 t 表示。

用 $p_{ijn}(t)$ 和 $p_{mn}(t)$ 分别表示第 t 时隙RRH j 在第 n 个子载波上分配给用户RUE i 的功率和HPN在第 n 个子载波上分配给用户HUE m 的功率。用 $g_{ijn}(t)$ 和 $g_{mn}(t)$ 分别表示第 t 时隙从RRH j 到RUE i ,从HPN到HUE m 在第 n 个子载波上的信道增益。信道状态信息包括了路径损耗,阴影衰落及噪声的影响,对于CSI而言,本文假设它们在每时隙内保持固定而在时隙间独立同分布。子载波的集合为 $\{1, 2, \dots, N\}$,为避免小区内部干扰,在同一个RRH或者HPN上每个子载波只能分给一个用户使用

$$\rho_{ijn}(t) = \begin{cases} 1, & \text{第}t\text{时隙RRH}j\text{将子载波}n\text{分配给RUE}i \\ 0, & \text{其他} \end{cases} \quad (1)$$

$$\rho_{mn}(t) = \begin{cases} 1, & \text{第}t\text{时隙HPN将子载波}n\text{分配给HUE}m \\ 0, & \text{其他} \end{cases} \quad (2)$$

在H-CRAN中,不同的RRH间以及RRH和HPN间的同频用户存在着干扰。假设系统采取连续的速率方案,可以用香农公式来表示传输速率,因此第 t 时隙RUE i 和HUE m 在子载波 n 上的传输速率分别为

$$R_{ijn}(t) = \frac{B}{N} \log_2 \left(1 + \frac{p_{ijn}(t)g_{ijn}(t)}{\sum_{j' \neq j} p_{ij'n}(t)g_{ij'n}(t) + p_{mn}(t)g_{mn}(t) + \sigma^2} \right) \quad (3)$$

$$R_{mn}(t) = \frac{B}{N} \log_2 \left(1 + \frac{p_{mn}(t)g_{mn}(t)}{\sum_{j \in J} p_{ijn}(t)g_{ijn}(t) + \sigma^2} \right) \quad (4)$$

其中, B 为系统的总带宽, σ^2 为高斯信道噪声。

在H-CRAN中,用户关联策略对提高无线资源的利用率非常关键。在第 t 时隙,1个用户只能关联到1个基站上,RUE用户和HUE用户的关联变量 $u_{ij}(t)$ 和 $u_m(t)$ 分别表示为

$$u_{ij}(t) = \begin{cases} 1, & \text{第}t\text{时隙RUE}i\text{关联到RRH}j\text{上} \\ 0, & \text{其他} \end{cases} \quad (5)$$

$$u_m(t) = \begin{cases} 1, & \text{第}t\text{时隙HUE}m\text{关联到HPN上} \\ 0, & \text{其他} \end{cases} \quad (6)$$

在RUE i 和HUE m 的传输速率分别为 $R_{ij}(t) = \sum_{n=1}^N u_{ij}\rho_{ijn}(t)R_{ijn}(t)$ 和 $R_m(t) = \sum_{n=1}^N u_m\rho_{mn}(t)R_{mn}(t)$ 。因此,可以得到网络的总速率和为 $R(t) = \sum_{j=1}^J \sum_{i=1}^{u_r} R_{ij}(t) + \sum_{m=1}^{u_h} R_m(t)$ 。类似地,RUE i 和HUE m 的传输功率分别为 $p_{ij}(t) = \sum_{n=1}^N u_{ij}\rho_{ijn}(t)p_{ijn}(t)$ 和 $p_m(t) = \sum_{n=1}^N u_m\rho_{mn}(t)p_{mn}(t)$,则RRH j 和HPN的传输功率分别为 $p_j(t) = \sum_{i=1}^{u_r} p_{ij}(t)$ 和 $p_H(t) = \sum_{m=1}^{u_h} p_m(t)$ 。因此,网络的总功率消耗为 $P_{\text{sum}}(t) = \sum_{j \in J} p_j(t) + p_j^c + p_H(t) + p_H^c$ 。其中 p_j^c 和 p_H^c 分别表示RRH和HPN上的静态功率消耗。

2.4 业务队列模型

针对用户业务到达的随机性和瞬时性,本文提出业务队列模型,用来表征业务队列的动态变化过程,使H-CRAN根据队列模型可以更加智能地分配无线资源,从而有效地控制业务排队时延。用 $Q_u(t), u \in \{u_r, u_h\}$ 表示第 t 时隙用户 u 的缓冲队列长度,用 $A_u(t), u \in \{u_r, u_h\}$ 表示第 t 时隙用户 u 的随机业务到达量。 $A_u(t)$ 在时隙间服从泊松分布,且不同数据包是相互独立的。因此用户 u 缓冲队列的动态变化为

$$Q_u(t+1) = \max\{Q_u(t) - \mu_u(t)\tau + A_u(t), 0\}, u \in \{u_r, u_h\} \quad (7)$$

为了保证业务队列的稳定性,根据网络稳定性^[8]的定义:当离散时间队列过程 $Q(t)$ 满足 $Q(t) = \lim_{t \rightarrow \infty} \sup \frac{1}{t} \sum_{t=0}^{t-1} E\{Q(t)\} < \infty$ 时,它是强稳定的。为了最大化基站吞吐量并保证业务队列的稳定性,在H-CRAN中解决由于大量的业务请求导致的业务队列拥塞问题被建模为如式(8)的随机优化问题

$$\left. \begin{aligned} & \max_{\{u, \rho, p\}} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T R(t) \\ & \text{s.t. C1: } u_{ij}(t) \in \{0, 1\}, \forall i \in U_r; u_m(t) \in \{0, 1\}, \\ & \quad \forall m \in U_h \\ & \text{C2: } \rho_{ijn}(t) \in \{0, 1\}, \forall n \in \{1, 2, \dots, N\} \\ & \text{C3: } \rho_{mn}(t) \in \{0, 1\}, \forall n \in \{1, 2, \dots, N\} \\ & \text{C4: } Q_j(t) < \infty, Q_m(t) < \infty, \forall j, m \\ & \text{C5: } p_j(t) \leq p_{\text{RRH}}^{\text{max}}, \forall j, t; p_m(t) \leq p_{\text{HPN}}^{\text{max}}, \forall m \end{aligned} \right\} \quad (8)$$

其中, C1表示每个用户只能关联到1个基站上; C2和C3分别表示在RRH和HPN中每个子载波只能分配给1个用户使用; C4保证每个用户的排队时延有限; C5限制了RRH和HPN的最大发射功率。所有队列通过控制策略耦合在一起, 以寻求最佳的资源优化动作来实现网络吞吐量最大化。

3 问题转化与算法描述

3.1 基于DQN的无线接入网资源优化算法

本文在H-CRAN架构下进行用户关联、子载波分配和功率分配, 根据文献[9]的分析, 采用传统的迭代运算方式, 系统的状态和动作空间是非常大的, 极有可能出现维度灾问题。受DQN成功地解决机器博弈问题的启发[2], 它可以自动地从高维的原始数据中提取特征, 并且DQN已经被用在无线网络中来提高网络性能[5]。因此, 本文将基站上联合拥塞控制和资源分配问题转化一个深度Q学习过程, 它可以在稳定业务队列的同时找到最优的资源分配策略。基站与无线网络环境进行交互过程中, 采用 ϵ 贪心策略进行探索, 并且选择回报最大的动作。

状态集定义为 $\mathbf{S}(t) = \{Q_{\text{avg}}(t), \text{CSI}(t), p_{\text{H}}(t), p_{\text{J}}(t)\}$, $Q_{\text{avg}}(t)$ 为 t 时刻业务平均队列长度, $\text{CSI}(t)$ 为 t 时刻信道状态信息, $p_{\text{H}}(t)$ 和 $p_{\text{J}}(t)$ 分别为在 t 时刻 HPN 和 RRH 的发射功率。为了降低算法的复杂度和状态空间大小, 还需要对 HPN 和 RRH 上的发射功率进行离散化处理。动作集用 $\mathbf{A}(t) = \{u_{ij}(t), u_m(t), \rho_{ijn}(t), \rho_{mn}(t), p_{ij}(t), p_m(t)\}$ 来表示, $u_{ij}(t)$ 和 $u_m(t)$ 分别表示 RRH 用户和 HPN 用户与基站进行关联的信息, $\rho_{ijn}(t)$ 和 $\rho_{mn}(t)$ 分别为 RRH 用户和 HPN 用户上子载波的使用情况, $p_{ij}(t)$ 和 $p_m(t)$ 分别表示 RRH 用户 i 和 HPN 用户 m 的功率分配信息, 其中 $p(t) \in \{-|p(t)|, 0, |p(t)|\}$ 。本文目标是最大化基站的吞吐量, 根据文献[10]可得网络的奖励函数表示为

$$r_k(t) = \begin{cases} 1 - e^{-R_{kj}(t)}, & k \in U_{\text{R}}, j \in J \\ 1 - e^{-R_k(t)}, & k \in U_{\text{H}} \\ -1, & \text{其他} \end{cases} \quad (9)$$

系统即时奖励是网络中所有用户的即时奖励总

和, 等价于 $r(t) = \sum r_k(t)$, $\mathbf{K} \in \{U_{\text{R}}, U_{\text{H}}\}$ 。在保证业务队列稳定的前提下, 为了最大化网络的总吞吐量, 必须选择合适的资源分配方案来最大化未来累积奖励

$$Q^*(s, a) = \max_{\pi} \left\{ \sum_{t=0}^{T-1} \gamma^t E(r(s, a)_t^{(i)} | s_t = s, a_t = a, \pi) \right\} \quad (10)$$

其中, $Q^*(s, a)$ 为最优的 Q 值函数, 通常情况下以递归方式获取函数 (s, a, r, s', a') , 根据式(11)进行更新

$$Q_{t+1}(s, a) = (1 - \alpha)Q_t(s, a) + \alpha(r + \gamma \max_{a'} Q_t(s', a')) \quad (11)$$

其中, α 是学习率, $\gamma \in [0, 1]$ 是折扣因子。式(10)在 $t \rightarrow \infty$ 时才能收敛到最优的 Q 值, 这种方法是不现实的。因此, 本文的DQN使用神经网络 $Q(s, a; w) \approx Q^*(s, a)$ 作为非线性近似器, 为缓解非线性网络中出现的稳定性问题, DQN使用经验回放池来减少训练样本的相关性, 利用目标网络来防止估值网络的预估值出现失控, 使训练更加稳定。基于TensorFlow架构的TensorBoard模块建立DQN模型的可视化, 如图3所示。

DQN在迭代过程中通过最小化损失函数 $\text{Loss}(w) = E[(r + \gamma \max_{a'} Q(s', a'; w^-) - Q(s, a; w))^2]$ 来训练网络参数, 其中 w^- 和 w 分别为目标网络和估值网络的参数, 需要对损失函数进行梯度下降更新

$$\nabla_w \text{Loss}(w) = E[(r + \gamma \max_{a'} Q(s', a'; w^-) - Q(s, a; w)) \nabla_w Q(s, a; w)] \quad (12)$$

在经验回放元组 D 中抽取一部分样本 $e(t) = (s(t), a(t), r(t), s(t+1))$, 根据当前抽取的样本来计算估值网络参数 w 的梯度 dw , 再使用自适应估计(Adaptive moment estimation, Adam)算法更新估值网络的参数 w

$$V_{\text{dw}} = \beta_1 \cdot V_{\text{dw}} + (1 - \beta_1) \cdot dw \quad (13)$$

$$S_{\text{dw}} = \beta_2 \cdot S_{\text{dw}} + (1 - \beta_2) \cdot dw^2 \quad (14)$$

式(13)用来计算指数加权平均数的1阶矩估计, 式(14)用来计算平方数的指数加权平均数的2阶矩估计。为了防止初始化权重更新出现的严重偏差, Adam对1阶矩和2阶矩估计都进行了修正

$$\bar{V}_{\text{dw}} = \frac{V_{\text{dw}}}{1 - \beta_1^t} \quad (15)$$

$$\bar{S}_{\text{dw}} = \frac{S_{\text{dw}}}{1 - \beta_2^t} \quad (16)$$

其中, $\bar{V}_{\text{dw}}$ 和 $\bar{S}_{\text{dw}}$ 分别1阶矩和2阶矩的偏差修正项, 再根据偏差修正项来更新估值网络的参数 w

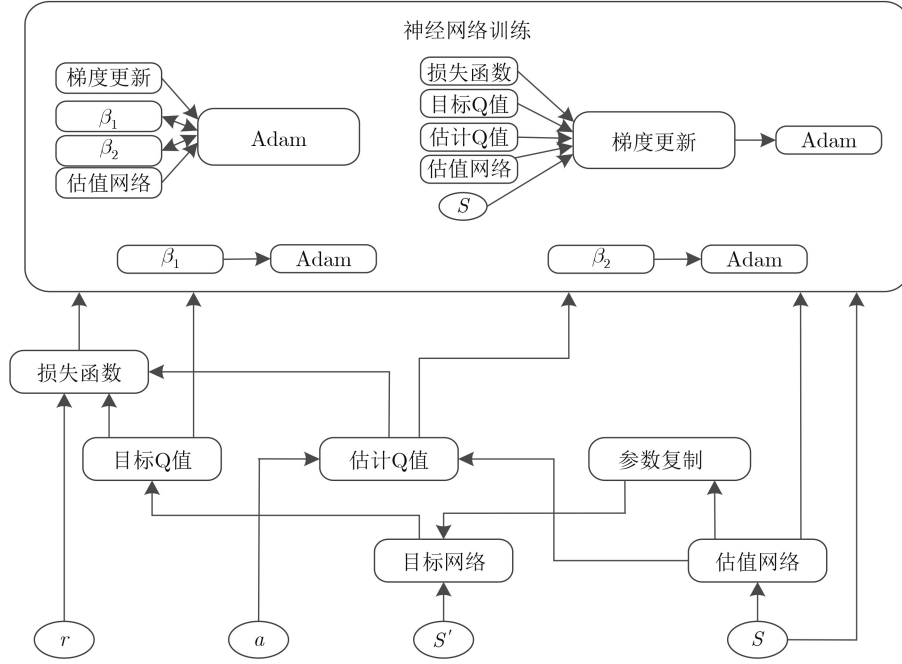


图3 DQN算法框图

$$w = w - \alpha \cdot \frac{\bar{V}_{dw}}{\sqrt{S_{dw} + \psi}} \quad (17)$$

其中, ψ 的取值为 10^{-8} , 为了获得估值网络的最优权重参数, 利用表1所示的算法1训练DQN模型的估值网络。

通过算法1将DQN模型训练好后, 可以获取DQN网络的最优权重参数 w 。H-CRAN根据最大信噪比准则完成用户关联的初始化, 在用户关联的基础上, 对基站上关联的用户进行均分子载波, 完成子载波的初始化, 最后根据注水算法对子载波上的功率进行初始化。最后, 通过DQN资源分配算法来获取最优资源分配策略, 从而最大化整个网络的吞吐量。本文算法时间复杂度由算法1决定, 由于算法1采用基于神经网络的深度强化学习。因此, 系统的时间复杂度为 $O(m \times (n_1 \times n_2 + n_2 \times n_3 + n_3 \times n_4 + n_4 \times n_5))$, 其中 m 为神经网络的输入, n_i 表示神经网络中第 i 层神经元的数目。

3.2 迁移学习

迁移学习对无线网络的资源分配而言, 是一个全新的学习领域。在本文中, 基站持续地与环境进行交互并做出决策, 这个过程是非常耗时的, 深度神经网络的强大特征表示和函数近似也常常伴随着高成本的计算。对于新接入H-CRAN网络的基站来说, 当其面对业务队列拥塞问题时, 需要重新与无线网络环境进行交互并重新训练DQN模型, 所以许多研究者都关注于可以加快训练速度的理论研究^[11]。

根据文献[12]中对迁移学习的研究, 为了使DRL模型快速地帮助新接入网络的基站处理业务队列拥

表1 算法1

算法1: DQN训练估值网络参数算法

- (1) 初始化经验回放池
- (2) 随机初始化估值网络中的参数 w , 初始化目标网络中的参数 w^- , 权重为 $w^- = w$
- (3) For episode $k = 0, 1, \dots, K - 1$ do
- (4) 随机初始化一个状态 s_0
- (5) For $t = 0, 1, \dots, T - 1$ do
- (6) 随机选择一个概率 p
- (7) if $p \leq \epsilon$ 资源管理器随机选择一个动作 $a(t)$
- (8) else 资源管理器根据估值网络选取动作 $a^*(t) = \arg \max_a Q(s, a; w)$
- (9) 执行动作 $a(t)$, 根据式(9)得到奖励值 $r(t)$, 并观察下一个状态 $s(t+1)$
- (10) 将元组 $(s(t), a(t), r(t), s(t+1))$ 存储到经验回放池中
- (11) 从经验回放池中随机抽取选取一组样本 $(s(t), a(t), r(t), s(t+1))$
- (12) 通过估值网络和目标网络的输出损失函数, 利用式(13), (14)计算1, 2阶矩
- (13) Adam算法通过式(15), 式(16)计算1阶矩和2阶矩的偏差修正项
- (14) 通过神经网络的反向传播算法, 利用式(17)来更新估值网络的权重参数 w
- (15) 每隔 δ 将估值网络中的参数 w 复制给参数 w^-
- (16) End for
- (17) End for
- (18) 获得DQN网络的最优权重参数 w

塞问题, 本文采用迁移学习来加速学习的进程。如图4所示, 首先, 迁移学习将源基站与环境进行交互得到的训练样本迁移到目标基站上, 减少了训练时间, 同时避免了目标基站由于样本不足导致的过

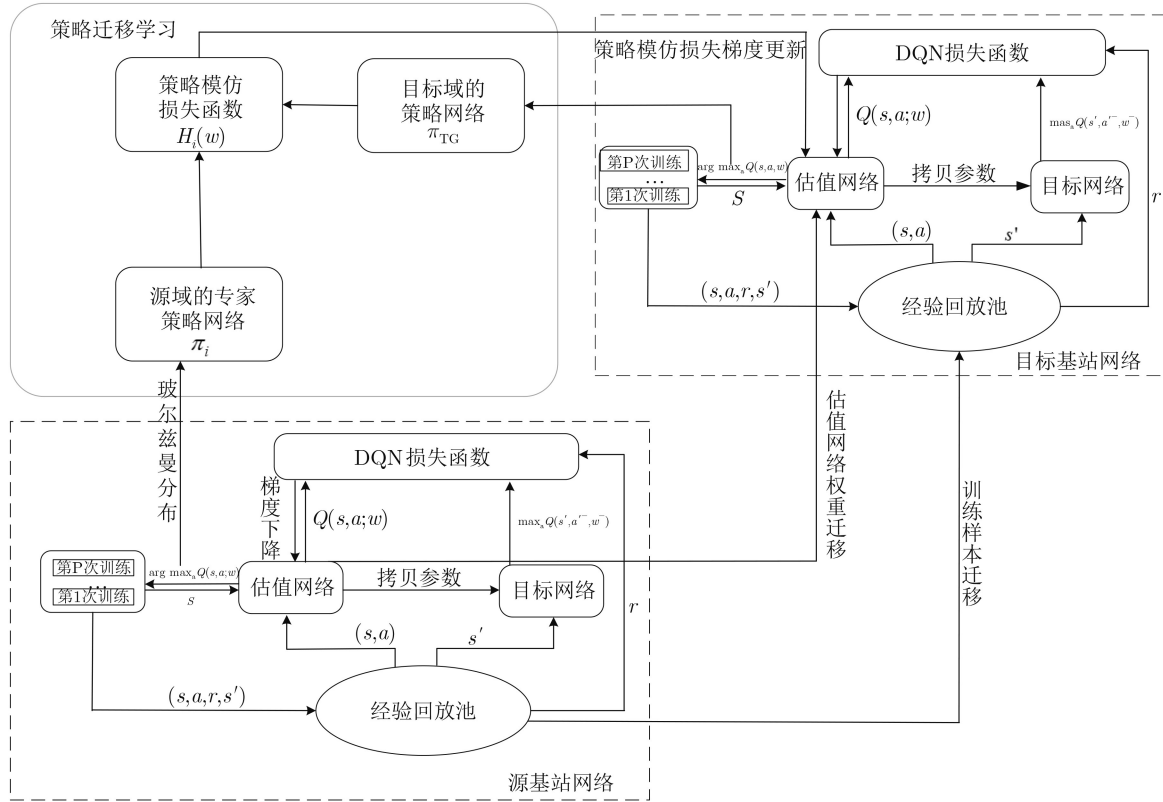


图4 迁移学习场景图

拟合问题；其次，通过重用已经训练好的估值网络权重参数，进一步加速算法的收敛性。对比于标准的DQN以随机的搜寻方法开始学习，使用迁移学习可以使学习的起点更接近于目标基站吞吐量最大化的最终解决方案。

在本文中，源基站通过不断地与环境进行交互并利用DQN来训练网络参数 w ，从而得到最佳的 Q 值。为了验证迁移学习的准确性，利用文献[13]的思想，本文采用策略模仿机制，确保源基站与目标基站之间的策略损失达到最小，状态-动作值 $Q_i(s, a)$ 的玻尔兹曼分布用来将源基站的深度学习网络转化为策略网络 $\pi_i(s, a)$ ，上述的策略模仿机制综合考虑了每个动作的 Q 值影响，源基站上的策略可以用 π_i 来表示

$$\pi_i(s, \bar{a}) = \frac{1}{Z_i} e^{\frac{1}{T} Q_i(s, \bar{a})}, Z_i = \sum_{\bar{a}' \in A_i} e^{\frac{1}{T} Q_i(s, \bar{a}')} \quad (18)$$

其中， T 表示温度参数， A_i 为源基站DQN的动作空间。相似地，目标基站的DQN中在 s 状态下的 Q 值估计也可以转化为策略网络

$$\pi_{TG}(s, a) = \frac{1}{Z_{TG}} e^{\frac{1}{T} Q_{TG}(s, a)}, Z_{TG} = \sum_{a' \in A_{TG}} e^{\frac{1}{T} Q_{TG}(s, a')} \quad (19)$$

源基站与目标基站处于同一个网络中，两者之间的策略模仿损失可以用交叉熵来进行衡量

$$H(w_i) = \sum_{a \in A_i} \pi_i(s, \bar{a}; w_i) \log_2 \pi_{TG}(s, a) \quad (20)$$

目标基站DQN的 Q 值估计表示为 $Q_{TG}(s, a)$ ，在策略损失的条件下， Q 值估计的更新规则为

$$\Delta H_i = [\bar{Q}^\pi(s_t, a_t) - Q_{TG}(s_t, a_t)] \frac{\partial Q_{TG}(s_t, a_t)}{\partial H_i} \quad (21)$$

其中， $\bar{Q}^\pi(s_t, a_t)$ 表示在策略 π 下 Q 值的无偏差估计值。迁移源基站策略的TLDQN算法如表2所示的算法2所示。

表2 算法2

算法2: 基于TLDQN的策略知识迁移算法

- (1) 初始化;
- (2) 源基站的DQN参数 w ，策略网络温度参数 T ，目标网络的DQN参数 w'
- (3) For 对于每个状态 $s \in \mathcal{S}$ ，源基站的动作 $\bar{a}$ ，目标基站可能采取的动作 a do
- (4) 执行算法1，得到估值网络的参数 w ，以及输出层对应的 Q 值函数
- (5) 根据式(18)将源基站上的 Q 值函数转化为策略网络 $\pi_i(\bar{a} | s)$
- (6) 根据式(19)将目标基站上的 Q 值函数转化为策略网络 $\pi_{TG}(a | s)$
- (7) 利用式(20)构建策略模仿损失的交叉熵 $H(w)$
- (8) 根据式(21)进行交叉熵的迭代更新,再进行策略模仿的偏导数的计算。
- (9) 直至目标基站选取出的策略达到 $Q_{TG}(s, a) \rightarrow Q^*_{TG}(s, a)$
- (10) End for
- (11) 目标基站获得对应的网络参数 w'
- (12) 执行算法1，目标基站得到最优资源分配策略

由于目标基站和源基站处于同一个网络中, 源基站的DQN很容易就可以得出函数 Q_i^* 的近似值和学习策略 π_i 。目标基站的在当前状态 s_0 得到的总奖励函数为 $R_{TG} = E \left(\sum_{t=1}^{\infty} \gamma^{t-1} r_t | s_0, \pi \right)$ 以及真实的值函数为 $Q_{TG}^*(s, a)$, R_{TG} 关于策略模仿损失的偏导数为 $\frac{\partial R_{TG}}{\partial H_i} = \sum_s Y^\pi(s) \sum_a \frac{\partial \pi_{TG}(s, a)}{\partial H_i} Q_{TG}^*(s, a)$, 其中 $Y^\pi(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi)$ 为折扣的转换概率。当满足 $\bar{Q}_{TG}^\pi(s, a) \approx Q_{TG}^*(s, a)$ 时, 根据策略损失的偏导数证明可以得到 $Q_{TG}(s, a) \rightarrow Q_{TG}^*(s, a)$, 对比于重新开始学习, TL可以避免由于随机选取动作导致的错误, 从而使目标基站在一开始就选出准确的资源分配策略。

4 仿真与讨论

在本节中, 通过与文献[14]中Q学习(Q-learning)算法和文献[15]中的注水(water-filling)算法的对比研究, 详细地介绍仿真结果来评估所建立算法的性能。

4.1 参数设置

本文设置的网络拓扑大小为 $500 \times 500 \text{ m}^2$, 1个HPN放置在网络中心位置, 15个RRH均匀分布在网络中, 用户均匀地分布在HPN和RRH上, HUE和RUE的用户数分别为10和90。系统的总带宽为5 MHz, 子载波数目 N 为32, 在仿真中, 仿真时隙 τ 取10 ms, 无线信道被建模为瑞利信道, HPN和RRH的路径损耗模型分别为 $31.5 + 40.0 \lg(d)$ 和 $31.5 + 35.0 \lg(d)$, d 以m为单位。网络上的随机业务数据包在每时隙服从均值为20的泊松分布, 数据包的大小为1 kbit/(s×packet)。噪声功率密度为-174 dBm/Hz, HPN和RRH的最大发射功率分别为43 dBm和30 dBm, HPN和RRH的初始发射功率都设置为15 dBm。此外, HPN和RRH的静态功率消耗分别为2 W和1 W。本文采用DRL解决无线资源的智能化分配问题, 还需要对神经网络参数进行训练。所选的 ε 贪心值为0.2, 学习率 α 为0.0001, Batch-size大小为32, 神经网络采用的优化器算法为Adam, Adam优化器中超参数 β_1 为0.9, β_2 为0.999。隐藏层的激活函数为ReLU, 神经元的dropout取0.2, 经验回放池的大小为10000, 目标网络参数更新的时间步长周期 δ 为500。

4.2 性能分析

本文通过DRL的训练讨论了估值网络中不同优化器对无线网络性能的影响, 从图5中可以得知本文采用的Adam优化器是最好的, 网络的总吞吐量为69 Mbps。当采用随机梯度下降(Stochastic Gradient Descent, SGD)优化器时, 无线网络吞吐

量的波动较明显。当采用均方根传播(Root Mean Square prop, RMSprop)优化器时, 引入了超参数 β , 消除了摆动幅度大的方向, 因此网络吞吐量更加稳定。但是较Adam优化器而言, 网络总吞吐量较低, 仅为63 Mbps。

图6的结果显示随着仿真时隙的增加, 平均队列长度起始迅速增加, 随后趋于稳定。这也说明了所提的DQN算法可以有效地保证系统队列稳定性, 满足系统时延的要求。与此同时, 随着到达率 λ 的增加, 平均队列长度会有一定的上升, 但最终都趋于一个稳定值。

由图7可知随着网络用户数的增加, 网络的总吞吐量逐渐增大, 而本文采用的DQN算法明显优于Q-learning和water-filling算法, 因为无线网络具有动态变化特性, 网络的状态空间非常庞大, 而DQN可以根据网络中用户的CSI和QSI进行智能的资源分配, 大大地提升了网络总吞吐量。图8也可以看出本文提出的DQN算法可以有效地降低平均队列时延, 使网络的稳定性得到了进一步地提升。如图9所示, 迁移学习下的平均队列长度值更小, 队列稳定所需的迭代次数更少。因为TL将DQN中源基站上估值网络的参数迁移到目标基站上, 目标基站的估值网络在迭代初期就可以取得较优的资源分配策略来应对网络的业务量变化。与此同时, 由于迁移学习采用策略模仿机制, 源基站上得到的策

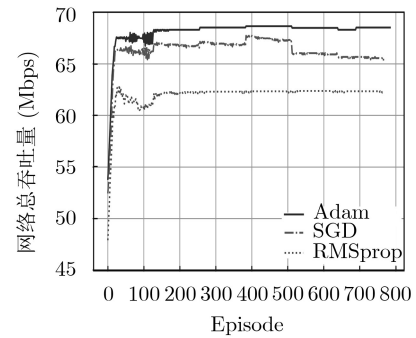


图5 DQN中不同优化器下的网络总吞吐量

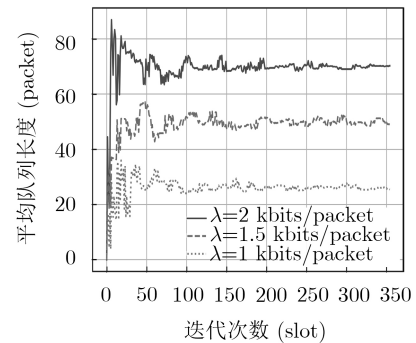


图6 不同到达率下的平均队列长度

略知识可以进一步指导目标基站选取较好的资源分配方案,所以迁移学习下的平均队列长度增长趋势更加平稳。

如图10所示,由于重用了源基站估值网络的权重参数,迁移学习下的损失函数值起点较DQN更

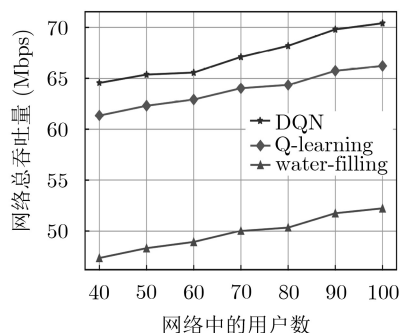


图7 网络用户数的总吞吐量

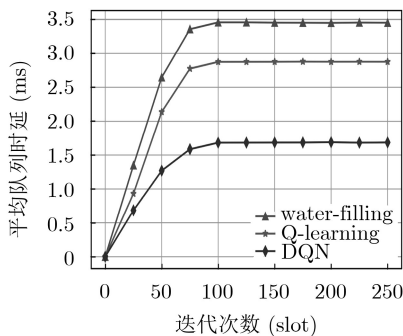


图8 网络的平均队列时延

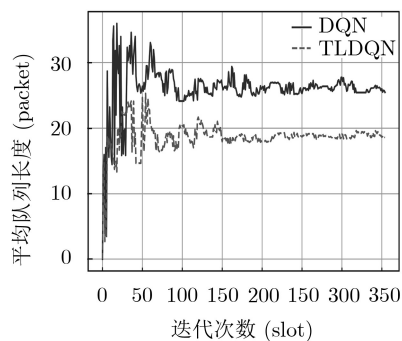


图9 迁移学习下的平均队列长度

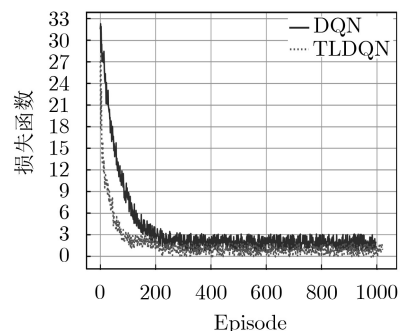


图10 迁移学习下的损失函数

低。以此同时,在迁移学习的指导下,TLDQN算法在100次Episode时算法就达到了收敛,而DQN算法在接近200次Episode时才开始收敛,因此,TLDQN较DQN重新开始训练模型而言,加速了算法的收敛速度。

5 结论

本文研究了H-CRAN中联合优化拥塞控制和无线资源分配的问题,将以最大化网络总吞吐量为优化目标的随机优化问题转化为深度Q学习过程。DQN算法基于网络中的CSI, QSI和基站的发射功率,在用户关联、子载波分配和功率分配方面进行快速、智能的分配。进一步地,利用TL的小样本学习特征,加快DQN的收敛速度。仿真结果表明,DQN和TLDQN在满足队列稳定的要求下,有效地提高了网络的吞吐量。

参考文献

- [1] MNIH V, KAVUKCUOGLU K, SILVER D, *et al.* Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529–533. doi: [10.1038/nature14236](https://doi.org/10.1038/nature14236).
- [2] SILVER D, HUANG A, MADDISON C J, *et al.* Mastering the game of go with deep neural networks and tree search[J]. *Nature*, 2016, 529(7587): 484–489. doi: [10.1038/nature16961](https://doi.org/10.1038/nature16961).
- [3] ZHANG Haijun, LIU Hao, CHENG Julian, *et al.* Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks[J]. *IEEE Transactions on Communications*, 2018, 66(4): 1705–1716. doi: [10.1109/TCOMM.2017.2763623](https://doi.org/10.1109/TCOMM.2017.2763623).
- [4] ZHANG Yuan, WANG Ying, and ZHANG Weidong. Energy efficient resource allocation for heterogeneous cloud radio access networks with user cooperation and QoS guarantees[C]. 2016 IEEE Wireless Communications and Networking Conference, Doha, Qatar, 2016: 1–6. doi: [10.1109/WCNC.2016.7565103](https://doi.org/10.1109/WCNC.2016.7565103).
- [5] HE Ying, ZHANG Zheng, YU F R, *et al.* Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks[J]. *IEEE Transactions on Vehicular Technology*, 2017, 66(11): 10433–10445. doi: [10.1109/TVT.2017.2751641](https://doi.org/10.1109/TVT.2017.2751641).
- [6] 唐伦, 魏延南, 马润琳, 等. 虚拟化云无线接入网络下基于在线学习的网络切片虚拟资源分配算法[J]. 电子与信息学报, 2019, 41(7): 1533–1539. doi: [10.11999/JEIT180771](https://doi.org/10.11999/JEIT180771).
- [7] TANG Lun, WEI Yannan, MA Runlin, *et al.* Online learning-based virtual resource allocation for network slicing in virtualized cloud radio access network[J]. *Journal of Electronics & Information Technology*, 2019, 41(7): 1533–1539. doi: [10.11999/JEIT180771](https://doi.org/10.11999/JEIT180771).
- [7] LI Jian, PENG Mugen, YU Yuling, *et al.* Energy-efficient

- joint congestion control and resource optimization in heterogeneous cloud radio access networks[J]. *IEEE Transactions on Vehicular Technology*, 2016, 65(12): 9873–9887. doi: [10.1109/TVT.2016.2531184](https://doi.org/10.1109/TVT.2016.2531184).
- [8] NEELY M J. Stochastic network optimization with application to communication and queueing systems[J]. *Synthesis Lectures on Communication Networks*, 2010, 3(1): 1–211. doi: [10.2200/S00271ED1V01Y201006CNT007](https://doi.org/10.2200/S00271ED1V01Y201006CNT007).
- [9] KUMAR N, SWAIN S N, and MURTHY C S R. A novel distributed Q-learning based resource reservation framework for facilitating D2D content access requests in LTE-A networks[J]. *IEEE Transactions on Network and Service Management*, 2018, 15(2): 718–731. doi: [10.1109/TNSM.2018.2807594](https://doi.org/10.1109/TNSM.2018.2807594).
- [10] SAAD H, MOHAMED A, and ELBATT T. A cooperative Q-learning approach for distributed resource allocation in multi-user femtocell networks[C]. 2014 IEEE Wireless Communications and Networking Conference, Istanbul, Turkey, 2014: 1490–1495. doi: [10.1109/WCNC.2014.6952410](https://doi.org/10.1109/WCNC.2014.6952410).
- [11] PAN S J and YANG Qiang. A survey on transfer learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345–1359. doi: [10.1109/TKDE.2009.191](https://doi.org/10.1109/TKDE.2009.191).
- [12] SUN Yaohua, PENG Mugen, and MAO Shiwen. 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. doi: [10.1109/JIOT.2018.2871020](https://doi.org/10.1109/JIOT.2018.2871020).
- [13] PAN Jie, WANG Xuesong, CHENG Yuhu, et al. Multisource transfer double DQN based on actor learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, 29(6): 2227–2238. doi: [10.1109/TNNLS.2018.2806087](https://doi.org/10.1109/TNNLS.2018.2806087).
- [14] ALQERM I and SHIHADA B. Sophisticated online learning scheme for green resource allocation in 5G heterogeneous cloud radio access networks[J]. *IEEE Transactions on Mobile Computing*, 2018, 17(10): 2423–2437. doi: [10.1109/TMC.2018.2797166](https://doi.org/10.1109/TMC.2018.2797166).
- [15] LI Yan, LIU Lingjia, LI Hongxiang, et al. Resource allocation for delay-sensitive traffic over LTE-Advanced relay networks[J]. *IEEE Transactions on Wireless Communications*, 2015, 14(8): 4291–4303. doi: [10.1109/TWC.2015.2418991](https://doi.org/10.1109/TWC.2015.2418991).
- 陈前斌：男，1967年生，教授，博士生导师，研究方向为个人通信、多媒体信息处理与传输、下一代移动通信网络等。
- 管令进：男，1995年生，硕士生，研究方向为网络功能虚拟化、无线资源分配、机器学习。
- 李子煜：女，1995年生，硕士生，研究方向为资源分配、机器学习。
- 王兆堃：男，1995年生，硕士生，研究方向为5G网络故障检测、自愈合、机器学习。
- 杨 恒：男，1993年生，硕士生，研究方向为网络切片及虚拟网络资源分配。
- 唐 伦：男，1973年生，教授，博士生导师，研究方向为新一代无线通信网络、异构蜂窝网络、软件定义无线网络等。