文章编号：2095-4980(2024)07-0792-08

基于深度强化学习的网络切片资源管理算法

王菲菲[1,2]，王　兰[1]，郑斯辉[3]，陈　翔[2,4]

(1.深圳大学 电子与信息工程学院，广东 深圳 518060；2.深圳清华大学研究院，广东 深圳 518057；3.清华大学 深圳国际研究生院，广东 深圳 528055；4.中山大学 电子与信息工程学院，广东 广州 510006)

摘　要：随着第五代通信技术(5G)的发展，各种应用场景不断涌现，而网络切片可以在通用的物理网络上构建多个逻辑独立的虚拟网络来满足移动通信网络多样化的业务需求。为了提高移动通信网络根据各切片业务量实现资源按需分配的能力，本文提出了一种基于深度强化学习的网络切片资源管理算法，该算法使用两个长短期记忆网络对无法实时到达的统计数据进行预测，并提取用户移动性导致的业务数据量动态特征，进而结合优势动作评论算法做出与切片业务需求相匹配的带宽分配决策。实验结果表明，相较于现有方法，该算法可以在保证用户时延和速率要求的同时，将频谱效率提高约7.7%。

关键词：5G；网络切片；深度强化学习；资源分配

中图分类号：TN929.5　　**文献标志码：**A　　**doi：**10.11805/TKYDA2022154

Resource management algorithm for network slicing based on deep reinforcement learning

WANG Feifei[1,2]，WANG Lan[1]，ZHENG Sihui[3]，CHEN Xiang[2,4]

(1.College of Electronic and Information Engineering，Shenzhen University，Shenzhen Guangdong 518060，China；
2.Research Institute of Tsinghua University in Shenzhen，Shenzhen Guangdong 518057，China；
3.Shenzhen International Graduate School，Tsinghua University，Shenzhen Guangdong 528055，China；
4.School of Electronics and Information Technology，Sun Yat-sen University，Guangzhou Guangdong 510006，China)

Abstract：With the development of the 5th Generation Mobile Communication Technology(5G), various application scenarios continue to emerge. Network slicing can construct multiple logically independent virtual networks on a common physical network to meet the diverse service requirements of mobile communication networks. In order to enhance the ability of mobile communication networks to allocate resources on demand according to the traffic of each slice, this paper proposes a network slicing resource management algorithm based on deep reinforcement learning. The algorithm uses two Long Short-Term Memory(LSTM) networks to predict statistical data that cannot be reached in real time, and extracts dynamic characteristics of business data volume caused by user mobility, and then makes bandwidth allocation decisions that match the needs of slice services in combination with the Advantage Actor-Critic(A2C) algorithm. Experimental results show that compared with existing methods, this algorithm can improve the spectral efficiency by about 7.7% while ensuring the user's delay and rate requirements.

Keywords：the 5th Generation Mobile Communication Technology(5G)；network slicing；deep reinforcement learning；resource allocation

在如今信息爆炸的时代，无线通信网络规模随着无线通信终端的激增而不断剧增。与此同时，以5G的3大业务场景即增强移动宽带(enhanced Mobile Broadband，eMBB)、大规模机器类型通信(massive Machine Type Communication，mMTC)和超高可靠超低时延通信(Ultra-Reliable and Low Latency Communication，URLLC)[1]为代

收稿日期：2022-08-22；**修回日期：**2022-10-10

基金项目：深圳市基础研究重点资助项目(JCYJ20200109143016563)

表，各种应用场景在速率、时延等方面要求不尽相同。因此运营商如果只构建一种网络，是难以同时满足所有场景需求的。为此，网络切片[2]技术应运而生，通过网络切片，运营商可以在通用的物理网络之上构建多个逻辑独立的虚拟网络，每个虚拟网络功能各异，从而灵活地应对不同服务需求。

网络切片由无线接入网、传输网、核心网子切片组合而成，并通过端到端切片管理系统进行统一的管理。本文主要研究接入网(Radio Access Network，RAN)层面的网络切片。在RAN网络切片中，核心问题是如何根据不同切片的服务等级协议（(Service Level Agreement，SLA)实现资源的按需配置，从而在保证各切片服务质量的同时，又能充分实现资源的高效共享。目前RAN网络切片的资源分配方案大致可以分为3类：静态资源分配、半静态资源分配和动态资源分配[3−4]。静态资源分配即根据经验、优先级或平均资源数为切片分配资源，并在整个生命周期内保持不变。半静态资源分配是在静态资源分配的基础上，预留出一部分资源，根据切片内的流量请求动态地将该预留资源进行分配。动态资源分配则考虑切片共享所有资源，根据流量请求与资源需求动态调整不同切片间的资源分配，该方案能够最大化系统资源利用率和保证更好的用户业务体验。从资源利用效率、服务定制的便利性和灵活性等角度考虑，对于5G以及Beyond 5G(B5G)移动通信网络而言，动态资源分配的潜力和优势都更大，值得更加深入地进行研究[5]。

针对网络切片的动态资源分配问题，传统方法往往通过严格的限定条件和推导形成特定优化问题，然后利用一系列优化算法或启发式方法进行求解[6−7]。然而，这些优化问题的灵活性与可扩展性有限，并且求解过程复杂，当环境参数(如切片数量、共享资源、用户发生移动等)或优化目标发生变化时，原有的求解方法可能并不再适用。强化学习(Reinforcement Learning，RL)可以通过与环境交互，观察环境的状态转移和获得的反馈变化来自主地学习最优的动作策略，尤其是结合了神经网络的深度强化学习(Deep Reinforcement Learning，DRL)具有很强的学习能力和灵活性，因此近年来涌现了不少基于DRL的网络切片资源管理研究工作。文献[8]采用深度Q网络(Deep Q Network，DQN)，以最大化平均用户体验质量(Quality of Experience，QoE)和频谱效率(Spectrum Efficiency，SE)为目标，实现了接入网切片无线资源的合理分配。文献[9]和[10]都在文献[8]的基础上做了改进，文献[9]将离散归一化优势函数引入到深度强化学习中，通过将Q值函数分离为状态值函数项和优势项，采用确定性策略梯度(Deterministic Policy Gradient，DPG)算法，避免了对每个状态动作都进行Q值计算，解决了文献[8]中频谱资源分配粒度过大的问题，并提高了神经网络训练时的收敛速度。文献[10]考虑随机性和噪声的影响，提出了基于生成对抗网络的深度分布式Q网络(GAN−DDQN)来学习动作值的分布，提高了算法的稳定性。

然而，上述工作均没有考虑无线通信网络中用户的移动性，这一因素将使得网络切片的动态资源分配问题更具挑战性。在文献[11]中，作者构建了一个基于长短期记忆(LSTM)网络和优势动作评论(Advantage Actor Critic，A2C)算法的LSTM−A2C框架，将特定时间窗口内的切片到达请求数视为状态，使智能体能够适配移动性带来业务请求动态性，在每个时隙均做出相对合理的带宽分配决策。尽管如此，在实际应用场景中，各切片业务请求数据的统计、获取和传输难以做到实时，存在一定的滞后性，这会导致LSTM−A2C算法做出的决策不再具备实时性，这种失配会严重影响网络资源的利用效率。为此，在本文中进一步提出了Doule−LSTM−A2C算法，通过额外引入一个LSTM网络对用户的移动性和切片请求信息进行预测，提高了在业务请求数据反馈存在时延情况下的系统性能；实验结果表明，Double−LSTM−A2C相较于不具备预测能力的LSTM− A2C算法，在切片请求信息存在多个时隙延迟的情况下，仍然能够在保证不同切片的平均用户QoE前提下，优化SE和系统效率(Utility)。

1　系统模型

1.1　系统场景

网络切片资源管理场景如图1所示，考虑单个基站且提供$\mathcal{N}=\{1,2,\cdots,N\}$种不同业务类型切片，其中第$n$种网络切片服务的用户集合记为$\mathcal{U}_n=\{u_1,u_2,\cdots,u_{M_n}\},n\in\mathcal{N}$。不同切片内用户使用不同的移动模式，每个切片上的所有用户采用轮询调度的方式获取切片分配到的资源。所有切片业务请求数量记为$d=\{d_1,d_2,\cdots,d_N\}$，其中d_n表示第n种切片类型总的业务请求数量，基站根据d_n的大小决定分配给切片n的带宽w_n。假设所有切片共享总频谱资源W，系统目标是找到最优的带宽资源分配方案，最大化系统效用。

1.2　系统目标

在本文中，系统的目标是尽可能满足各类切片的QoE以及提高SE。因此，可将优化目标定义为如下具有加权形式的效用函数：

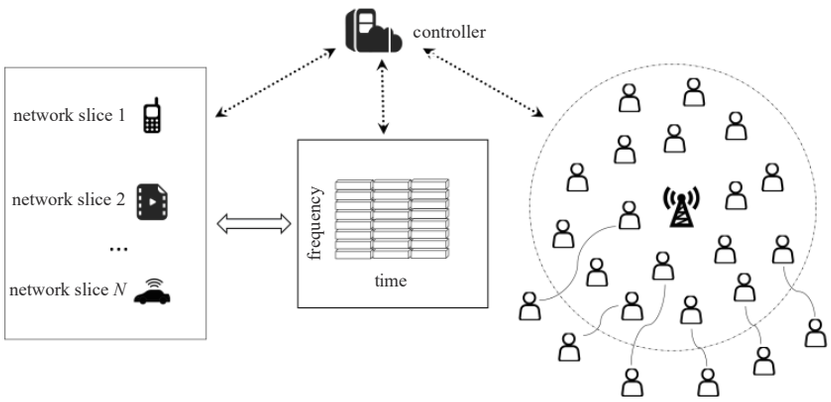

Fig.1 Scenario of resource management for network slicing
图1 网络切片资源管理场景

$$J=\alpha S+\sum_{n\in\mathcal{N}}\beta_n q_n \tag{1}$$

式中：S表示SE；q_n表示切片n的平均QoE，将在后文分别给出它们的表达式；同时，$\alpha\in(0,1)$和$\beta=[\beta_1,\beta_2,\cdots,\beta_N]$分别为SE和QoE的重要性系数。

$$S=\frac{\sum_{n\in\mathcal{N}}\sum_{u_n\in\mathcal{U}_n} r_{u_n}}{W} \tag{2}$$

式中r_{u_n}表示用户u_n的数据传输速率。对于同一切片内的用户，假设它们在不同时隙分别占用整个切片带宽进行通信，因此可以基于如下香农公式

$$r_{u_n}=w_n\log(1+R_{\mathrm{SN},u_n}),\forall u_n\in U_n \tag{3}$$

求得其通信速率，其中用户u_n的传输信噪比

$$R_{\mathrm{SN},u_n}=\frac{g_{u_n}P_{\mathrm{t}}}{N_0 w_n} \tag{4}$$

式中：g_{u_n}表示信号从基站到用户u_n经过路径损耗和阴影衰落后的平均信道增益；P_{t}表示基站的发射功率；N_0为单边噪声谱密度。

切片n的平均QoE则定义为成功传输(满足速率和时延要求)的数据包占基站传输总数据包的比例：

$$q_n=\frac{\sum_{u_n\in\mathcal{U}_n}\sum_{q_{u_n}\in Q_{u_n}} i_{q_{u_n}}}{\sum_{u_n\in\mathcal{U}_n} Q_{u_n}} \tag{5}$$

式中：Q_{u_n}表示基站和用户u_n之间传输的所有数据包集合；$i_{q_{u_n}}=1$表示数据包q_{u_n}传输成功，$i_{q_{u_n}}=0$表示该数据包传输失败。这里，数据包传输成功与否取决于传输过程是否满足用户的速率和时延要求。

基于以上分析，可将网络切片资源分配的优化目标表示为:

$$J=\alpha S+\sum_{n\in\mathcal{N}}\beta_n q_n=\alpha\frac{\sum_{n\in N}\sum_{u_n\in\mathcal{U}_n} r_{u_n}}{W}+\sum_{n\in N}\beta_n\frac{\sum_{u_n\in\mathcal{U}_n}\sum_{q_{u_n}\in Q_{u_n}} i_{q_{u_n}}}{\sum_{u_n\in\mathcal{U}_n} Q_{u_n}} \tag{6}$$

$$s.t.\sum_{n\in N} w_n=W \tag{7}$$

$$\sum_{u_n\in\mathcal{U}_n} Q_{u_n}=d_n \tag{8}$$

$$w_n=k_n\Delta,\ k_n\in\mathbb{Z}^+,\ n=1,2,\cdots,N \tag{9}$$

上述资源分配模型中，式(7)表示所有切片共享总资源W且在每个调度时隙均将总资源全部分给所有切片，

不剩余任何频谱资源。式(9)表示系统以带宽资源块(Resource Block，RB)Δ为最小单位进行分配，k_n表示分配给切片n的RB数。

值得注意的是，各切片的流量需求不仅取决于流量模型，而且还取决于用户在不同基站之间移动时的动态用户分布。然而，无论是流量需求还是用户分布都是先验未知的，这使得式(6)难以直接求解。为此，可以将上述问题转化为马尔可夫决策过程(Markov Decision Process，MDP)，采用RL来解决。发现只依赖当前时隙的流请求RL很难学到好的策略，同时考虑到状态信息的统计与收集存在滞后性，因此可以利用LSTM网络对滞后信息进行补偿，捕捉过去一定时间内用户请求的时间变化规律，并结合过去一段时间的历史请求数据，从而辅助RL寻找最优的资源分配策略。

2　基于深度强化学习的资源管理方案

2.1 马尔可夫决策过程建模

首先将网络切片的资源分配问题建模为MDP，MDP过程可用四元组$<\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R}>$表示，其中$\mathcal{S}$表示状态空间，$\mathcal{A}$表示动作空间，$\mathcal{P}$表示状态转移概率，$\mathcal{R}$表示奖励函数。状态转移概率$p(s_{t+1}|s_t,a_t)\in P$表示$t$时刻在状态$s_t$下执行动作$a_t$后转移到下一状态$s_{t+1}$的概率。由于此概率无法事先得到，因此本文将该模型建立为无模型MDP，并对状态空间、动作空间和奖励函数有如下定义：

状态空间$\mathcal{S}$：定义第t个调度周期(时隙)的观测向量$p_t=\{d_1^{(t)},d_2{}^{(t)},\cdots,d_N^{(t)}\}$为该周期内各切片数据包请求数量。在实际应用中，该观测向量的收集不可避免存在时延，因此在时刻t进行决策时实际上无法获得实时的观测向量。假设观测向量的收集时延为x，为了让智能体做出更准确的判断，时刻t的状态由已收集的最新T个时隙的历史观测状态$s_t{}^{(h)}=\{p_{t-T-x},\cdots,p_{t-1-x}\}$和$x$个时隙的预测量$s_t{}^{(p)}=\{p'_{t-x},\cdots,p'_{t-1}\}$组合而成，即$s_t=s_t{}^{(h)}\cup s_t{}^{(p)}$。

动作空间$\mathcal{A}$：由于带宽资源以最小单位Δ进行分配，故总共有$z=W/\Delta$个资源块，将这z个资源块无保留地分配给各个切片。假设从$\{1,2,\cdots,z\}$中随机选择3个和为z的整数，总共有M种可能的方案，将第m种方案记为a_m，则动作空间为$\mathcal{A}=\{a_1,a_2,\cdots,a_M\}$，每个调度周期$t$，智能体选择其中某个策略$a_t\in\mathcal{A}$作为动作执行。

奖励函数$\mathcal{R}$：奖励是最重要的部分之一，需要专门设计奖励函数指导智能体寻找最优策略。将奖励函数设计为：

$$r_t=\kappa_1(1-I)+I(\kappa_2+\kappa_3 S\cdot 1_{S\geqslant P_S}) \tag{10}$$

式中：$\kappa_1<0,\kappa_{2,3}>0$均为常值系数；$1_{S\geqslant P_S}$则是一个指示函数，当条件$S\geqslant P_S$成立时其取值为1，否则取值为0。

式(10)由两项组成，分别和QoE、SE有关系。其中，I是一个指示变量，当所有网络切片的QoE满足SLA要求时I=1，否则I=0，因此式(10)的结构表示：当有任意一个切片的QoE不满足要求时即$I=0$，都会带来一个惩罚κ_1，且关于SE的所有奖励都不会被考虑；当所有切片QoE都满足要求时，惩罚消失，同时视SE的大小给予奖励$\kappa_2+\kappa_3 S\cdot 1_{S\geqslant P_S}$，其中$\kappa_2$为基础性奖励，$P_S$是频谱效率的预期阈值，当$S$大于该值时，认为智能体选择的动作对于提高频谱效率具有实质性作用，因此给予额外奖励$\kappa_3 S$。基于这样的奖励函数，实际上希望引导智能体优先满足切片的QoE要求，在此基础上再尽可能提高SE。

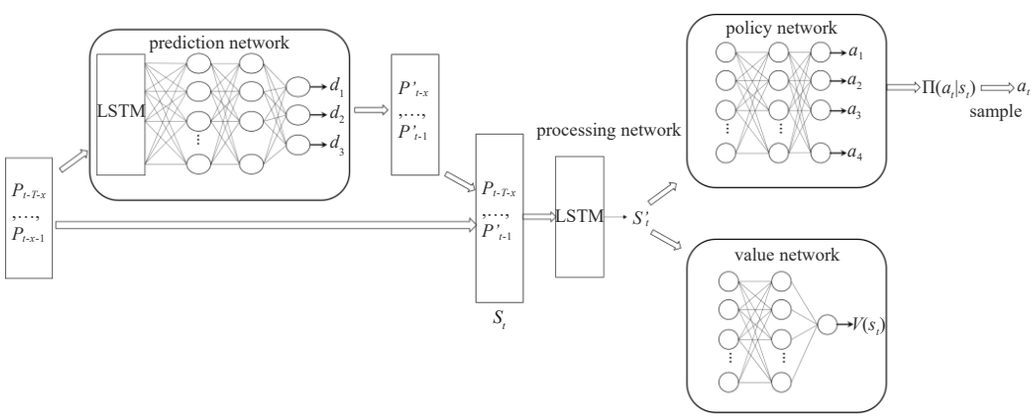

Fig.2 Overall framework of the algorithm

图2 算法整体框架

2.2 算法总体框架

算法的总体框架如图2所示，包括预测网络、数据处理网络、策略网络和值网络。

预测网络负责对因收集时延导致的无法获取的最近x个时隙的业务请求数据进行预测，由4层LSTM网络及2层全连接层组成。在t时刻，预测网络以历史状态$s_t^{(h)}$作为输入，递归地得到预测量$s_t^{(p)}$，此处的递归指首先基于$\{p_{t-T-x},\cdots,p_{t-1-x}\}$预测$p'_{t-x}$，进而基于$\{p_{t-T-x+1},\cdots,p_{t-1-x},p'_{t-x}\}$预测$p'_{t-x+1}$；以此类推。

数据处理网络作为观测变量和A2C网络之间的中间层，用于从观测变量中提取特征以便A2C网络能够更好地收敛，该网络由单层的LSTM组成，以历史和预测状态拼接成的完整状态$s_t=\{p_{t-x-T},\cdots p_{t-x-1},p'_{t-x},\cdots,p'_{t-1}\}$作为输入，输出LSTM的最后一个时间步的结果$s'_t$作为后续A2C网络的输入。

策略网络负责根据当前的状态特征s'_t选择相应的动作，由两层全连接层组成，分别使用tanh和softmax作为激活函数，输出所有动作的选择概率，并按照该概率分布随机采样出动作a_t，并结合值网络得到的状态价值，对策略进行评估，不断更新不同动作被选择的概率，直到收敛。

值网络负责对不同状态的价值进行评估，用于指导策略网络选择好的动作。该网络同样由两层全连接层组成，以状态特征s'_t为输入，最终输出对该状态价值的估计$V(s'_t)$。基于状态价值函数，可以计算出表征一个状态–动作对相对于平均状态–动作对的好坏程度的优势函数，如果优势函数为正，则增加当前动作被选取的概率，否则减小动作被更新的幅度，从而指导策略网络选取最优的策略。

2.3 损失函数与网络更新

在上述算法框架中，共包含3个不同的网络，每个网络的目标各不相同，均需要设置相应的损失函数并据此更新网络参数，从而引导其向预期的功能发展。下面介绍每个网络的损失函数和参数更新过程：

1) 预测网络

预测网络的训练目标是最小化预测值和真实值之间的误差。如前所述，在时隙t，基站共做出x次预测得到$\{p'_{t-x},\cdots,p'_{t-1}\}$；但由于在下一时隙$t+1$，基站仅可获取真实的$p_{t-x}$，因此只基于$t-x$时刻的预测误差对预测网络进行更新。假设预测值$p'_{t-x}=\{d'_1,d'_2,\cdots,d'_N\}$，真实值$p_{t-x}=\{d_1,d_2,\cdots,d_N\}$，则损失函数定义为二者的均方误差：

$$\mathrm{MSE}(p_{t-x},p'_{t-x})=\frac{1}{N}\sum_{i=1}^{N}(d_i-d'_i)^2 \tag{11}$$

并使用Adam优化器更新神经网络参数。

2) 策略网络

策略网络训练目标是优化动作选择策略，更新过程受优势函数指导。定义t时刻动作选择策略为参数θ_P的函数，记作$\pi(a_t|s_t;\theta_\mathrm{P})$；状态–动作价值函数为$Q(s_t,a_t)$，状态价值函数为$V(s_t)$，优势函数定义为：

$$A(s_t,a_t)=Q(s_t,a_t)-V(s_t)=E\left[r_t+V(s_t+1)\right]\approx r_t+\gamma V(s_t+1)-V(s_t)=\delta_t(s_t) \tag{12}$$

式中γ为未来潜在收益在当下的折扣因子。在智能体与环境进行交互中，如果策略探索不足，容易陷入局部最优解，学不到最优策略，因此为了鼓励探索，将熵正则化项$H(\cdot)$添加到参与者网络的损失函数中，以η作为熵值权重控制探索的程度。因此，策略网络的损失函数可定义为：

$$L(\theta_\mathrm{P})=-[\delta_t(s_t)\log\pi(a_t|s_t;\theta_\mathrm{P})+\eta H(\pi(a_t|s_t;\theta_\mathrm{P})] \tag{13}$$

式中负号是因为希望最大化优势函数，进而可用梯度下降法对网络参数进行更新：

$$\theta_\mathrm{P}\leftarrow\theta_\mathrm{P}+\delta_t(s_t)\frac{\partial\log\pi(a_t|s_t;\theta_\mathrm{P})}{\partial\theta_\mathrm{P}}+\eta\frac{\partial H\pi(a_t|s_t;\theta_\mathrm{P})}{\partial\theta_\mathrm{P}} \tag{14}$$

3) 值网络

训练目标是最小化估计状态值函数与真实状态值函数之间的误差，实际训

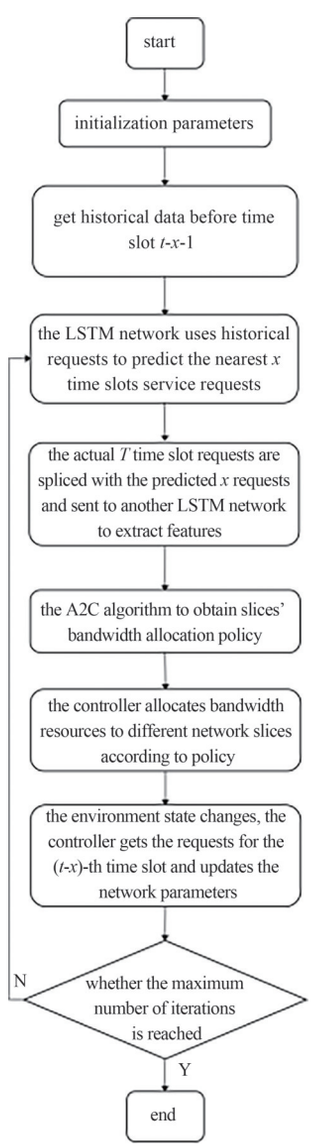

Fig.3 System workflow
图3 系统流程图

练过程中无法获取状态值函数的真值，因此用$r_t+\gamma V(s_{t+1})$进行近似，因而误差恰好就是式(12)中的$\delta(s_t)$。假设值网络是参数θ_{V}的函数$\mathrm{V}(s_t;\theta_{\mathrm{V}})$，显然误差也和$\theta_{\mathrm{V}}$有关系，可以将其也写为$\delta(s_t;\theta_{\mathrm{V}})$，则值网络的损失函数可以表示为：

$$L(\theta_{\mathrm{V}})=\frac{1}{2}\delta_t^2(s_t;\theta_{\mathrm{V}}) \tag{15}$$

同样采用梯度下降法对值网络的参数更新，即：

$$\theta_{\mathrm{V}}\leftarrow\theta_{\mathrm{V}}-\delta_t(s_t;\theta_{\mathrm{V}})\frac{\partial V(s_t;\theta_{\mathrm{V}})}{\partial\theta_{\mathrm{V}}} \tag{16}$$

基于以上损失函数和网络更新方法，本文所提的Double-LSTM-A2C算法流程图可总结如图3所示。

3　仿真结果与性能分析

3.1　仿真环境设置

在本节中，基于仿真结果对本文模型与算法的有效性进行评估和验证。考虑在1.2 km×1.2 km的区域内有一个基站，基站的覆盖半径为200 m，总带宽为10 MHz，带宽资源块大小为200 kHz；系统包括长期演进语音承载(Voice over Long-Term Evolution，VoLTE)、eMBB和URLLC共3种不同类型的服务/切片，共有800个用户，3种服务的用户比例为1:2:2，假设同一种服务/切片的用户具有相同的移动模式(速度和方向)，当用户移动至模拟区域的边界时，其方向将发生反转。各种类型切片所服务的用户设备(User Equipment，UE)详细参数配置如表1所示。

表1 网络切片具体参数设置

Table1 Parameter settings of the network slices

parameter	VoLTE	eMBB	URLLC
total number of users	160	320	320
user moving speed/(m/s)	6	6	6
distribution of packet arrival interval/ms	uniform [0,160]	truncated Pareto [shape parameter=1.2, mean=6, max=12.5]	exponential [mean=180]
distribution of packet size/Byte	constant [40]	truncated Pareto [shape parameter=1.2, mean=100, max=250]	constant [0.3M]
communication rate requirement/bps	51 k	100 M	10 M
transmission delay requirement/ms	10	10	1

在实验中，优化目标的权重系数α和β分别设置为0.01和[1, 1, 1]；每个切片内部不同用户基于0.5 ms的调度时隙轮询使用整个切片带宽；奖励函数中的常值系数κ_1、κ_2、κ_3和P_{S}分别设置为−5、4、0.01和160；业务请求数据收集延迟x(即预测观测变量的维度)设为5，状态变量的总维度设为15；预测网络、策略网络和价值网络的学习率分别设置为0.001、0.005和0.008，奖励折扣因子$\gamma=0.9$，熵值权重$\eta=0.001$；A2C神经网络的训练次数为5 000。为了验证算法的性能，本文以文献[11]中提出的不带预测的LSTM-A2C算法作为基准线进行性能对比。

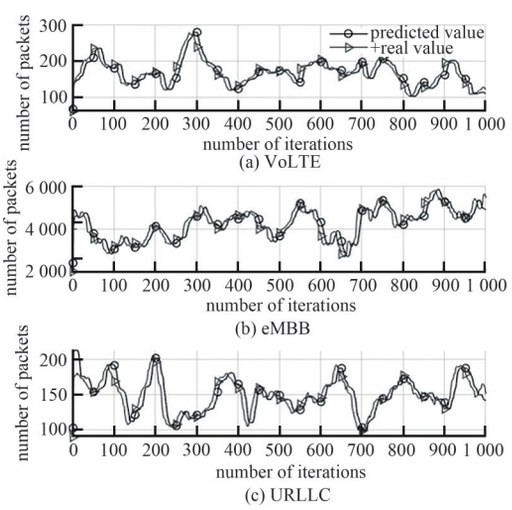

Fig.4 Comparison of predicted and true values

图4 预测值与真实值对比图

3.2　算法性能分析

图4(a)、4(b)、4(c)给出了基站内各类型服务真实请求值与预测值的对比。实验结果表明，对于VoLTE、eMBB和URLLC 3种切片，其请求规模因用户总数不同而存在差异，同时变化规律因用户移动模式不同而不同，但本文使用的LSTM预测网络均可以准确预测和跟踪，因而，当切片的业务请求数据具有滞后性的情况下，本文额外使用的LSTM预测网络

将有助于辅助强化学习进行实时决策。

图 5(a)、5(b)、5(c)为本文提出的 Double-LSTM-A2C 与不考虑预测补偿的 LSTM-A2C 算法平均用户 QoE 对比图。可见，两种算法均保证 3 种切片的平均 QoE 基本达到 1，尤其是 VoLTE 的要求较低，几乎在迭代开始的时候就已经满足了，这是因为在奖励函数上使保障 QoE 的优先级高于 SE。

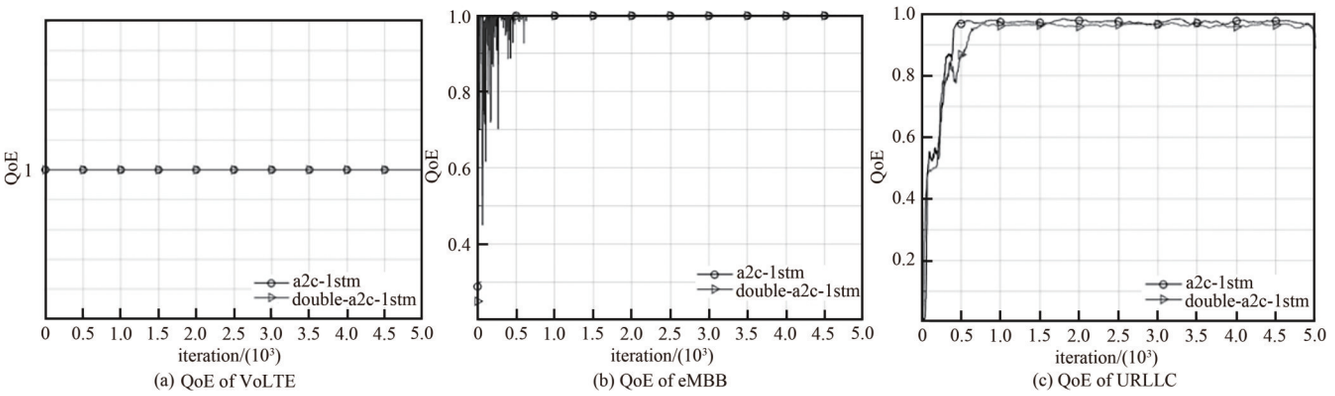

Fig.5 Comparison of different algorithms on average user QoE

图 5 不同算法平均用户体验质量对比图

图 6(a)、6(b)为两种算法的 SE 和系统效用性能对比图，从中可以明显看出，本文的算法性能显著优于 LSTM-A2C，这是因为当请求数据的收集存在滞后性时，无法准确捕捉当前时刻各切片的实时资源需求，导致强化学习无法学到最优的带宽分配策略；而本文所提出的算法可以较准确地预测因滞后性导致缺失的请求数据，从而使得状态处理网络可以准确地捕捉切片请求的变化规律，辅助强化学习学到最优的带宽分配策略，在资源有限的情况下，提升频谱利用效率。

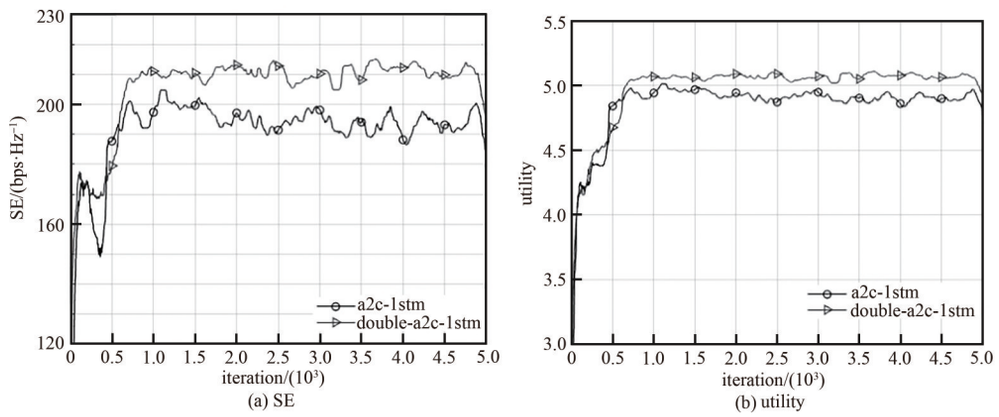

Fig.6 Comparison of different algorithms on SE and utility

图 6 不同算法的频谱效率和系统利用率对比图

4 结论

本文研究面向 5G/B5G 的网络切片动态资源分配问题，构建了由 LSTM 网络和深度强化学习 A2C 算法组成的 Double-LSTM-A2C 框架，其中两个 LSTM 网络可以捕获用户移动性导致的业务请求波动性变化的时间规律，同时克服网络切片请求数据的统计与收集存在滞后性对算法性能的影响；A2C 算法则可以在此基础上，学习在不同业务请求状态下做出合理的带宽资源分配策略，满足各切片的服务质量需求并提高系统效率。仿真结果表明，在频谱资源受限情况下，本文所提算法可以有效满足各切片平均用户体验质量要求，同时相较于不具备预测能力的算法，可以在业务请求数据收集存在滞后性的情况下取得更高的频谱效率和系统利用率。

参考文献：

[1] International Telecommunication Union. IMT Vision-Framework and overall objectives of the future development of IMT for 2020 and beyond:ITU-R M. 2083-0[S]. Geneva,Switzerland:ITU, 2015.

[2] FOUKAS X, PATOUNAS G, ELMOKASHFI A, et al. Network slicing in 5G: survey and challenges[J]. IEEE Communications

Magazine, 2017,55(5):94–100. doi:10.1109/MCOM.2017.1600951.

[3] 刘鹏. 面向5G网络切片无线资源分配[J]. 中国新通信, 2018,20(13):154–155. (LIU Peng. Radio resource allocation for 5G network slicing[J]. China New Telecommunications, 2018,20(13):154–155.) doi:10.3969/j.issn.1673–4866.2018.13.127.

[4] 粟欣,龚金金,曾捷. 面向5G网络切片无线资源分配[J]. 电子产品世界, 2017,24(4):30–32,40. (SU Xin,GONG Jinjin,ZENG Jie. Wireless resources allocation for 5G network slicing[J]. Electronic Engineering & Product World, 2017,24(4):30–32,40.) doi:10.3969/j.issn.1005–5517.2017.3.007.

[5] MARQUEZ C, GRAMAGLIA M, FIORE M, et al. How should I slice my network? A multi–service empirical evaluation of resource sharing efficiency[C]// Proceedings of the 24th Annual International Conference on Mobile Computing and Networking. New Delhi,India:Association for Computing Machinery, 2018:191–206. doi:10.1145/3241539.3241567.

[6] VO P L, NGUYEN M N H, LE T A, et al. Slicing the edge: resource allocation for RAN network slicing[J]. IEEE Wireless Communications Letters, 2018,7(6):970–973. doi:10.1109/LWC.2018.2842189.

[7] SUN Yao,FENG Gang,ZHANG Lei,et al. User access control and bandwidth allocation for slice–based 5G–and–Beyond Radio access networks[C]// 2019 IEEE International Conference on Communications(ICC). Shanghai, China: IEEE, 2019: 1–6. doi: 10.1109/ICC.2019.8761841.

[8] LI Rongpeng,ZHAO Zhifeng,SUN Qi,et al. Deep reinforcement learning for resource management in network slicing[J]. IEEE Access, 2018(6):74429–74441. doi:10.1109/ACCESS.2018.2881964.

[9] QI Chen,HUA Yuxiu,LI Rongpeng,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. doi:10.1109/LCOMM.2019.2922961.

[10] HUA Yuxiu,LI Rongpeng,ZHAO Zhifeng,et al. GAN–based deep distributional reinforcement learning for resource management in network slicing[C]// 2019 IEEE Global Communications Conference(GLOBECOM). Waikoloa,HI,USA:IEEE, 2019:1–6. doi: 10.1109/GLOBECOM38437.2019.9014217.

[11] LI Rongpeng,WANG Chujie,ZHAO Zhifeng,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. doi: 10.1109/LCOMM. 2020. 3001227.

作者简介：

王菲菲(1997–)，女，在读硕士研究生，主要研究方向为无线通信.email:wffarn@163.com.

郑斯辉(1997–)，男，在读博士研究生，主要研究方向为无线通信、联邦学习.

王　兰(1979–)，女，博士，讲师，主要研究方向为移动通信系统、无线资源管理.

陈　翔(1980–)，男，博士，教授，主要研究方向为无线与移动通信、卫星通信、物联网、软件无线电.