计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000078

基于深度强化学习的无人机辅助弹性视频多播机制

成昭炜[1,2]　沈　航[1,2]　汪　悦[1]　王　敏[1]　白光伟[1]
1 南京工业大学计算机科学与技术学院　南京 211816
2 南京大学计算机软件新技术国家重点实验室　南京 210093
(18052559504@163.com)

摘　要　文中提出了一个异构网络下无人机基站辅助的弹性视频多播机制。结合SVC编码,将无人机动态部署和资源分配问题联合考虑,目的是最大化用户整体的视频质量。考虑到宏基站覆盖范围内用户的移动会使网络拓扑结构发生改变,传统的启发式算法难以应对用户移动的复杂性。对此,采用基于深度强化学习的DDPG算法训练神经网络来决策无人机的最佳部署位置和带宽资源分配比重。在模型收敛后,学习代理可以在较短的时间内找到最优的无人机部署和带宽分配策略。仿真结果表明,所提方案达到了预期目标并且优于现有的基于Q-learning的方案。

关键词:可伸缩视频编码;多播;深度强化学习;无人机;移动互联网
中图法分类号　TP393

Deep Reinforcement Learning Based UAV Assisted SVC Video Multicast

CHENG Zhao-wei[1,2],SHEN Hang[1,2],WANG Yue[1],WANG Min[1] and BAI Guang-wei[1]
1 College of Computer Science and Technology,Nanjing Tech University,Nanjing 211816,China
2 State Key Laboratory for Novel Software Technology,Nanjing University,Nanjing 210093,China

Abstract　In this paper,a flexible video multicast mechanism assisted by the UAV base station is proposed. In combination with SVC encoding,the dynamic deployment and resource allocation of UAV are considered jointly in order to maximize the overall number of enhancement layers received by users. The traditional heuristic algorithm is difficult to deal with the complexity of user movement,considering that the user movement within the range of macro station will change the network topology. To this end,the DDPG algorithm based on deep reinforcement learning is used to train the neural network to decide the optimal location and bandwidth allocation proportion of UAV. After the model converges,the learning agent can find the optimal UAV deployment and bandwidth allocation strategy in a short time. The simulation results show that the proposed scheme achieves the expected goal and is superior to the existing scheme based on Q-learning.

Keywords　Scalable video coding(SVC),Multicast,Deep reinforcement learning,Unmanned aerial vehicles,Mobile Internet

1　引言

近年来,视频流量的快速增长导致无线网络资源的紧缺加剧。为了保证用户的视频质量,前人在异构网络的基础上做出了诸多尝试。多播是有效利用无线网络资源的技术之一[1],是一种同时将数据传输到一组终端设备的可行的有效的解决方案。多播使得请求同一视频资源的用户共享频谱资源。当多播组中的用户都能正确接收到数据时,多播组中信道条件最差的用户成为了制约系统性能的关键。为了满足不同用户的视频质量需求,将可伸缩视频编码(Scale Video Coding,SVC)技术引入到无线视频多播中。采用SVC编码将视频分为一个基础层和多个增强层。用户可以根据不同的信道条件接收增强层,信道条件好的用户可以接收基础层和更多的增强层。虽然引入多播和SVC编码能够有效利用网络资源,但不能减小宏基站的压力。

为了缓解宏基站(Macro Base Station,MBS)的压力,在异

到稿日期:2020-10-14　返修日期:2021-03-15　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:国家自然科学基金项目(61502230);江苏省自然科学基金项目(BK20201357);江苏省"六大人才高峰"高层次人才资助项目(RJFW-020);计算机软件新技术国家重点实验室资助项目(KFKT2017B21);江苏省研究生科研与实践创新计划项目(KYCX20_1079,SJCX20_0351);国家教育部2019年第二批产学合作协同育人项目(201902182003)
This work was supported by the National Natural Science Foundation of China (61502230),Natural Science Foundation of Jiangsu Province (BK20201357),Six Talent Peaks Project in Jiangsu Province (RJFW-020),State Key Laboratory Program for Novel Software Technology (KFKT2017B21),Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX20_1079,SJCX20_0351) and University-Industry Collaborative Education Program of the Ministry of Education(201902182003).
通信作者:沈航(hshen@njtech.edu.cn)

构蜂窝网络中引入了小型固定基站(small-cell base station)[2]，然而在现有的研究文献中小型基站的部署主要基于对通信量长期时空分布的预测。对于不可预知的时空分布，固定基站在服务移动用户时缺乏灵活性，导致用户整体的视频质量下降。近期的一些工作提出在常规网络中部署无人机基站(Drone-mounted Base Station，DBS)[3-5]，以提高无线网络的效率和灵活性[6]。在用户的位置难以预测和无法被宏基站覆盖的情况下，无人机基站可以提供支持。例如协助宏基站解决自然灾害和大型公共活动导致的网络拥堵等问题[6-8]。

不同于传统的小型固定基站，无人机移动基站能够更快、更廉价地部署。文献[9]考虑了用户对延迟的容忍和敏感程度，提出了一种无人机的三维定位算法，还研究了用户-基站关联和无线回程的带宽的分配问题，以最大程度地提高网络效用。文献[6]研究了无人机基站的下行覆盖性能，在无人机辅助的无线网络下，无人机的位置部署和轨迹设计影响着系统的整体覆盖性能。文献[10]研究了无人机辅助车辆网络中的多维资源管理问题，通过多接入边缘计算优化终端接入和资源划分。文献[11]允许无人机携带缓存，以最大限度提高网络吞吐量为目标，对无人机轨迹和缓存内容的放置进行决策。文献[12-13]提出了无人机辅助的边缘计算体系架构，用于支持延迟敏感的计算任务卸载应用。然而，现有的无人机部署和资源分配机制很少从用户移动角度统筹考虑无人机位置和资源分配，因此有必要设计终端设备移动性感知的无人机动态部署机制，以促进资源的优化配置。

本文提出了一种无人机基站辅助的弹性视频多播机制。基于 SVC 编码，将流媒体视频资源分割为多层，基础层由宏基站向多播组提供，增强层由宏基站和无人机基站联合提供。无人机位置和资源分配决定了无人机基站和宏基站的增强层覆盖效率。在基站覆盖范围内，为了最大化用户整体的增强层接收层数，综合无人机动态部署和资源分配，本文提出了联合优化问题，对面向 SVC 视频分发的资源分配和无人机部署联合优化问题进行建模。在求解优化问题时，本文考虑了传统启发式算法的计算复杂度和时延，设计了基于深度强化学习的 DDPG(Deep Deterministic Policy Gradient)算法并训练神经网络。该神经网络根据移动用户的位置分布进行决策，以获得无人机的位置和带宽资源分配。为提高训练稳定性、加快模型收敛和优化目标，本文提供了 3 种代表性的神经网络结构。仿真结果表明，该无人机部署和资源分配策略可以达到预期目标并且优于现有的强化学习算法方案。

2 系统模型

2.1 视频分发网络架构

如图 1 所示，本文考虑了一个由单个宏基站和单个无人机移动基站组成的异构无线网络。宏基站和无人机基站各自服务其覆盖范围内的多播组。将视频的 SVC 编码分为基础层和增强层两层，宏基站提供基础层和增强层，无人机基站为位置相对偏远的移动用户提供增强层，用户首先收到宏基站的基础层，再根据所处的位置和视频接收速率决定从属，接收从属的基站提供的增强层。

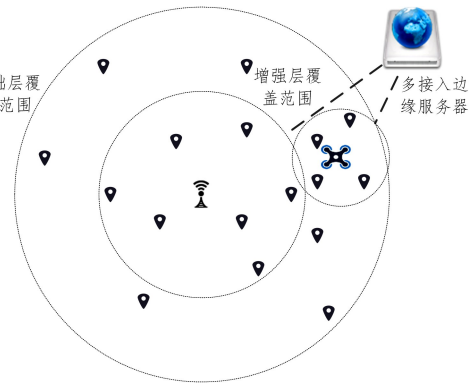

图 1 网络系统架构

Fig. 1 Framework of network system

宏基站覆盖范围内的用户随时间不断运动。在时间点 t，系统假定用户处于静止状态。随时间不断变化的用户分布可被分割为一个个连续时间点下的静态分布。通过当前的静态用户分布，系统根据当前用户位置和上一个时间点的环境状态来决策当前无人机的位置和资源分配策略。当进入下一个时间点 $t+1$ 时，用户位置分布发生改变，系统再次进行决策，从而为移动用户提供自适应服务。

设用户集合为 $\mathcal{N}$，总带宽资源为 B，系统首先分配带宽 B_b 来向用户多播组提供基础层，将剩余带宽资源 $B-B_b$ 分为 B_m 和 B_d，B_m 分配给宏基站投递增强层，B_d 分配给无人机基站投递增强层。

2.2 通信模型

无人机基站 d 和移动用户 i 之间的无线传播信道可以通过 LoS(Line of Sight)概率信道来建模，无人机基站和用户 i 之间 LoS 连接的概率为：

$$p^{(\mathrm{LoS})}=\frac{1}{1+\alpha\,\mathrm{e}^{-\beta(\theta_i-\alpha)}} \tag{1}$$

其中，θ_i 为 $\arctan\left(\frac{h_d}{v_i}\right)$，是用户 i 到无人机的仰角，h_d 是无人机的高度，v_i 是用户 i 与无人机之间的水平距离，α 和 β 为 Sigmoid 曲线参数。用户和无人机之间的 LoS 连接路径损失为：

$$\eta_i^{(\mathrm{LoS})}=\xi^{(\mathrm{LoS})}+\gamma^{(\mathrm{LoS})}\log_{10}\left(\sqrt{(v_i)^2+(h_d)^2}\right) \tag{2}$$

用户和无人机之间的 NLoS 连接路径损失为：

$$\eta_i^{(n\mathrm{LoS})}=\xi^{(n\mathrm{LoS})}+\gamma^{(n\mathrm{LoS})}\log_{10}\left(\sqrt{(v_i)^2+(h_d)^2}\right) \tag{3}$$

其中，$\xi^{(\mathrm{LoS})}$ 和 $\gamma^{(\mathrm{LoS})}$ 分别为 LoS 连接下参考距离的路径损耗补偿和路径损耗指数；$\xi^{(n\mathrm{LoS})}$ 和 $\gamma^{(n\mathrm{LoS})}$ 分别为 NLoS(None Line of Sight)连接下参考距离的路径损耗补偿和路径损耗指数；$\sqrt{(v_i)^2+(h_d)^2}$ 表示宏基站和用户 i 之间的三维距离。无人机基站 d 和用户 i 之间的平均路径损耗 $l_{d,i}$ 为：

$$l_{d,i}=p^{(\mathrm{LoS})}\cdot\eta^{(\mathrm{LoS})}+(1-p^{(\mathrm{LoS})})\cdot\eta^{(n\mathrm{LoS})} \tag{4}$$

信道增益 $g_{d,i}$ 为：

$$g_{d,i}=10^{-\frac{l_{d,i}}{10}} \tag{5}$$

3 无人机动态部署和资源分配联合优化问题

3.1 基础层资源分配

在宏基站覆盖范围内请求视频的用户都要从宏基站处获得基础层。设被请求视频资源基础层的接收速率为 γ_0，为了

节省分配的带宽并满足基础层接收速率的要求，可计算出投递基础层所要分配的最小带宽。令 $\eta_{m,i}$ 为宏基站 m 到用户 i 之间的平均路径损失。

$$\eta_{m,i}=\xi^{(\mathrm{LoS})}+\gamma^{(\mathrm{LoS})}\log_{10}(\sqrt{(z_{m,i})^2+h_m^2}) \tag{6}$$

其中，$z_{m,i}$ 是用户 i 与宏基站 m 的水平距离，h_m 为宏基站 m 的高度。宏基站与用户 i 之间的信道增益为：

$$g_{m,i}=10^{\frac{-\eta_{m,i}}{10}} \tag{7}$$

因为基础层多播组的信道增益 $g_m^{(\min)}$ 由该分组内信道增益最差的用户决定，因此有：

$$g_m^{(\min)}=\min_{i\in N}\{g_{m,i}\} \tag{8}$$

根据香农公式，投递基础层所需要的带宽为：

$$B_b=\frac{\gamma_0}{\log_2\left(1+\frac{p_m g_m^{(\min)}}{\sigma^2}\right)} \tag{9}$$

其中，p_m 为宏基站 m 的发射功率，σ^2 为高斯噪声。

3.2 增强层资源分配

增强层由宏基站和无人机基站联合提供，无人机基站为宏基站无法覆盖的用户提供服务。根据式(1)和式(2)以及香农公式，用户 i 到宏基站的信道容量为：

$$c_{m,i}=B_m\times\log\left(1+\frac{p_m g_{m,i}}{\sigma^2}\right) \tag{10}$$

根据式(4)，无人机基站 d 和用户 i 之间的信道增益为：

$$g_{d,i}=10^{-\frac{l_{d,i}}{10}} \tag{11}$$

根据香农公式可以计算出用户到无人机的信道容量为：

$$c_{d,i}=(B-B_b-B_m)\log\left(1+\frac{p_d g_{d,i}}{\sigma^2}\right) \tag{12}$$

其中，p_d 为无人机基站 d 的发射功率。

3.3 问题建模

为了最大化系统中用户整体接收的视频质量，需要获得无人机的最优部署位置和带宽分配比重，使覆盖范围内的用户整体收到的 SVC 层数最多。为便于实验和性能评估，本文考虑两层的 SVC 编码方式，保证所有请求视频资源的用户都收到基础层，那么优化目标转化为使用户整体收到的增强层数最多。对应的优化问题如下：

$$\max\sum_{i\in\mathcal{N}}(\beta_{m,i}+\beta_{d,i}) \tag{13}$$

$$\text{s.t } 0<B_d<B-B_b \tag{14}$$

$$x^{(\min)}\leqslant x_d\leqslant x^{(\max)} \tag{15}$$

$$y^{(\min)}\leqslant y_d\leqslant y^{(\max)} \tag{16}$$

$$z^{(\min)}\leqslant z_d\leqslant z^{(\max)} \tag{17}$$

$$\beta_{m,i}+\beta_{d,i}\leqslant 1 \tag{18}$$

$$\beta_{m,i}\in\{0,1\} \tag{19}$$

$$\beta_{d,i}\in\{0,1\} \tag{20}$$

约束条件(15)－(17)中，x_d，y_d 和 z_d 为无人机的三维坐标。约束条件(19)－(20)中，$\beta_{m,i}$ 和 $\beta_{d,i}$ 属于 0-1 变量。令 γ_1 为增强层的接收速率，$\beta_{m,i}=1$ 表示用户 i 可以收到来自宏基站 m 的增强层；反之表示未收到，即：

$$\beta_{m,i}=\begin{cases}1, & \text{if } c_{m,i}\geqslant\gamma_1\\ 0, & \text{otherwise}\end{cases} \tag{21}$$

而 $\beta_{d,i}$ 表示用户 i 能否收到无人机基站 d 的增强层。

$$\beta_{d,i}=\begin{cases}1-\beta_{m,i}, & \text{if } c_{d,i}\geqslant\gamma_1\\ 0, & \text{otherwise}\end{cases} \tag{22}$$

用户的位置随时间变化，且使用启发式算法在计算优化问题时会因重复运算带来极高的计算复杂度，本文采用深度强化学习算法来解决该问题。实验结果证明，在一定的约束条件下，通过足够时间的学习，该方法可以在离散的环境下获得最优解。在足够时间的训练后，学习代理可以在很短的时间内找到最优的无人机部署位置和带宽分配策略，这一特性对于解决用户的移动性问题来说至关重要。考虑到真实环境下用户分布的不稳定性，这种可以针对环境变化而迅速调优的能力十分重要。

4 基于 DDPG 的无人机部署和资源分配策略

4.1 基于 DDPG 的算法设计

本文提出基于 DDPG[14] 的 DDPG-UAV 算法(简称 DU 算法)来解决无人机动态部署和带宽分配问题。DU 算法是 Actor-Critic 算法的变种，优点在于能够在连续动作上进行更有效的学习。DU 算法包含 4 个网络：Critic 当前网络、Critic 目标网络、Actor 当前网络和 Actor 目标网络。目标网络是当前网络的复制，Actor 当前网络负责策略参数 θ 的更新，根据当前状态 S 选择当前动作 A，用于与环境交互生成下一个状态 S' 和奖励 R。Actor 目标网络负责根据重放缓存(replay buffer)中采样的下一状态 S' 选择下一个最优动作 A'，其网络参数 $\theta^{\mu'}$ 定期从 Actor 当前网络的参数 θ^{μ} 中复制。Critic 当前网络负责价值网络参数 θ^Q 的更新，计算当前的 Q 值 $Q(S,A,\theta^Q)$。Critic 目标网络负责计算目标 Q 值中的下一状态 S' 动作 A' 的 Q 值 $Q'(S',A',\theta^{Q'})$，目标 Q 值为 $R+\gamma Q'(S',A',\theta^{Q'})$。每次迭代后使用当前网络更新目标网络，更新采用软更新(soft update)的方式：

$$\theta^{Q'}\leftarrow\tau\theta^Q+(1-\tau)\theta^{Q'} \tag{23}$$

和

$$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'} \tag{24}$$

其中，τ 取值为 0.001。由于该场景下动作空间是连续的，构造一个探索策略 μ'，通过给动作策略添加噪声的方式来实现强化学习中的探索过程，本文沿用文献[14]中采用的 Ornstein-Uhlenbeck process[15] 来生成噪声。

本文场景下，Actor 网络以所有用户的二维位置信息 s_t 为输入。Critic 网络将用户的位置信息和 Actor 网络的输出动作作为输入，输出得分。算法的执行架构如图 2 所示。Reward(R)的设计采用增强层的宏基站和无人机基站服务率的加权平均的形式，计算式为：

$$R=\frac{(1-\rho)\cdot\sum_{i\in N}\beta_{m,i}+\rho\cdot\sum_{i\in N}\beta_{d,i}}{N} \tag{25}$$

其中，N 为用户的数量。为了鼓励模型探索更好的策略，给予 ρ 较大的比重，一般大于 0.5。实验证明，$\rho=0.6$ 时取得了最好的性能。模型训练的流程如算法 1 所示。

算法 1 DU 算法

1. 随机初始化 Critic 网络 $Q(s,a|\theta^Q)$ 和 Actor 网络的 $\mu(s|\theta^{\mu})$ 权重参数
2. 初始化目标网络 Q' 和 μ' 的权重 $\theta^{Q'}\leftarrow\theta^Q$，$\theta^{\mu'}\leftarrow\theta^{\mu}$
3. 初始化 replay buffer R 和用户环境 E
4. 获得观测到的初始用户分布，将用户二维位置信息归一化得到状态输入 s_1

5. for t=1→M do
6. 根据 Actor 网络和噪声生成动作 $a_t=\mu(s_t|\theta^\mu)+$noise
7. 执行动作并计算 reward r_t，观察新的状态 s_{t+1}
8. 将 (s_t, a_t, r_t, s_{t+1}) 存入缓存 R 中
9. 从 R 中随机采样 N 组数据组成 minibatch
10. 计算 $y_t=r_t+\gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'})$
11. 使用 smooth l1 损失函数最小化 y_t 和 $Q(s_t, a_t|\theta^Q)$ 的距离，并更新 Critic 网络参数 θ^Q
12. 更新 Actor policy
13. 更新目标网络：
$\theta^{Q'} \leftarrow \tau\theta^Q+(1-\tau)\theta^{Q'}$
$\theta^{\mu'} \leftarrow \tau\theta^\mu+(1-\tau)\theta^{\mu'}$；
14. 令 $s_t=s_{t+1}$
15. end

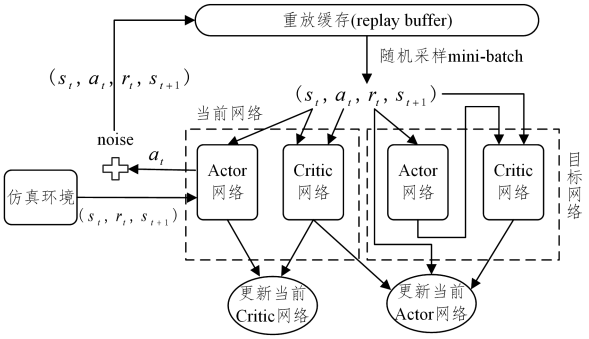

图 2 DU 算法的执行架构

Fig. 2 Execution architecture of DU

4.2 神经网络结构设计

Actor 网络的输入是归一化的用户二元位置信息 s_t。s_t 组织为一个三维矩阵，3 个维度分别表示批量数、用户位置的 x 坐标和 y 坐标。无人机的三维位置和增强层带宽分配比重作为输出的 action a_t，组织为一个五维矩阵，5 个维度分别表示批量数、无人机的 x 坐标 x_t、y 坐标 y_t、z 坐标 z_t 和带宽分配比重 ε_t。如图 3 所示，Actor 网络由 3 个网络单元结构 Actor Block 堆叠而成。Actor Block 是全连接层、批归一化层(BatchNorm)[16]和带泄露修正线性单元(LeakyReLU)连接而成的基本块结构。批归一化层的作用是在深度神经网络训练过程中使得每一层神经网络的输入保持相同分布，以保证训练的稳定性并缓解收敛慢的问题。LeakyReLU 是最常见的激活函数线性整流函数(ReLU)[17]的变体。实验证明，Actor 网络采用 LeakyReLU，性能优于 ReLU 激活函数。激活函数采用双曲正切函数(tanh)，将输出动作值的范围约束在(−1,1)之间。

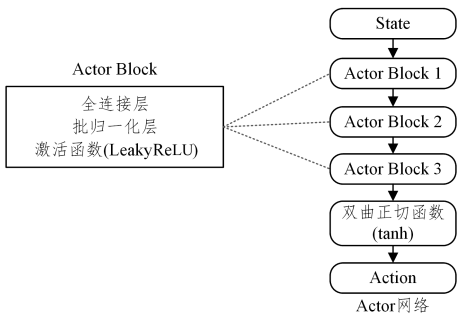

图 3 Actor 网络结构

Fig. 3 Framework of Actor network

图 4 给出了 Critic 网络结构。该网络的输入是归一化的用户位置信息 s_t 和动作 a_t。网络单元结构 Critic Block 与 Actor Block 类似，只是激活函数采用了 ReLU 函数。状态 s_t 通过一个 Critic Block 提取特征信息后，将特征信息和 a_t 进行连接操作，再将组合成的特征送到下一层。最终网络的输出是对当前用户状态 s_t 和动作 a_t 的评分，评分是一个二维矩阵，两个维度分别表示批量数和得分。

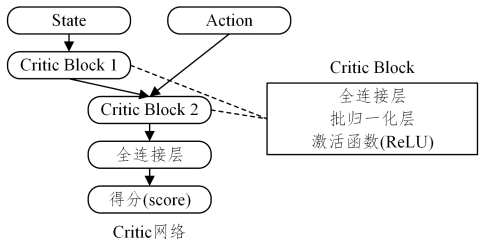

图 4 Critic 网络结构

Fig. 4 Framework of Critic network

Critic 网络的损失函数采用均方损失函数，Actor 网络和 Critic 网络均采用 Adam 优化器，Actor 网络的学习率为 0.001，Critic 网络的学习率为 0.0001。实验表明，采用该模型能够在稳定收敛的条件下取得较好的性能，增强层覆盖率优于传统地面基站的异构网络。为缓解在训练初期 Actor 网络输出的不稳定性，并保证模型能够稳定收敛，实验中对模型网络结构和超参数设计进行了探索。在上文描述的模型结构的基础上，提出了另外两种结构。

(1) DU-Sig：使用 Sigmoid 激活函数替换 Actor 网络的 tanh 激活函数；

(2) DU-LN：使用 LayerNorm 层来替换 BatchNorm 层以稳定训练过程[18]。

图 5 给出了 3 种方案训练近 1 万次迭代的平均 rewards 的变化趋势。实验表明，3 种方案都能在 1 万次迭代内有效收敛。DU-Sig 网络结构虽然比 DU 更快收敛，但平均 rewards 远远落后于其他两种结构。DU-LN 网络结构的 rewards 训练曲线相比 DU 更加光滑，更早收敛，但最终平均 rewards 略低于 DU 结构。

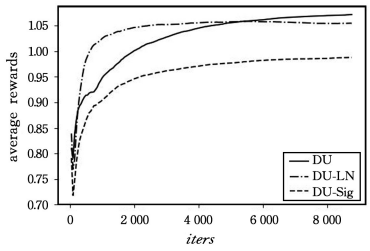

图 5 训练趋势

Fig. 5 Training trend

5 性能评价

5.1 实验设计

利用仿真环境验证所提文算法的性能，考虑一个宏基站同无人机小基站协同工作的场景。实验开始前用户位置分布

服从泊松点过程，用户的移动遵循上文提出的 Random Walk 模型，暂不考虑用户出入宏基站的覆盖范围，无人机不会飞出宏基站覆盖范围。宏基站的下行发射功率为 46 dBm，无人机基站的发射功率为 24 dBm，表 1 列出了重要仿真参数。

表 1　实验环境的参数设置

Table 1　Experimental parameters

参数名称	参数值/范围
宏基站高度h_m/m	10
路径损耗补偿(LoS)ζ_{los}/dB	103.4
路径损耗指数(LoS)γ_{los}/(dB/km)	24.2
路径损耗补偿(NLoS)$\zeta_{n\text{los}}$/dB	131.4
路径损耗指数(NLoS)$\gamma_{n\text{los}}$/(dB/km)	42.8
噪声σ^2/dBm	−104
宏基站传输功率p_m/dBm	46
无人机基站传输功率p_d/dBm	24
Sigmoid 曲线参数 α,β	11.95,0.136
无人机高度区间$h_{\min}$/m,$h_{\max}$/m	20,50

为了客观地评估所提方案的性能，将该方案和常见的基于 Q-learning[19-20] 的方法(命名为 QL)进行比较。在相同的神经网络基本架构下，将训练完成的模型在仿真环境下迭代 1 万次之后，比较各个模型能接收到基础层和增强层的平均用户数。为了方便统计和计算，在模拟环境下，用户数量固定为 50。

从图 6 中可以看出，采用 DU 算法的模型性能明显优于常见的采用 QL 算法的模型。对比接收宏基站增强层的用户数量，基于 DU 算法的 3 种方案与基于 QL 算法的差别较小。但对比接收无人机基站增强层的用户数，基于 DU 算法的 3 种结构远强于基于 QL 的算法。

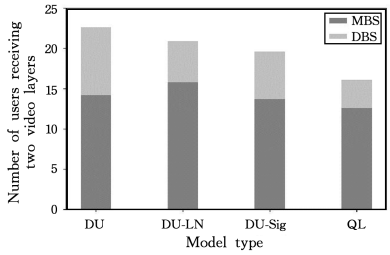

图 6　接收到两层的平均用户数

Fig. 6　Number of users receiving two video layers

在仿真环境下用不同模型运行 1 万次迭代后，统计用户接收视频图像的峰值信噪比(PSNR)。在基础层数据率为 180 KBPS，增强层数据率为 440 KBPS 的情况下，平均峰值信噪比的核密度估计(Kernel Density Estimation，KDE)和累积分布函数(Cumulative Distribution Function，CDF)如图 7 所示。从 PSNR 核密度估计图中可以看出，DU 模型的 PSNR 主要分布于 36.65 dB 和 37.25 dB 之间，而 QL 模型的 PSNR 主要分布于 36.35 dB 和 37.00 dB 之间，采用 DU 方案在用户接收视频质量的分布上优于基于 Q-learning 的方案。这是由于本文提出的神经网络直接决策下一个时间点无人机的位置，相比基于 QL 的模型，决策无人机的动作更加准确，使无人机的部署更加合理。

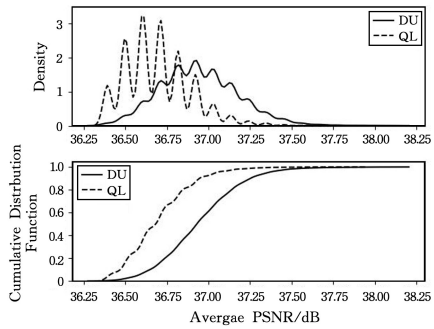

图 7　平均 PSNR 的核密度估计和累积分布

Fig. 7　KDE and CDF of average PSNR values

5.2　性能分析

在神经网络的训练和超参数的调试过程中可以发现，在网络结构不变的情况下，影响模型收敛和性能的瓶颈是超参数 ρ。当超参数设置过小时，有可能导致模型无法探索更优的策略，从而将带宽的绝大部分分配给宏基站以提供增强层服务，使无人机基站处于无法服务任何用户的空转状态。当超参数设置过大时，模型在训练过程中难以收敛且输出严重单一化。图 8 给出了模型在不同超参数 ρ 下的性能。

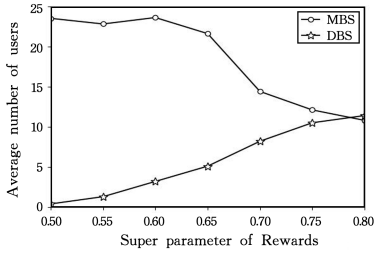

图 8　Reward 超参数性能比较

Fig. 8　Comparison of hyper parameter performance

将模型设置不同的超参数 ρ 后，分别训练 1 万次，以获得各个超参数下模型增强层的用户平均覆盖数。由图 8 可知，较小的超参数 ρ 下无人机的平均覆盖数很小。当 ρ 接近 0.8 时，虽然无人机用户服务数和宏基站用户服务数相近，但模型训练无法收敛。为了鼓励模型探索更好的无人机部署位置并保证稳定收敛，将权重 ρ 设为 0.6。

图 9 给出了测试数据中几个时间点无人机的三维坐标和覆盖半径，以宏基站位置为坐标中心。覆盖半径是由当前时间点能接收到无人机增强层的最远用户与无人机的水平距离决定的。从图中的数据可以看出，当用户随时间移动时，无人机调整自身位置为覆盖范围内的用户提供增强层服务。

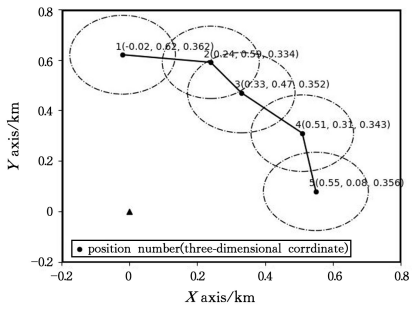

图 9　无人机轨迹

Fig. 9　Drone moving trajectory

图 10 给出了迭代 10 万次之后的无人机位置的分布热度图。图 10 中，将宏基站覆盖范围的二维空间分成 32 * 32 的网格，每个网格的长宽均为 50 m。实验统计了每个网格范围内无人机的数量后生成了这张热度图，从图中可以看到，无人机大部分位置都落在了宏基站覆盖范围的靠边缘区域，以达到为远离宏基站的移动用户提供服务的目标。

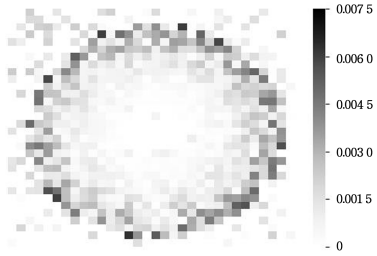

图 10 无人机分布热度

Fig. 10 Drone distribution heatmap

为了便于观察飞行高度对性能的影响，这里同时给出了飞行高度固定时的仿真结果。如图 11 所示，无人机飞行高度被分别固定为 20 m，30 m，40 m，45 m 和 50 m(命名为 DU-1，DU-2，DU-3，DU-4 和 DU-5)。固定无人机高度将直接影响从无人机处接收增强层的用户数量，实验表明无人机在 40～45 m 之间移动能服务较多用户，但明显低于自适应无人机高度的 DU 模型。在这些模型中，从宏基站接收增强层的用户数变化不大，而 DU 模型因为将更多带宽资源分配给无人机基站，所以宏基站服务的用户数略少，但不影响整体性能的优异。

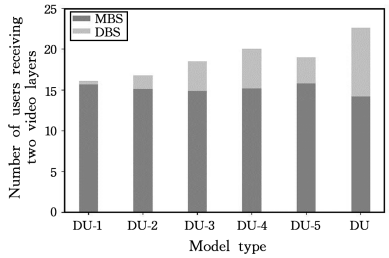

图 11 不同无人机高度下模型性能的对比

Fig. 11 Model performance under different UAV heights

降低延迟主要依赖于合理的资源分配，在考虑公平性的基础上尽可能降低传输延迟。图 12 给出了不同策略传输单位体积文件所耗费的延迟。从图中可以看出，本文提出的 DU 模型具有更低的数据传输延迟，这得益于多播和 SVC 编码灵活复用带宽资源，使用户整体尽可能多地接收增强层，获得更快的传输速率，从而降低了单位体积的数据传输延迟。

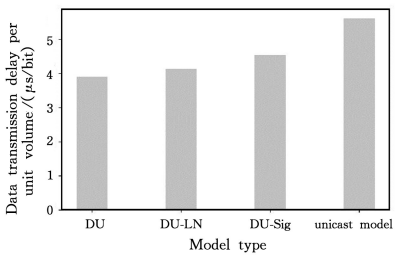

图 12 不同策略的传输延迟

Fig. 12 Transmission delay for different policies

结束语 本文提出了一种无人机基站辅助的弹性视频多播机制。在无线网络中，将无人机基站和 SVC 多播相结合，研究了无人机三维空间位置部署和带宽资源分配的联合优化问题。在基站覆盖范围内，最大化用户整体增强层的接收层数。采用 DU 算法训练神经网络，根据用户位置进行决策，获得无人机位置和带宽分配。仿真结果表明，基于深度强化学习的无人机辅助弹性视频多播机制可以根据不断变化的用户分布调整无人机的位置，为部分移动用户提供增强层服务，其增强层覆盖率优于基于 QL 算法的方案。

本文方案未考虑利用无人机的机动性进行动态视频内容缓存，并将缓存策略融入到机器学习模型中。此外，在多无人机环境下，如何使无人机避免干扰和碰撞，协调无人机动态部署以最大化用户整体接收的视频质量也是后续研究要解决的问题。

参 考 文 献

[1] ARANITI G, CONDOLUCI M, SCOPELLITI P, et al. Multicasting over emerging 5G networks: Challenges and perspectives [J]. IEEE Network, 2017, 31(2): 80-89.

[2] AGIWAL M, ROY A, SAXENA N. Next generation 5G wireless networks: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1617-1655.

[3] GHOSH A, MANGALVEDHE N, RATASUK R, et al. Heterogeneous cellular networks: From theory to practice[J]. IEEE Communications Magazine, 2012, 50(6): 54-64.

[4] BOR-YALINIZ R I, EL-KEYI A, YANIKOMEROGLU H. Efficient 3-D placement of an aerial base station in next generation cellular networks[C]// 2016 IEEE International Conference on Communications (ICC). IEEE, 2016: 1-5.

[5] GUO W, DEVINE C, WANG S. Performance analysis of micro unmanned airborne communication relays for cellular networks [C]// 2014 9th International Symposium on Communication Systems, Networks & Digital Sign (CSNDSP). IEEE, 2014: 658-663.

[6] MOZAFFARI M, SAAD W, BENNIS M, et al. Drone small cells in the clouds: Design, deployment and performance analysis [C]// 2015 IEEE Global Communications Conference (GLOBECOM). IEEE, 2015: 1-6.

[7] BOR-YALINIZ I, YANIKOMEROGLU H. The new frontier in RAN heterogeneity: Multi-tier drone-cells[J]. IEEE Communications Magazine, 2016, 54(11): 48-55.

[8] DERUYCK M, WYCKMANS J, MARTENS L, et al. Emergency ad-hoc networks by using drone mounted base stations for a disaster scenario[C]// 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2016: 1-7.

[9] KALANTARI E, BOR-YALINIZ I, YONGACOGLU A, et al. User association and bandwidth allocation for terrestrial and aerial base stations with backhaul considerations[C]// 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). IEEE, 2017: 1-6.

[10] PENG H, SHEN X. Multi-agent reinforcement learning based

resource management in MEC- and UAV-assisted vehicular networks[C]// IEEE Journal on Selected Areas in Communications. 2021:131-141.

[11] WU H,LYU F,ZHOU C,et al. Optimal UAV caching and trajectory in aerial-assisted vehicular networks: A learning-based approach[C]// IEEE Journal on Selected Areas in Communications. 2020:2783-2797.

[12] CHENG N,LYU F,QUAN W,et al. Space/aerial-assisted computing offloading for IoT applications: A learning-based approach[J]. IEEE Journal on Selected Areas in Communications, 2019,37(5):1117-1129.

[13] ZHOU C,WU W,HE H,et al. Delay-aware iot task scheduling in space-air-ground integrated network[C]// IEEE GLOBECOM. 2019:1-6.

[14] LILLICRAP T P,HUNT J J,PRITZEL A,et al. Continuous control with deep reinforcement learning[J]. arXiv: 1509.02971,2015.

[15] StackExange. Implementing Ornstein-Uhlenbeck in Matlab [OL]. (2017-09-22) [2020-05-20]. https://math.stackexchange.com/questions/1287634/implementing-ornstein-uhlenbeck-in-matlab.

[16] ROTA BULÒ S,PORZI L,KONTSCHIEDER P. In-place activated batchnorm for memory-optimized training of dnns[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:5639-5647.

[17] GLOROT X,BORDES A,BENGIO Y. Deep sparse rectifier neural networks[C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence Andstatistics. 2011: 315-323.

[18] BA J L,KIROS J R,HINTON G E. Layer normalization[J]. arXiv:1607.06450,2016.

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

[20] MNIH V,KAVUKCUOGLU K,SILVER D,et al. Playing atari with deep reinforcement learning[J]. arXiv:1312.5602,2013.

CHENG Zhao-wei, born in 1995, postgraduate. His main research interests include space-air-ground integrated networks and so on.

SHEN Hang, born in 1984, Ph.D, associate professor. His main research interests include network slicing and space-air-ground integrated networks.