IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MOBILE MULTIMEDIA: METHODOLOGY AND APPLICATIONS

Received November 16, 2019, accepted December 8, 2019, date of publication December 17, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960191

Deep Reinforcement Learning-Based Video Quality Selection and Radio Bearer Control for Mobile Edge Computing Supported Short Video Applications

WENJUN WU, (Member, IEEE), YANG GAO, TIANQI ZHOU, YINHUA JIA, HAO ZHANG, TINGTING WEI, AND YANG SUN, (Member, IEEE)
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Corresponding author: Yang Sun (sunyang@bjut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U1633115, in part by the Science and Technology Foundation of Beijing Municipal Commission of Education under Grant KM201810005027, in part by the Beijing Post-Doctoral Funding under Project Q6042001201903, and in part by the Chaoyang District Post-Doctoral Funding under Project Q1042001201901.

ABSTRACT With the rapid development of mobile communication technology, short video applications, which combine the features of both social and multimedia applications, have become more and more popular. However, the transmission of short videos poses great challenges to the existing mobile networks. In this paper, mobile edge computing is adopted to provide content caching of short videos close to end users. To improve the quality of service, we take both the quality level of the video and the long-term wireless network transmission performance into consideration. The joint video quality selection and radio bearer control optimization problem is formulated as a Markov decision process, aiming at maximizing the long-term video quality profit and minimizing the cost of bearers and the penalty of latency. Deep reinforcement learning is used as the solution and the policy gradient based quality selection and radio bearer control method is proposed. The REINFORCE algorithm with baseline is used to train the policy network, and the episodic simulations are built to obtain the training samples. Different weight coefficients of the objective function are configured. Training results show that the proposed method can achieve the best accumulated value among all the comparison methods. When the weight coefficients are changed, the training processes can lead the policy networks to obtain proper trade-off between different objective factors. Moreover, the performance of the trained policy network is evaluated with different short video request arriving rates. Testing results show that the proposed method performs well when the arriving rates vary in a certain range.

INDEX TERMS Deep reinforcement learning, video quality selection, radio bearer control, short video applications, mobile edge computing.

I. INTRODUCTION

With the rapid development of 5G technology, multimedia services and applications have experienced unprecedented explosive growth, e.g. Internet Protocol Television (IPTV), social networks, mobile multimedia, immersive multimedia and virtual reality games, etc [1]. At present, these multimedia applications have become an indispensable part of daily life and are expected to grow exponentially. At the same time, the current popular multimedia applications, especially the short video applications, will pose great challenges to the service capabilities of existing mobile networks. In order to solve the above challenges, it is necessary for multimedia service providers to develop and utilize various new technologies to meet the quality of service (QoS) of users.

The huge traffic demand brought by the growth of multimedia services will not only bring tremendous pressure to the radio access network, but also to the backhaul network

The associate editor coordinating the review of this manuscript and approving it for publication was Dapeng Wu.

connecting the base station (BS) and the core network. Therefore, the improvement of the wireless resource utilization efficiency and the alleviation of backhaul congestion have become very critical research issues. With the emergence of mobile edge computing (MEC), the local service processing can be realized in traditional radio access networks. It can provide services with high bandwidth and low latency, effectively reduce the traffic load of the core network and improve QoS [2], [3]. These advantages brought by MEC are critical to multimedia streaming services. By introducing the concept of content caching into MEC, the industry can effectively reduce access latency and traffic congestion on backhaul links by caching contents at MEC servers in advance, thus improve the network performance and the QoS performance of users [4].

Multimedia service distribution and resource allocation supported by MEC have been widely researched and discussed [5]–[19]. To guarantee the high multimedia service performances for users, some researchers investigate the cache deployment and cache decision optimization problems to reduce the service load of base station effectively [5]–[11], while some researchers focus on designing cooperative and efficient video transmission mechanisms to improve the video distribution performance [12]–[17]. To improve the service quality of video streaming service, the authors in [18] design an effective caching scheme and hybrid video transmission mechanism by jointly taking users' similarity, sharing willingness, location distribution and user requirements into consideration. To offload and alleviate the heavy traffic load, the authors in [19] jointly optimize the cache size allocation and multicast beamformer at the centralized processor to minimize the expected file downloading time and maximize the expected file downloading rate. The related resource allocation algorithms that take video quality into account generally based on the traditional methods such as ant colony optimization [7], game theory [9], simulated annealing [14] and dichotomy [17]. However, most of the existing works only consider the optimization problem from the aspect of MAC layer and physical layer, taking the content and the cache resources into account. Besides, most of them assume the wireless transmission performance requirements are predefined according to the characteristics of the multimedia source data and the upper layer bearer control.

In reality, the QoS of multimedia service is a long-term performance indicator in most cases. Take the short video application as an example. The quality level of the video and the wireless network transmission performance are integrated together affecting the QoS [20]. Since the Radio Bearer Control (RBC) is a relatively long-term wireless network transmission performance guarantee, it should be addressed. Therefore, the joint optimization of video quality selection and RBC is important for improving QoS. It can also provide guidance to more detailed wireless resource allocation crossing MAC layer and physic layer. But this joint optimization problem is a very complex long-term decision-making problem, which has not been fully studied.

Generally, the solutions of the long-term decision-making problem are dynamic programming or other mixed-integer programming methods. However, the solution space and the complexity of the problem are relatively large, so using traditional optimization method to solve it becomes very difficult. Deep reinforcement learning (DRL) turns out to be an effective method to solve the above challenge which only needs to learn from the sample data of the environment in advance and makes decisions quickly when it is applied [21], [22]. The application of DRL in the optimization of wireless network has become one of the hottest research fields, which is widely used to solve the problem such as scheduling [23], relay selection [24], traffic offloading [25] and network slicing [26].

In this paper, the MEC supported short video applications are addressed. Assume the original source files of copyrighted short videos are cached at the MEC server. When a UE request for a video, only the authentication information is sent to the remote application server, and the MEC server can transmit the video data with low latency. The problem of quality selection and RBC for sequentially arriving requests of videos is formulated as a Markov Decision Process (MDP). The objective is to maximize the long-term video quality profit while minimizing the cost of bearers and the penalty of latency. The policy is defined as a mapping from the system states to the quality selection and RBC decision at each time step. A DRL based solution architecture is proposed with the policy represented by a deep neural network. To train the policy network, the episodic simulations are built, and the REINFORCE algorithm with baseline is used. The complexity of the proposed method is analyzed, and the test results with different environment parameters confirm the effectiveness and the generalization ability of the proposed method.

The main contributions of this paper are summarized as follows.

1) We model the MDP corresponding to the video quality selection and RBC problem for sequentially arriving user requests. The radio bearer (RB) setup, RB reconfiguration and RB release in RBC are considered. These different RBC procedures are modeled as different actions of the MDP. Base on the proposed MDP model, the long-term system performance can be optimized, which is a more practical performance metric for the upper layer signaling process.
2) We formulate the DRL based solution using the policy gradient method. Taking the advantage of MDP, the decision is made according to the current state, not the process itself. Therefore, when the state spaces are the same, the policy trained in a specified scenario can be used to scenarios with different parameter settings. In the formulated MDP, the definition of states is given. The changes in some request parameters do not change the state space.

   Our results confirm that the well-trained policy network of PG-QR method performs well when the arriving rates vary in a certain range.

The whole paper is organized as follows. We form the system model in section II. The problem formulation is given in section III, and we further transform the optimization problem into an MDP process and give the definition of the states, actions and reward. In section IV, the solution based on DRL is proposed, and the simulation results of training and testing performance with the complexity analysis are given in section V. Finally, some conclusions are drawn in section VI.

II. SYSTEM MODEL

The system with one MEC server and several UEs is considered in this paper. The third-party service providers deploy the short video applications at the MEC server, where the original source files of copyrighted short videos are cached. Only the authentication procedure must be processed at the remote application server. When a UE requests a video, the authentication information is sent to the remote application server first. Then, the MEC server chooses a certain video quality level for this request and set up the RB accordingly. Since the source files are cached at the MEC, the transmitting latency and the traffic of the core network can be greatly reduced. The system architecture is given in Fig.1.

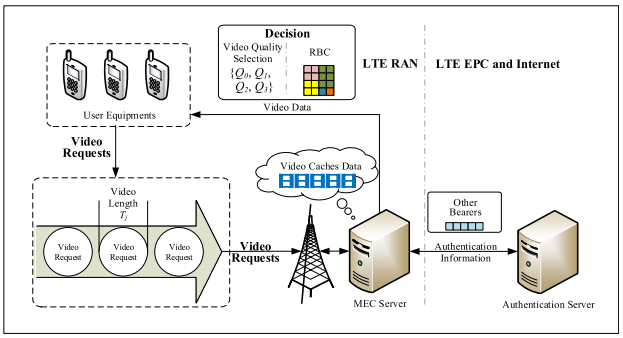


FIGURE 1. System architecture.

A. RESOURCE MODEL

The transmission resources considered in this paper include default RBs, dedicated RBs and other bearers consist of S1 bearers, S5/S8 bearers and external bearers. To facilitate the problem formulation, the bearers are quantified as units. Here we use N_R and N_E to denote the number of default RB units and other bearer units, respectively. To encourage reserving bearers for other applications, the common resource cost w_t must be calculated if the number of occupied bearer units exceeds the pre-defined threshold $N_{R,0}$ and $N_{E,0}$. Define

$$w_t = \max\left[0,\, \rho\left(N_{R,0} - n_{R,t}\right)\right] + \max\left[0,\, \rho\left(N_{E,0} - n_{E,t}\right)\right], \quad (1)$$

where, $n_{R,t}$ and $n_{E,t}$ are the number of occupied bearer units at time step t, and ρ is the cost ratio.

As the default RB only provide services with best efforts (Non-GBR), the transmission ability of each RB unit varies considering the randomness of wireless channel and the fluctuations of traffic from other applications. This varying ability is quantified as C_t. The number of dedicated RBs are not limited in our model. When C_t of the default RBs can not satisfy the transmission rate requirement of the current video, the dedicated RBs will be used and the extra resource cost h must be calculated.

B. SHORT VIDEO MODEL

Generally, the user experience highly depends on the quality of the video. In this paper, four typical quality levels of online videos are considered, i.e. 240P, 480P, 720P and 1080P. The corresponding quality indicators are denoted by $\{Q_0, Q_1, Q_2, Q_3\}$. Obviously, better quality of video needs more transmission resources. The number of RB units that the video required is defined as $\{B_0, B_1, B_2, B_3\}$ according to the different quality indicators. Similarly, the video quality profits UEs can obtain are defined as $\{G_0, G_1, G_2, G_3\}$. Meanwhile, each accepted request will occupy one other bear for the authentication information.

Assume the requests of videos from UEs arrive as a Poisson process with density λ. The length of the j-th requested video is denoted by T_j. If this request is accepted with quality $q_j = Q_{x_t}$, the consumed RB units and the video quality profits for the UE at time step t during the transmission are $b_{j,t} = B_{x_t}$ and $g_{j,t} = G_{x_t}$, respectively. The extra resource cost of the j-th request can be calculated as

$$h_{j,t} = \begin{cases} 0, & Q_{x_t} \le C_t \\ H\left(Q_{x_t}, C_t\right), & Q_{x_t} > C_t, \end{cases} \quad (2)$$

where $H\left(Q_{x_t}, C_t\right)$ is a discrete function denoting the extra resource cost caused by the video with quality level Q_{x_t} while the transmission ability of the default RB is C_t.

C. LATENCY CALCULATION MODEL

When the j-th request comes at time step t, it will join in the queue and wait for the setup of the bearer, and the latency of waiting is denoted by τ_j^{w}. Once it is scheduled and the corresponding bear is built, the authentication information will be sent to the remote application server and the MEC server will process the video source files according to the video quality indicator. The latency of this phase is roughly modeled as a constant denoted by τ^{a_0}. Therefore, the total latency can be calculated as

$$\tau_j = \tau_j^{\mathrm{w}} + \tau^{a_0}. \quad (3)$$

The procedure of RB reconfiguration is also considered in this paper. For those reconfigured requests, the change of video quality level will cause additional video processing. For each RB reconfiguration procedure, an additional latency τ^{r} is taken into account. If a request is reconfigured for N_{RF} times, the latency of video processing is

$$\tau^{\mathrm{a}} = \tau^{a_0} + N_{RF}\tau^{\mathrm{r}} \quad (4)$$

and

$$\tau_j = \tau_j^{\mathrm{w}} + \tau^{\mathrm{a}}. \quad (5)$$

III. PROBLEM FORMULATION

A. OPTIMIZATION PROBLEM

To provide better short video services to UEs with less resource cost in the proposed system, the long-term video quality profit, the penalty of latency and the cost of bearers are considered. The optimization problem can be expressed as

$$\begin{aligned}\max\ f(\boldsymbol{Q},\boldsymbol{Z}) = &\hat{g}\sum_{j=1}^{J_M}\sum_{t=t_j}^{t_j+T_j-1} g_{j,t}(\boldsymbol{Q},\boldsymbol{Z})\\ &-\hat{\tau}^{\mathrm{w}}\sum_{j=1}^{J_M}\tau_j^{\mathrm{w}}(\boldsymbol{Q},\boldsymbol{Z}) - \hat{\tau}^{\mathrm{a}}\sum_{j=1}^{J_M}\tau_j^{\mathrm{a}}(\boldsymbol{Q},\boldsymbol{Z})\\ &-\hat{h}\sum_{j=1}^{J_M}\sum_{t=t_j}^{t_j+T_j-1} h_{j,t}(\boldsymbol{Q},\boldsymbol{Z}) - \hat{w}\sum_{t=0}^{T_M} w_t(\boldsymbol{Q},\boldsymbol{Z})\\ \text{s.t. } & n_{R,t}(\boldsymbol{Q},\boldsymbol{Z}) \le N_R\\ & n_{E,t}(\boldsymbol{Q},\boldsymbol{Z}) \le N_E,\end{aligned} \tag{6}$$

where, T_M is the number of time steps of an observed system period, J_M is the number of requests in T_M, t_j is the start time of the j-th request, $\hat{g}$, $\hat{\tau}^{\mathrm{w}}$, $\hat{\tau}^{\mathrm{a}}$, $\hat{h}$ and $\hat{w}$ are the weight coefficients, $\boldsymbol{Z} = (z_{t,j})_{T_M\times J_M}$ denotes the maintaining of RBs. If there is a RB maintained for the j-th request at time step t, $z_{t,j} = 1$, otherwise, $z_{t,j} = 0$. $\boldsymbol{Q} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_j, \ldots, \boldsymbol{Q}_{J_M})^T$ denotes the matrix of video quality levels. As the RB reconfiguration procedure is considered, the video quality levels of a certain request might be different from time step to time step. Therefore, $\boldsymbol{Q}_j = (q_{j,t_j}, q_{j,t_j+1}, \ldots, q_{j,t_j+T_j-1})$. Since the video quality levels, the admission of requests and the occupation of bearers depend on the video quality selection and RB control, $g_{j,t}$, τ_j^{w}, τ_j^{a}, $h_{j,t}$, w_t, $n_{R,t}$ and $n_{E,t}$ are functions of $\boldsymbol{Q}$ and $\boldsymbol{Z}$.

B. MARKOV DECISION PROCESS MODEL

The optimization problem mentioned above is a long-term process optimization. To explore more efficient solutions, it can be transformed to a MDP, which is denoted by $(\mathcal{S}, \mathcal{A}, P^a_{ss'}, R^a_{ss'})$. $\mathcal{S}$ is the set of all possible states, $\boldsymbol{s}_t \in \mathcal{S}$, and $\mathcal{A}$ is the set of actions, which contains the video quality selection and RB control, $a_{t,j} = \{q_{j,t}, z_{t,j}\} \in \mathcal{A}$, ($j \in [1, J_M]$, $t \in [1, T_M]$). $P^a_{ss'} = P(\boldsymbol{s}'|\boldsymbol{s}, a)$ denotes the transition probability from state $\boldsymbol{s}' \in \mathcal{S}$ to $\boldsymbol{s}' \in \mathcal{S}$ by taking action $a \in \mathcal{A}$, and $R^a_{ss'}$ denotes the related reward. In the rest of this section, the definition of states, actions and reward are given.

1) STATES

The state of the MDP mainly consists of the states of bearers, requests and the transmission ability of each default RB unit at each time step. To limit the complexity, only the system state from time step t to $t + T - 1$ is observed, where t denotes the current time and $T > \max_j(T_j)$ denotes the maximum observed time length. The states of RBs and other bearers are denoted by matrix $\boldsymbol{s}_R = \left[s_{R,mn}\right]_{T\times\zeta(N_R)}$ and $\boldsymbol{s}_E = \left[s_{E,mn}\right]_{T\times\zeta(N_E)}$, respectively, where $\zeta(x)$ is the number of binary bits representing the decimal number x. Each row of the matrix represents the number of remaining available default RB units and other bearer units in binary to compress the size of the state. For the state of waiting requests, matrix $\boldsymbol{s}_w = \left[s_{w,mn}\right]_{T\times 2}$ is used, where $s_{w,m1}$ and $s_{w,m2}$ are the length and waiting time of the m-th requested video, respectively. For the processing requests, matrix $\boldsymbol{s}_p = \left[s_{p,mn}\right]_{T\times 2}$ denotes their states, where $s_{p,m1}$ and $s_{p,m2}$ denote the remaining time length and the video quality level at the current time step. To be aware of the transmission ability of default RBs in wireless environment, $\boldsymbol{s}_a = \left[s_{a,mn}\right]_{T\times 1}$ is used to denote the change of C_t over a period of time in the future. As C_t is a quantified transmission ability indicator, it is designed in accord with the transmission rates required by different video quality levels, and $s_{a,mn} \in \{Q_0, Q_1, Q_2, Q_3\}$. Therefore, the state of the MDP can be expressed as

$$\boldsymbol{s}_{T\times U} = [\boldsymbol{s}_R, \boldsymbol{s}_E, \boldsymbol{s}_w, \boldsymbol{s}_p, \boldsymbol{s}_a], \tag{7}$$

where $U = \zeta(N_R) + \zeta(N_E) + 5$.

According to the above definition, the number of states in set $\mathcal{S}$ is

$$N_S = |\mathcal{S}| = 2^{(\zeta(N_R)+\zeta(N_E))T}[\max_j(T_j)]^{2T} {T_M}^T 4^{2T}. \tag{8}$$

2) ACTIONS

At each time step, MEC server allocates the arrived requests and reallocates the processing requests simultaneously, specifically chooses the video quality level and configures the RBs. Originally, when the number of waiting requests is D_t^{w} at time step t and the number of processing requests is D_t^{p}, the number of actions in set $\mathcal{A}$ is $5^{D_t^{\mathrm{w}}+D_t^{\mathrm{p}}}$ since there are four video quality levels and an additional action denoting do nothing. To avoid the number of possible actions changing exponentially with time and simplify the MDP model, we observe the first D^{w} waiting requests and the first D^{p} processing requests at each time step. Refer to [27], we break down the joint action at each time step into a sequence of decisions, and each decision only chooses one request and its video quality level. To clearly describe the decision-making process, X_t is used to represent the number of separated decisions at time step t, and the action can be expressed as $\boldsymbol{a}_t = (a_{t,1}, a_{t,2}, \ldots, a_{t,x}, \ldots, a_{t,X_t})$, $a_{t,x} = a_{t,j} \in \mathcal{A}$. Then, the number of actions in set $\mathcal{A}$ is reduced to

$$N_A = |\mathcal{A}| = 4D^{\mathrm{w}} + 4D^{\mathrm{p}} + 1. \tag{9}$$

For the long-term cumulative value, serial decisions can achieve the same performance as the parallel decision when the quality selection and RBC are intelligent enough.

3) REWARD

In (6), the constraints of default bearer resources have been considered in the design of state. The objective function $f(\boldsymbol{Q},\boldsymbol{Z})$ will be decomposed into the reward r_t of each time step to formulate the MDP, and r_t can be expressed as

$$\begin{aligned}r_t(a_t) = &\hat{g}\sum_{j\in\mathcal{J}_1} g_{j,t}(a_t) - \hat{\tau}^{\mathrm{w}}\xi(\mathcal{J}_2) - \hat{\tau}^{\mathrm{a}}\xi(\mathcal{J}_3)\\ &-\hat{h}\sum_{j\in\mathcal{J}_1} h_{j,t}(a_t) - \hat{w}w_t(a_t),\end{aligned} \tag{10}$$

where $\mathcal{J}_1$ is the set of processing requests, $\mathcal{J}_2$ is the set of arrived and not allocated requests, $\mathcal{J}_3$ is the set of allocated requests at time step t, and $\xi(\cdot)$ is the counting function. Comparing with (6), the optimization function can be converted to the reward, r_t, which can be accumulated over time steps. Meanwhile, the value function can also be accumulated from initial state, which can be expressed as

$$v(s_0) = f(\boldsymbol{Q}, \boldsymbol{Z}) = \sum_{t=0}^{T_M} \gamma^t r_t(a_t) \tag{11}$$

where γ is the discount factor, and $\gamma = 1$ in our model.

IV. DEEP REINFORCEMENT LEARNING BASED SOLUTION

For MDP problem, classical dynamic programming is one of the traditional solution methods. To obtain the global optimal solution, the computational complexity of classical dynamic programming is $O(N_S^2)$ [28], where N_S is given in (8). Besides, classical dynamic programming is a model-based method, which means the transition probability $P^a_{ss'}$ must be completely known. However, due to the large value of N_S and the random changes of system state related to wireless characteristics, a perfect model of the proposed MDP is difficult to obtain. Therefore, classical dynamic programming is not appropriate for our problem.

In this section, we propose the DRL based method aiming at selecting video quality level and controlling RBs simultaneously. The policy gradient method is adopted, and an algorithm named as policy gradient based quality selection and RBC (PG-QR) is introduced to solve the problem. Since the decision making is extremely complex, a deep neural network $\pi_{\boldsymbol{\theta}}$ with parameter $\boldsymbol{\theta}$ is used to represent the policy function. According to the MDP we formulated, the state at each time step is the input of $\pi_{\boldsymbol{\theta}}$, and the output is the probability distribution of all the possible actions in $\mathcal{A}$, as shown in Fig.2.

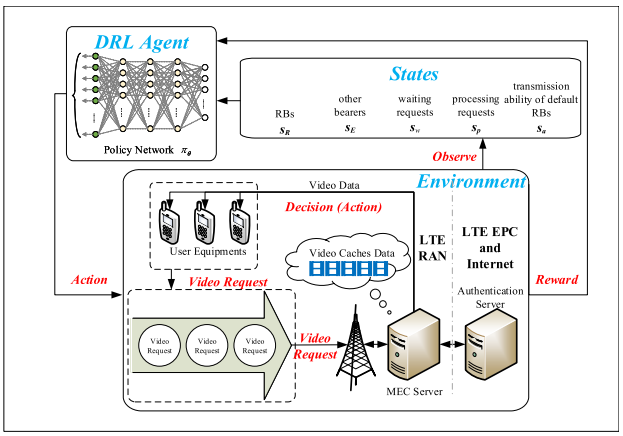


FIGURE 2. Deep Reinforcement Learning based video quality selection and radio bearer control process.

After taking the action a_t with the roulette wheel selection method, the environment will generate a series of reactions, which can be used to obtain the reward r_t. The policy network will be trained based on the trajectories which include the related states $\boldsymbol{s}$, actions a and rewards r, and the policy parameter $\boldsymbol{\theta}$ of the neural network will be updated accordingly.

According to the formulation of MDP, a decision consists of two parts, i.e. selecting a request waiting in queue or being processed and selecting the video quality level. To implement this design of action with one policy network, we tactfully code the actions as a matrix $\boldsymbol{M}$, which can be expressed as

$$\boldsymbol{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,4} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,4} \\ \vdots & \vdots & m_{i,j} & \vdots \\ m_{D^{\mathrm{W}}+D^{\mathrm{P}},1} & m_{D^{\mathrm{W}}+D^{\mathrm{P}},2} & \cdots & m_{D^{\mathrm{W}}+D^{\mathrm{P}},4} \end{bmatrix} \tag{12}$$

where $m_{i,j} = 4(i-1)+j$. The first D^{W} rows denote the waiting requests, and the following rows denote the requests being processed at the current time. Thus, $\mathcal{A} = \{m_{i,j}, m_0\}$, where $a^i_{t,x} = m_0$ means do nothing, and $a^i_{t,x} = m_{i,j}$ means the i-th request with j-th video quality level is selected and the corresponding RB is built or reconfigured.

The Monte Carlo REINFORCE algorithm with baseline in Section 13.4 in [22] is used to train the policy network $\pi_{\boldsymbol{\theta}}$ of PG-QR. We generate N_J short video request sequences for training. For each request sequence, I episodes are simulated in each training iteration to get I trajectories of $\{\boldsymbol{s}^i_{t,x}, a^i_{t,x}, r^i_{t,x}\}$, $i \in [1, I]$, $t \in [0, T_M]$ and $x \in [1, X_t]$. For $x = X_t$, $r^i_{t,x}$ is calculated using (10), otherwise $r^i_{t,x} = 0$. Each trajectory is used as a training sample in the training process. And the process of one episodic simulation is given in Algorithm 1.

Algorithm 1 One Episodic Simulation in PG-QR

```
run episode i = 1, ..., I:
while time step t < T_M do
    get one action a^i_{t,x} based on π_θ;
    if a^i_{t,x} is none then
        compute r^i_{t,x} for the current time based on (10);
        new requests arrive;
        time step t moves forward;
    else if a^i_{t,x} = m_{i,j} ∈ [1, 4D^W + 4] then
        α = ⌊a^i_{t,x}/4⌋;
        β = a^i_{t,x} − 4⌊a^i_{t,x}/4⌋;
        setup RB for the α-th waiting request with video
        quality level β;
    else if a^i_{t,x} = m_{i,j} ∈ [4D^W + 5, 4(D^W + D^P)] then
        α = D^W + ⌊a^i_{t,x}/4⌋;
        β = a^i_{t,x} − 4⌊a^i_{t,x}/4⌋;
        reconfigure RB for the α-th request being processed
        with video quality level β;
    end if
    update s_{T×U};
    record trajectories s^i_{t,x} = s_{T×U}, a^i_{t,x} and r^i_{t,x};
end while
```

As there are a series of decisions at each time step, a decision index l is introduced to facilitate the description of training. For the x-th decision at time step t, $l = x + \sum_{\tau=0}^{t-1} X_\tau$

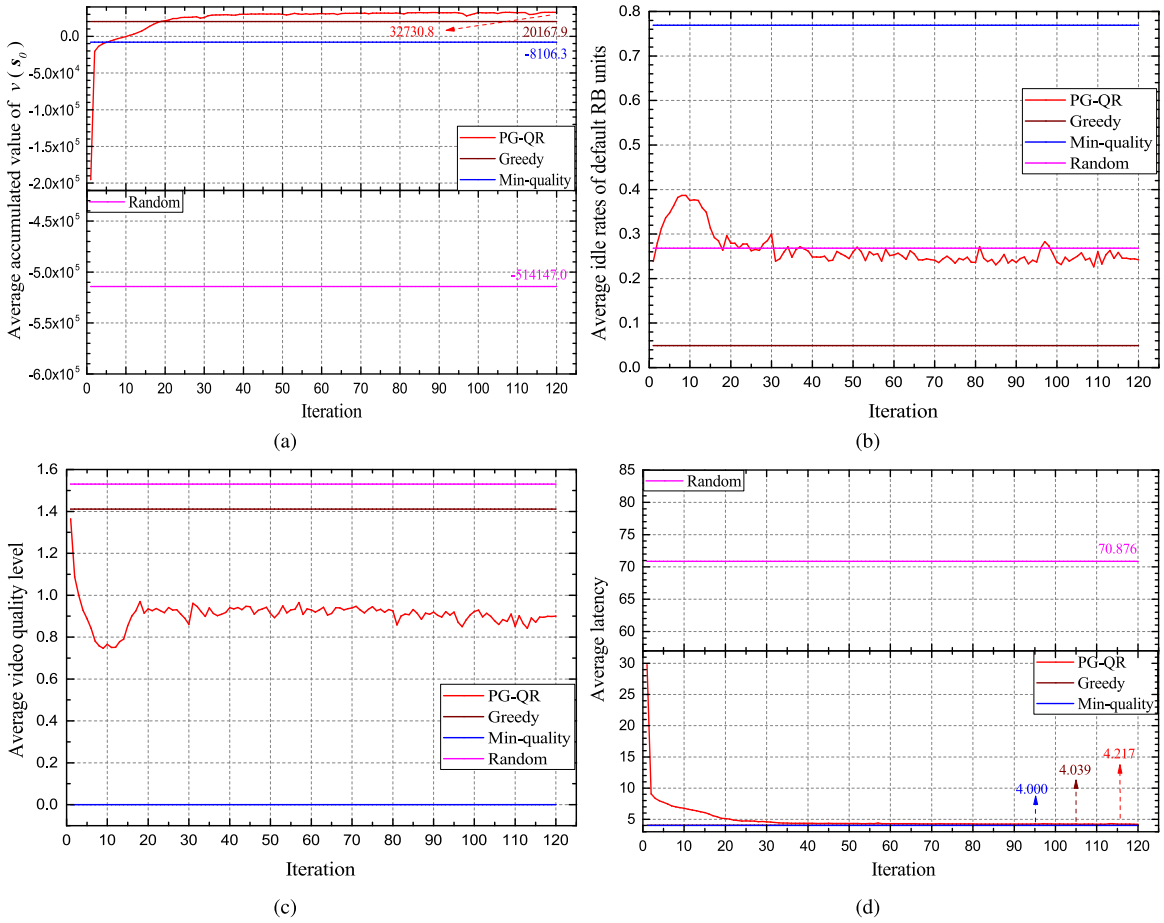


FIGURE 3. The training performance with $\hat{\tau}^{\mathbf{w}} = 8$ and $\hat{\tau}^{\mathbf{a}} = 8$.

and $L_M = \sum_{\tau=0}^{T_M} X_\tau$. Then, the value of the l-th decision of the i-th trajectory can be calculated as $v_l^i = \sum_{\tau=l}^{L_M} \gamma^{\tau-l} r_\tau^i$. The update of the parameters of the neural network, $\Delta\boldsymbol{\theta}$, can be calculated as [27]

$$\Delta\boldsymbol{\theta} = \sum_{l=0}^{L_M} \sum_{i=1}^{I} \nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(s_l^i, a_l^i)(v_l^i - b_l^i) \tag{13}$$

where $b_l^i = 1/I \sum_i v_l^i$ is a baseline.

V. EVALUATIONS AND ANALYSES

In this section, the training performance of the PG-QR method is given and analyzed. Then, the testing performances of the well-trained policy network under various short video request settings are given to evaluate the generalization. Greedy, min-quality and random are considered as the comparison methods. The greedy method firstly chooses the request with the longest waiting time, and then selects as high quality as possible for it. At each time step, the selection of request and video quality level are repeated until the bearer units are not enough or the queue is empty. The min-quality method always chooses the lowest video quality level for the requests waiting in queue to ensure that as many requests as possible are served without waiting.

A. TRAINING PERFORMANCE

The training parameters are shown in Table 1. The policy network in PG-QR method has 2 fully connected hidden layers of $N_H = 32$ neurons. The learning rate is empirically set to 0.001. We generate 50 request sequences with the ratio of small video requests $\eta = 80\%$. Each episodic simulation lasts for $T_M = 600$ time steps, and we run $I = 5$ episodes for each request sequence in each training iteration.

Fig.3 shows the training performance of PG-QR algorithm with the comparison of greedy, min-quality and random methods. The latency penalty coefficient is set to 8. In Fig.3(a), the average accumulated value $v(\boldsymbol{s_0})$ increases significantly over iterations, and achieves to about 32700, which is the best among all the methods and about 68% better than greedy. It can also be observed that $v(\boldsymbol{s_0})$ tends to converge after about 40 iterations, which is relatively quick.

In Fig.3(b) and In Fig.3(c), the average idle rates of default RB units and the average video quality level of all the requests are given. Obviously, the greedy method occupies more default RB units and gains better video quality than the min-quality method. According to the definition and parameter settings of reward in training, the more using of default RBs will cause the more cost. Meanwhile, the better video quality level will gain more profit. This leads the PG-QR

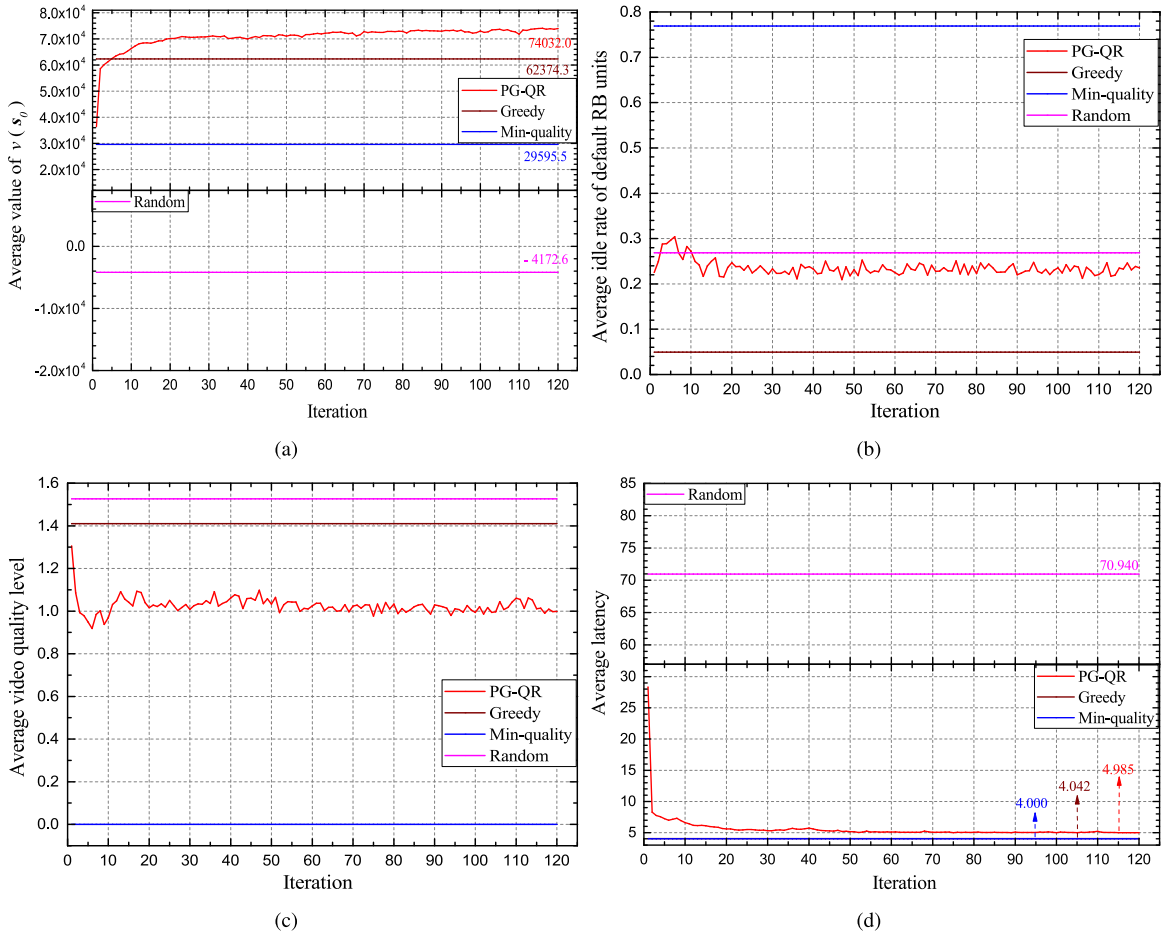


FIGURE 4. The training performance with $\hat{\tau}^{\mathrm{w}} = 1$ and $\hat{\tau}^{\mathrm{a}} = 1$.

TABLE 1. Training parameters.

Symbol	Description	Value
N_R	default RB units	256
N_E	other bearer units	256
$N_{R,0}$ $N_{E,0}$	threshold of occupied bearers	192
ρ	cost ratio of excess bearers	-1
C_t	varying transmission ability	$\{Q_0,Q_1,Q_2,Q_3\}$
$\{B_0,B_1,B_2,B_3\}$	number of required RB units	$\{1,3,5,8\}$
$\{G_0,G_1,G_2,G_3\}$	video quality profits	$\{1,2,4,8\}$
$H(Q_{x_t},C_t)$	extra resource cost	$\begin{pmatrix}0&0&0&0\\1&0&0&0\\2&1&0&0\\4&2&1&0\end{pmatrix}$
λ	arrive rate of requests	2
T_j	length of j-th requested video	$[10,30]$,$[60,90]$
τ^{a_0}	constant latency for the communication with the remote application server	3
τ^{r}	additional latency for the reconfiguration of RB	3
$\{\hat{g},\hat{h},\hat{w}\}$	weight coefficients	$\{1,1,1\}$
$\{\hat{\tau}^{\mathrm{w}},\hat{\tau}^{\mathrm{a}}\}$	weight coefficients	$\{8,8\}$,$\{1,1\}$

method to learn the economical decisions which occupy moderate number of default RB units and obtain moderate video quality profit. The economical decisions finally achieve a better balance between different reward factors and gain the best-accumulated value. As shown in Fig.3(d), the latency performance of the proposed PG-QR method converges to those of greedy and min-quality methods, which is extremely low.

To explore the intelligence of PG-QR method, we arrange another setting of reward parameters, which reduce the penalty of latency from 8 to 1. Comparing the results in Fig.3 and Fig.4, we can observe the similar trends. But the latency is increased with the reducing of latency-penalty and the average video quality level is increased slightly. This comparison reflects that the training processes can lead the policy networks to obtain proper trade-off between different objective factors when the weight coefficients are changed. Thus, the proposed PG-QR become an economical minded method with the guidance of the reward function.

B. TESTING PERFORMANCE

PG-QR algorithm is tested with the arriving rates of short video requests from 1.4 to 2.4 and $\hat{\tau}^{\mathrm{w}} = \hat{\tau}^{\mathrm{a}} = 8$. For each test setting, 300 request sequences are generated, and for each sequence $T_M = 600$ time steps are simulated. Since the results of random method are far below other methods in the training results, we only give the results of PG-QR, greedy and min-quality methods in Fig.5.

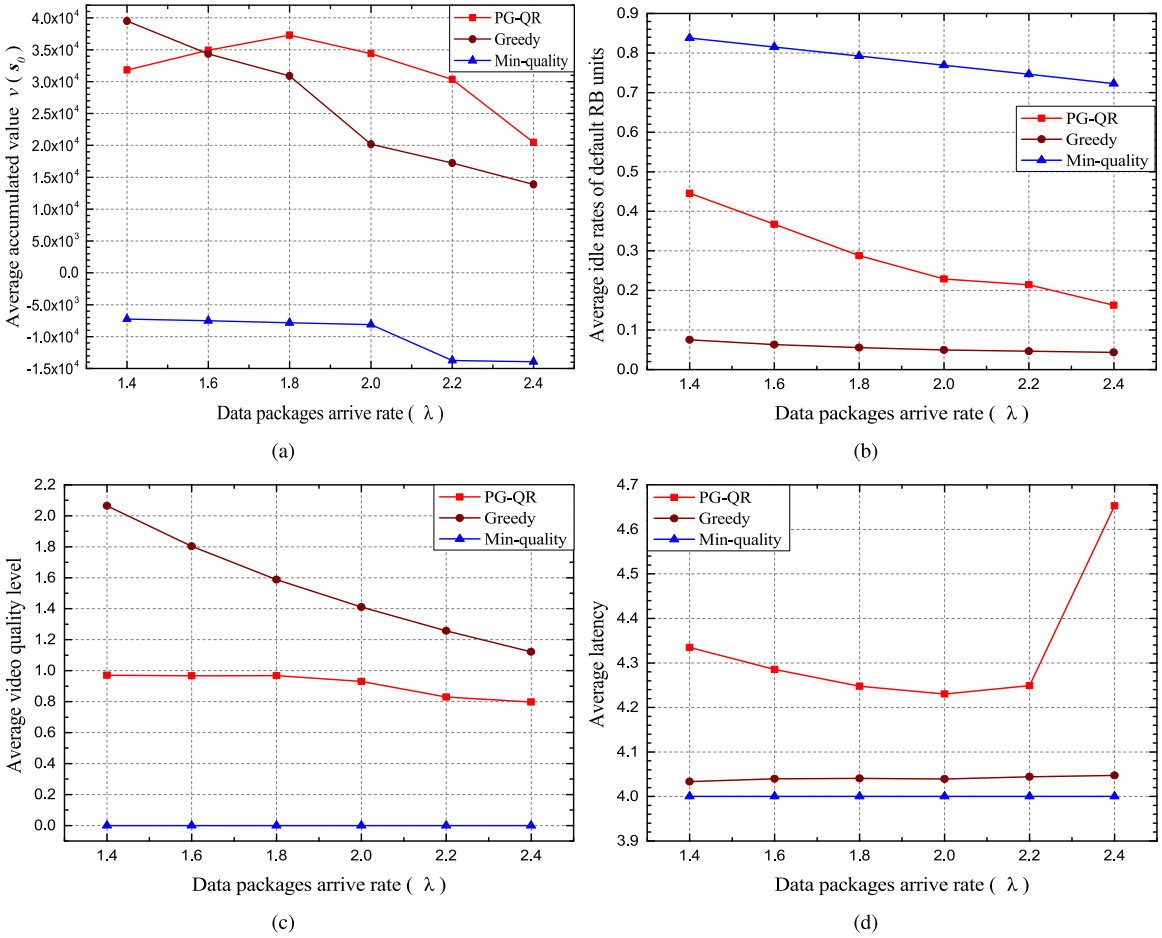


FIGURE 5. Performance comparison with various requests arrive rates.

It is worth noting that although the policy network is trained by 50 request sequences with a specific environment setting ($\lambda = 2.0$, $\eta = 80\%$, $\hat{\tau}^{\mathrm{w}} = \hat{\tau}^{\mathrm{a}} = 8$), it performs the best when λ various from 1.8 to 2.4. When $\lambda = 1.8$, the average value of $v(\boldsymbol{s}_0)$ obtained by PG-QR is the best as shown in Fig.5(a). Therefore, the results denote that the proposed PG-QR method has a good generalization feature in a certain range.

The performance of average idle rate of the default RB units is shown in Fig.5(b). The greedy method uses the RB units as many as possible and keeps the average idle rate below 0.1. By contrast, the min-quality method occupies the least RB units, and the value of average idle rate is large. It decreases linearly with the linear increase of arrival rate. The average idle rate performance of PG-QR method is between that of greedy and min-quality and achieves a better balance between the cost and the profit.

Fig.5(c) gives the results of average video quality level. The min-quality method always keeps the video quality at the lowest level. For the greedy method, the video quality decreases almost linearly with the linear increase of arrival rate due to the limited default RB units. At the same time, the proposed PG-QR method keeps the average video quality level between 0.8 to 1.0 with various arriving rates.

For the latency performance given in Fig.5(d), there is not much difference between different methods. This is because the latency penalty coefficients ($\hat{\tau}^{\mathrm{w}} = \hat{\tau}^{\mathrm{a}} = 8$) is large and the traffic load has not exceeded the system capacity.

C. COMPLEXITY ANALYSIS

According to the training and testing results, greedy is the second-best method. In the greedy method evaluated here, the reconfiguration procedure of the requests being processed is omitted, and it needs to observe the states of bearers, the requirements of waiting requests and the real-time transmission ability of each default RB unit. Thus, the computational complexity for each decision using greedy algorithm is $O(N_G)$, where $N_G = (\zeta(N_R) + \zeta(N_E) + 3)T$.

$$O(N_G) = O[(\zeta(N_R) + \zeta(N_E) + 3)T] \tag{14}$$

For PG-QR method, the training stage is time-consuming. However, using the well-trained policy network in practical is efficient. Since a deep neural network with 2 fully connected hidden layers is used as the policy network, the computational complexity is related to the number of elements in the input state, the number of neurons in the hidden layer and the number of elements in the output action probability vector. It can be calculated as

$$O[N_I N_H + N_H^2 + N_H N_A] \tag{15}$$

where $N_I = (\zeta(N_R) + \zeta(N_E) + 5)T$, N_A is given in (9). According to the setting of parameters in this paper, $N_A = 33$ and $N_H = 32$. Thus, for each decision using the policy network, the computational complexity is comparable to that of greedy method. Therefore, the proposed PG-QR method is an effective intelligent joint video quality selection and RBC method with low implementation complexity.

VI. CONCLUSION

In this paper, we present a scenario of MEC supported short video applications. For the short video requests arriving in a certain time, the video quality selection and the RBC procedures of RB setup, RB reconfiguration and RB release are jointly optimized. The problem is formulated as an MDP with the consideration of three factors, i.e., the long-term video quality profit, the cost of bearers and the penalty of latency. A DRL based solution method PG-QR is proposed. Simulation results show that PG-QR presents a remarkable training performance that converges quickly. In comparing with the greedy method, when the weight coefficients of reward are changed, the training gain of PG-QR is about 12000. This means the reinforce signal provided by the reward can lead the policy networks to obtain proper trade-off between different objective factors. The policy network trained by a specific setting of the scenario parameters is tested with the arriving rates of short video requests from 1.4 to 2.4. Results show that it performs the best when the arriving rates are between 1.8 and 2.4. But when arriving rate is less than 1.6, the performance is worse than that of greedy. Although the generalization ability of the PG-QR method only appears in a limited range, it is still a meaningful exploration of the generalization ability of the policy gradient method in such a complicated communication scenario. How to improve the training process to extend the generalization ability to more scenarios needs further study.

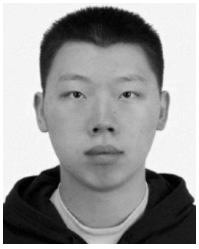
REFERENCES

[1] R. Trestian, I.-S. Comsa, and M. F. Tuysuz, "Seamless multimedia delivery within a heterogeneous wireless networks environment: Are we there yet?" *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 945–977, 2nd Quart., 2018.
[2] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.
[3] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.
[4] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 7, pp. 1594–1607, Jul. 2019.
[5] C. Li, L. Toni, J. Zou, H. Xiong, and P. Frossard, "QoE-driven mobile edge caching placement for adaptive video streaming," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 965–984, Apr. 2018.
[6] G. Ma, Z. Wang, M. Zhang, J. Ye, M. Chen, and W. Zhu, "Understanding performance of edge content caching for mobile video streaming," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1076–1089, May 2017.
[7] Z. Li, J. Chen, and Z. Zhang, "Socially aware caching in D2D enabled fog radio access networks," *IEEE Access*, vol. 7, pp. 84293–84303, 2019.
[8] J. Martín-Pérez, L. Cominardi, C. J. Bernardos, A. de la Oliva, and A. Azcorra, "Modeling mobile edge computing deployments for low latency multimedia services," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 464–474, Jun. 2019.
[9] Z. Su, Q. Xu, F. Hou, Q. Yang, and Q. Qi, "Edge caching for layered video contents in mobile social networks," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2210–2221, Oct. 2017.
[10] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. S. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.
[11] L. Pu, L. Jiao, X. Chen, L. Wang, Q. Xie, and J. Xu, "Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1751–1767, Aug. 2018.
[12] D. Wu, J. Yan, H. Wang, D. Wu, and R. Wang, "Social attribute aware incentive mechanism for device-to-device video distribution," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1908–1920, Aug. 2017.
[13] D. Wu, Q. Liu, H. Wang, D. Wu, and R. Wang, "Socially aware energy-efficient mobile edge collaboration for video distribution," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2197–2209, Oct. 2017.
[14] Z. Zhang and L. Wang, "Social tie-driven content priority scheme for D2D communications," *Inf. Sci.*, vol. 480, pp. 160–173, Apr. 2019.
[15] Z. Li, Q. Wang, and H. Zou, "QoE-aware video multicast mechanism in fiber-wireless access networks," *IEEE Access*, vol. 7, pp. 123098–123106, 2019.
[16] L. Sun, H. Pang, and L. Gao, "Joint sponsor scheduling in cellular and edge caching networks for mobile video delivery," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3414–3427, Dec. 2018.
[17] Z. Li, Y. Jiang, Y. Gao, D. Yang, and L. Sang, "On buffer-constrained throughput of a wireless-powered communication system," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 283–297, Feb. 2019.
[18] D. Wu, Q. Liu, H. Wang, Q. Yang, and R. Wang, "Cache less for more: Exploiting cooperative video caching and delivery in D2D communications," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1788–1798, Jul. 2019.
[19] B. Dai, Y.-F. Liu, and W. Yu, "Optimized base-station cache allocation for cloud radio access network with multicast backhaul," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1737–1750, Aug. 2018.
[20] A. Campbell, G. Coulson, and D. Hutchison, "A quality of service architecture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 24, no. 2, pp. 6–27, 1994.
[21] N. D. Nguyen, T. Nguyen, and S. Nahavandi, "System design perspective for human-level agents using deep reinforcement learning: A survey," *IEEE Access*, vol. 5, pp. 27091–27102, 2017.
[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[23] J. Zhu, Y. Song, D. Jiang, and H. Song, "A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2375–2385, Aug. 2018.
[24] P. Zhang, X. Kang, X. Li, Y. Liu, D. Wu, and R. Wang, "Overlapping community deep exploring-based relay selection method toward multi-hop D2D communication," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1357–1360, Oct. 2019.
[25] H. Zhang, W. Wu, C. Wang, M. Li, and R. Yang, "Deep reinforcement learning-based offloading decision optimization in mobile edge computing," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–7.
[26] R. Li, Z. Zhao, Q. Sun, C. Yang, C.-L. I, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for resource management in network slicing," *IEEE Access*, vol. 6, pp. 74429–74441, 2018.
[27] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. ACM Workshop Hot Topics Netw.*, 2016, pp. 50–56.
[28] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

WENJUN WU received the B.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, China, in 2007 and 2012, respectively. From 2012 to 2015, she was a Postdoctoral Researcher with Beihang University, China. She is currently an Associate Professor with the Beijing University of Technology, China. Her research interests include mobile edge computing, blockchain, and deep reinforcement learning.

YANG GAO received the B.S. degree from the Faculty of Information Technology, Beijing University of Technology, China, in 2018, where she is currently pursuing the M.S. degree. Her current research interests include resource allocation, mobile edge computing, deep reinforcement learning, and blockchain.

TIANQI ZHOU is currently pursuing the M.S. degree with the Faculty of Information Technology, Beijing University of Technology, China. Her current research interests include resource allocation, mobile edge computing, and deep reinforcement learning.



YINHUA JIA is currently pursuing the B.S. degree with the Faculty of Information Technology, Beijing University of Technology, China. His current research interests include edge computing, deep learning, and orthogonal time frequency space modulation.

HAO ZHANG received the B.S. degree from the Information Department, Beijing University of Technology, in 2019. He is currently pursuing the M.S. degree with the Faculty of Engineering, University of Ottawa. His current research interests include edge computing, crossed layer design, smart automobile, and the IoT technology.

TINGTING WEI is currently pursuing the M.S. degree with the Faculty of Information Technology, Beijing University of Technology, China. Her current research interests include computation offloading, fog radio access networks, mobile edge computing, reinforcement learning, and resource allocation.

YANG SUN received the Ph.D. degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China, in 2018. She is currently a Lecturer with the Beijing University of Technology. Her current research interests include ultra-dense heterogeneous networks, interference management, massive MIMO, and green telecommunications.