Received June 30, 2019, accepted August 2, 2019, date of publication August 15, 2019, date of current version August 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2935545

Resource Allocation in Information-Centric Wireless Networking With D2D-Enabled MEC: A Deep Reinforcement Learning Approach

DAN WANG[1], HAO QIN[1], BIN SONG[1], (Senior Member, IEEE), XIAOJIANG DU[2], (Senior Member, IEEE), AND MOHSEN GUIZANI[3], (Fellow, IEEE)

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2]Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
[3]Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Corresponding author: Bin Song (bsong@mail.xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772387, in part by the Fundamental Research Funds of Ministry of Education and China Mobile under Grant MCM20170202, in part by the National Natural Science Foundation of Shaanxi Province under Grant 2019ZDLGY03-03, in part by the Graduate Innovation Fund of Xidian University under Grant 5001-20109195456, and in part by the ISN State Key Laboratory.

ABSTRACT Recently, information-centric wireless networks (ICWNs) have become a promising Internet architecture of the next generation, which allows network nodes to have computing and caching capabilities and adapt to the growing mobile data traffic in 5G high-speed communication networks. However, the design of ICWN is still faced with various challenges with respect to capacity and traffic. Therefore, mobile edge computing (MEC) and device-to-device (D2D) communications can be employed to aid offloading the core networks. This paper investigates the optimal policy for resource allocation in ICWNs by maximizing the spectrum efficiency and system capacity of the overall network. Due to unknown and stochastic properties of the wireless channel environment, this problem was modeled as a Markov decision process. In continuous-valued state and action variables, the policy gradient approach was employed to learn the optimal policy through interactions with the environment. We first recognized the communication mode according to the location of the cached content, considering whether it is D2D mode or cellular mode. Then, we adopt the Gaussian distribution as the parameterization strategy to generate continuous stochastic actions to select power. In addition, we use softmax to output channel selection to maximize system capacity and spectrum efficiency while avoiding interference to cellular users. The numerical experiments show that our learning method performs well in a D2D-enabled MEC system.

INDEX TERMS ICWN, MEC, D2D, resource allocation.

I. INTRODUCTION

In addition to advances in information and communications technology, the proliferation of smart mobile devices is undergoing unprecedented growth [1]. Mobile applications in devices such as face recognition, natural language processing, and augmented reality are emerging constantly, resulting in ever-increasing data traffic [2]. Therefore, data services are expected to become information-centric communications to meet multimedia file sharing and video transmission [3] in future fifth-generation (5G) networks. However, traditional wireless cellular networks have gradually become incapable of meeting the strong demands not only in high network capacity but also in high computational capabilities [4]. Consequently, a network with a flexible structure is desirable.

Information-centric wireless networking (ICWN) is a promising next-Internet architecture that has better scalability and robustness. The goal is to evolve the Internet infrastructure to directly support information distribution by introducing uniquely named data as a core Internet principle [5]. ICWN enables network nodes to have computation and caching capabilities to accommodate the increasingly growing traffic of mobile data in the 5G high-speed communication networks [6]. Recently, the ICWN approach has

The associate editor coordinating the review of this article and approving it for publication was Balázs Sonkoly.

been explored by a number of researchers. Compared with traditional networks, ICWN provides network node caching capabilities in many implementations to further improve the network performance. However, the technical issues and challenges created by the ICWN network require in-depth research and thinking, such as the high and variable latency of transmitted high-volume quantities of data to the cloud for data processing. Thus, this approach causes a heavy burden on the network, while network congestion and high network demands need to be considered, such as computing, caching and communicating (3C).

Designing ICWNs face various challenges related to the capacity and traffic. To address the above issues, one prevalent method is to employ mobile edge computing (MEC) and device-to-device (D2D) communications, which can offload the core network and increase the capacity of the network [7]. In the recent ICWN paradigm, the D2D-enabled MEC can collaborate with cached popular contents on various nearby devices, helping to improve spectrum efficiency and decrease traffic congestion [8].

The emerging MEC is a promising approach for moving a portion of the data/computation to the edge of the network instead of sending it to the cloud datacenters [9]. MEC provides mobile users (MUE) with highly reliable, low-latency computing and communication services. In addition, D2D communications have been applied in MEC systems. D2D communications can be beneficial to MEC in two aspects: using the terminal device for content caching, and using the D2D link to aid the MEC node in performing service data transmission [10], which can efficiently reduce the high cost of base station (BS) transmission, reduce users' download time and improve users' QoE. Hence, efficiently allocating limited communication resources and optimizing the policy of power control and resource allocation in communication is still an urgent issue in integrating these two techniques in ICWN. In the D2D-enabled MEC system, our motivation to study efficient resource allocation and power control algorithms is two-fold. First, because of the continuous establishment of D2D links, MEC and D2D user collaborative content caching will improve cache efficiency in ICWNs, but spectrum reuse may cause serious inter-user interference [11]. Second, communication resource allocation directly affects the quality of communication links. Therefore, the problem of reasonably establishing links and allocating communication resources cannot be ignored.

In this paper, we consider a multiuser D2D-enabled MEC system in ICWN, as shown in Fig. 1. There are numerous small cells. The MEC servers and MUE are deployed in cells. When considering the communication resource allocation of D2D-enabled MEC in an ICWN, we employ a novel deep reinforcement learning (DRL) approach to automatically optimize resource allocation and power control decisions. The contributions of this paper are as follows:

(1) We first introduce the system model and optimization goal. We determine the communication mode based on the location of the cached content, whether it is the D2D communication mode or cellular communication mode.
(2) Then, a resource allocation with a policy gradient method is proposed, which is a joint resource allocation and power control algorithm for a D2D-enabled MEC.
(3) Optimization is a two-objective problem. We use the Gaussian distribution as a parameterization strategy to generate continuous stochastic actions to select power. Moreover, we use a softmax output channel selection to maximize system capacity and spectrum efficiency while minimizing interference.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work of resource allocation and power control for D2D-enabled MEC in ICWNs. Next, we describe the system model and optimization goal in section 3. Then, we propose the method of resource allocation and power control with DRL in section 4. In section 5, the performance of the proposed algorithm is verified by experiments. Finally, section 6 concludes the paper.

II. RELATED WORK

In this section, we investigate the recent work in ICWNs. To address the challenges caused by combining resource allocation and power adaptation in ICWN, a number of novel research technologies have been proposed in ICWN. Most recently, several approaches based on edge service frameworks have been the popular research topic in ICWNs. Considerable work has been performed on integrating wireless networks and information-centric networking. For instance, Liang *et al.* in [6] proposed an ICWN virtualization architecture for integrating wireless network virtualization with information-centric networking (ICN) and developed the key components of this architecture. TalebiFard *et al.* in [12] provided a framework for supporting service-centric networks, while they considered that the interaction service latency, customization, and contextualization will be at the network edge.

Moreover, to fully develop the potential of ICWNs, exploiting MEC, in-network caching, and D2D communication has become a popular research area. In-network caching is one of the key features of ICWN. He *et al.* in [13] considered the allocation of resources in trust-based MSNs with MEC, caching and D2D when the conditions of the network resources vary with time. In addition, a paper studied a novel device-to-device (D2D)-enabled multihelper MEC system in which a local user solicits its nearby WDs, serving as helpers for cooperative computation [14]. They primarily provided a joint task assignment and resource allocation for D2D-enabled mobile edge computing.

In the ICWN, since the wireless spectrum is still a bottleneck resource, the research on D2D-enabled MEC is significant for efficient wireless access. Specifically, when the communication resource allocation is resolved in the D2D-enabled MEC, it is necessary to consider the interference problem existing in D2D communication. Thus far, D2D interference management in traditional D2D communications

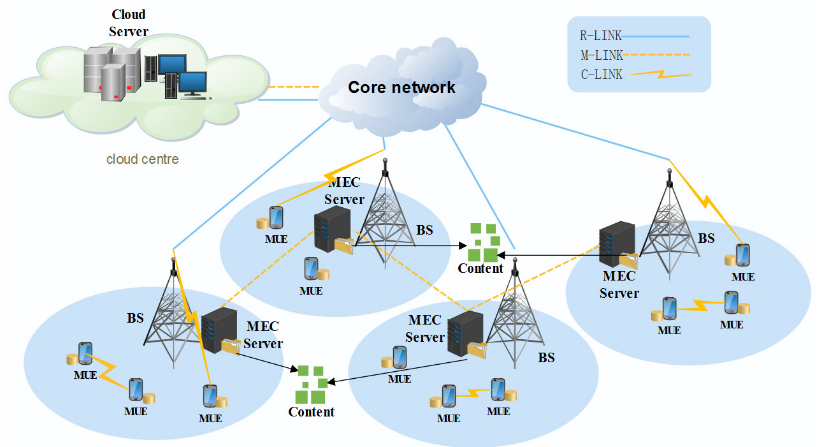

FIGURE 1. A multiuser D2D-enabled MEC system in ICWNs. (The R-LINK is the radio link, the M-LINK is the connection link of edge server and the C-LINK is the caching link.)

has received much attention. There are three main aspects: mode selection, resource allocation, and power selection.

To address the issues above, increasingly new methods have been proposed to reduce communication interference in D2D communication. In addition to traditional optimization methods, game theory and RL methods have been become prevalent methods to address interference management problem in wireless communication, especially the distributed decision-making problem and networking management [15], [16]. Zhang *et al.* [17] developed a coalitional game with transferable utility in which each user had the incentive to cooperate with other users to form a strengthened user group to increase the opportunity to win their preferred spectrum resources. Furthermore, the RL method has been used to achieve resource allocation, mode selection and power control by modeling these problems as Markov decision processes (MDPs). Qiu *et al.* [18] developed a joint mode selection and power adaptation approach using a multiagent Q-learning algorithm based on conjecture. Zhao *et al.* [19] proposed power control for D2D communication, which uses multiagent reinforcement learning (MARL) to maximize system throughput by adjusting the transmit power of each D2D user.

To summarize, there is still a demand to explore and investigate the proposed communication resource allocation algorithms for D2D-enabled MEC systems in ICWNs. In contrast to all existing works, in this paper, we focus on communication resource allocation with deep reinforcement learning (DRL) in D2D-enabled MEC, enabling mobile users to automatically learn allocation policies based only on their cached content and channel information.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, the system model used in this paper is described. We first illustrate the network model description. Then, we briefly introduce the MEC, in-network caching, and D2D communications model in ICWN. Finally, we formulate the optimization problem in detail.

A. NETWORK MODEL

As shown in Fig. 1, the network model consists of N small cells. The cells are connected to the Internet through the core network of the cellular communication system [2]. MEC servers are placed in the BS to provide data services to the MUE. The set of small cells is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$, and we set $\mathcal{M}_n = \{1, 2, \ldots, M_n\}$ to represent the number of BS and MEC servers. We assume that a BS is associated with the K_n MUE and an MEC server. The $\mathcal{K}_n$ is defined as $\mathcal{K}_n = \{1, 2, \ldots, K_n\}$, and K_n refers to the kth MUE in nth cell.

The D2D-enabled MEC system provides an offloading method for the core of the cellular network. It can handle tasks as far away as possible from the core network. We assume that distributed MEC nodes can cooperate with content caching in ICWNs and allow D2D communication. In this scenario, both the MEC server and the MUE deployed on the BS have a content caching capability. Each MUE can offload the cached content by selecting a communication mode, including cellular mode and D2D mode. The D2D mode can be implemented by D2D communication, and it can perform tasks without involving a cellular network [7].

In the communication network architecture, when an MUE requests data content, it can usually be implemented in two communication modes. 1) D2D mode: the D2D user in the communication range has buffered the requested content and then directly transmits it to the requesting user through the D2D link. 2) Cellular mode: the local cell's MEC server buffer has the requested content and can then be sent to the user. Generally, the MUE sends a request by broadcast to determine whether the local MUE has cached the content. D2D communication mode achieves low latency, reduces the traffic load through the network, and improves the cooperation of MUE at the edge of the network.

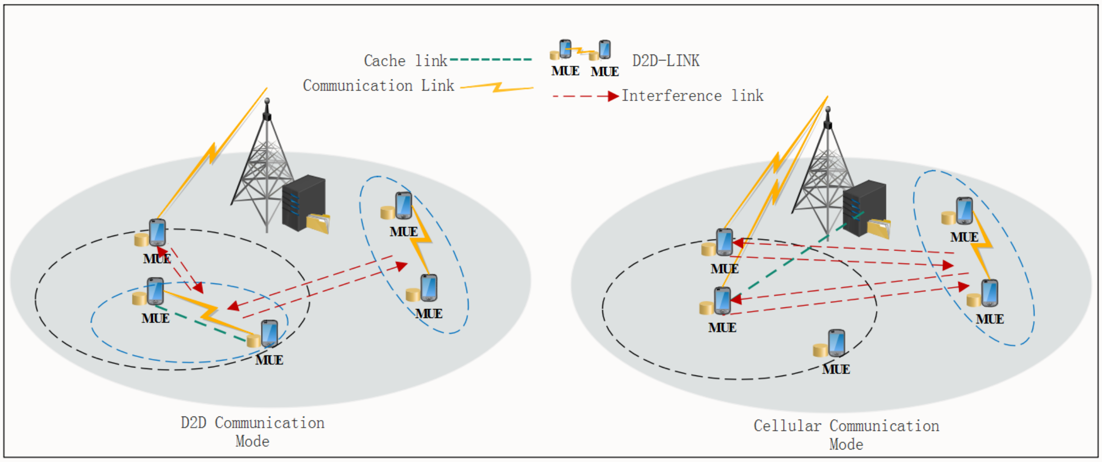

FIGURE 2. The communication mode of the D2D-enabled MEC system. ((a) D2D communication mode (b) cellular communication mode.)

B. SYSTEM MODEL

In this work, we focus on the communication model, considering a joint channel and power allocation algorithm with DRL, which can be used to solve the resource allocation problem in the D2D-enabled MEC system. Specifically, we consider resource allocation in the D2D communication mode and the cellular communication mode. As illustrated in Fig. 2, we describe the scenario of two communication modes. Each mode includes a D2D link, a cache link, a communication link and an interference link. In each cell, we assume that there are $\mathcal{K}_n$ MUE, denoted as $\mathcal{K}_n = \{1, 2, \ldots, K_n\}$. These MUEs can choose whether to become a D2D transmission user depending on the content cache. Fig. 2(a) is the D2D communication mode, and Fig. 2(b) is a cellular communication mode. When an MUE requests data content, it can usually be implemented in two communication modes. We describe these modes as follows:

1) CELLULAR MODE

Mobile user equipment communicates with another MUE through the BS. In this mode, the local cell's MEC server buffer has the requested content and can then be sent to the user.

2) D2D MODE

Mobile user equipment communicates directly with another MUE through direct traffic [20]. In this mode, the D2D user in the communication range has buffered the requested content and then directly transmits it to the requesting user through the D2D link.

Moreover, we assume that 1) in D2D communication, cellular users utilize the downlink (DL) resources of the cell, while D2D pairs reuse the downlink resources non-orthogonally; 2) a cellular user and D2D pairs share the same resource block and each resource block is allocated to one cellular user and shared with multiple D2D pairs. Here, we assume that D2D pairs reuse the downlink resource in the central cell. Therefore, there are three types of interference: D2D-to-cellular interference, cellular-to-D2D interference and D2D-to-D2D interference [21].

C. PROBLEM FORMULATION

We assume that the BS can use the resource scheduler to allocate D2D users to different channels and that the user can select different powers to avoid interference. Let $\mathcal{B}$ denote the channel bandwidth of the D2D-enabled MEC system, which can be divided into $\mathcal{H}$ PRBs. Each PRB is expressed as $\mathcal{b}_i = \frac{\mathcal{B}}{\mathcal{H}}, \mathcal{i} \in \{1, 2, \ldots, H\}$. In this scenario, we consider the problem of choosing a mode. Let $\mathcal{V} = \{\mathcal{v}_c, \mathcal{v}_d\}$ denote the communication mode of MUE. Next, we refer to one assumption for mode selection [22]. The data center provides $\mathcal{H}$ different contents, denoted as $\mathcal{C} = \{c_1, c_2, \ldots, c_{\mathcal{L}}\}$. The content caching matrix is defined as $\mathcal{X} = \{x_{k,f} \in (0, 1) \,|u_k \in \mathcal{U}, c_f \in \mathcal{C}\}$. $x_{k,f} = 1$ indicates that the content c_f is cached in the MUE and $x_{k,f} = 0$ indicates that the content c_f is cached in the local MEC server. When a mobile user sends a content request, we first need to determine if it is satisfied by the D2D user or the MEC service.

In addition, in D2D communication mode, we assume that a D2D pair can reuse multiple channels to ensure successful transmission of packets while meeting the QoS requirements of the entire communication system with minimum power consumption. We assume that there are $\mathcal{m}$ mobile users choosing to become D2D users and $\mathcal{n}$ mobile users choosing to became cellular users. Let $\gamma_{\mathcal{i}}$ denote the interference plus noise ratio (SINR) of the cellular user. For successful transmission, the SINR is higher than γ^*:

$$\gamma_{\mathcal{i}} > \gamma^*, \quad \forall \mathcal{i} \in N \tag{1}$$

where γ^* is a threshold of SINR to maintain communication. Generally, the SINR of the ℓth cellular user on the $\mathcal{i}$th channel is denoted as:

$$\gamma_\ell^i = \frac{P_l^i \cdot h_\ell^i}{I + \sum_{j=1}^{m} P_{\ell,j}^i \cdot h_{\ell,j}^i} \tag{2}$$

where P_i^ℓ is the transmission power of the ℓth cellular user of the ith channel, and h_i^ℓ is the link gain of the ℓth cellular user. Here, we denote $h_i^\ell = \mathscr{P} \cdot d^{-r}$, $\mathscr{P}$ is the pathloss, and r is the constant. $P_{i,j}^\ell$ denotes the transmission power of the jth D2D user that reuses the ith channel. The link gain of the D2D user is denoted as $h_{i,j}^\ell = \mathscr{Q} \cdot d^{-s}$, and $\mathscr{Q}$ is the pathloss, s is the constant. Here, I represents the power of the additive white Gaussian noise (AWGN). We assume that there is no interference from neighboring cells because we assume that the neighboring cells use channel resources of different bandwidths. Let the SINR of the jth D2D links on the ith channel be:

$$\gamma_j^i = \frac{P_{\ell,j}^i h_{\ell,j}^i}{I + (P_\ell^i \cdot h_\ell^i + \sum_{j' \in m}^{j' \neq j} P_{j',j}^i \cdot h_{j',j}^i)} \tag{3}$$

where $h_{\ell,j}^i$ is the link gain of the jth D2D user reusing the ith channel, h_ℓ^i is the link gain of the ℓth cellular user in ith channel, and $P_{\ell,j}^i$ is the transmission power of the jth D2D user. $P_{j',j}^i$ is the transmission power of the j'th D2D user of the ith channel, and P_ℓ^i is the transmission power of the ℓth cellular user of the ith channel. Similarly, I denotes the power of the AWGN. In a communication system, we define the capacity of a cellular user in the D2D-enabled MEC system as follows:

$$\mathfrak{C}_c = b_i log_2 \left(1 + \gamma_\ell^i\right) \tag{4}$$

In addition, the capacity of D2D users is given by

$$\mathfrak{C}_d = b_i log_2 \left(1 + \gamma_j^i\right) \tag{5}$$

The total system capacity of MUE is defined as:

$$\mathfrak{C} = \mathfrak{C}_c + \mathfrak{C}_d \tag{6}$$

Therefore, in both modes of communication, our optimization goal is to make allocation decisions based on channel quality between mobile users and BSs and interference between D2D users while maximizing total system capacity.

IV. RESOURCE ALLOCATION ALGORITHM

In the previous section, we formulated the optimization problem in the communication mode of the D2D-enabled MEC system. Here, we devise a resource allocation method based on a policy gradient algorithm to address the proposed problem. We divided the method into two subtasks as follows.

1) In the first subtask, we design the selection mechanism of the communication mode according to the cache matrix. When there is cached content in the MUE, the MUE selects the D2D communication mode. Otherwise, the cellular communication mode is selected.
2) In the second subtask, when the mobile users select the D2D communication mode, the D2D users aid the mobile user in offloading content. Here, since D2D users reuse channels, the increase in transmission power causes more interference for cellular users. We design each D2D pair to adaptively learn multichannel selection and power control strategies to maximize the capacity of the system and minimize interference. When the cached content is on the MEC server side, our optimization goal is also to optimize the system capacity of mobile cellular users.

A. DEEP REINFORCEMENT LEARNING

We use Markov decision processes (MDP) to model the optimization problem mentioned in the previous section. Generally, an MDP can be defined as a tuple {S, A, P, R, Υ} where S is a state space, A is an action space, P is a state transition probability, R is a reward function, and Υ is a discount factor. In MUE environments, the state transition probabilities and expected rewards for all states are usually unknown. Hence, we formulate that the resource allocation problem in a D2D-enabled MEC system is a model-free reinforcement framework in which the MDP has a continuous state and action space. The target of MDP is to find the optimal policy and then solve the decision-making problem to maximize the expected reward. In a reinforcement framework, the agent learns policy by interacting with the environment. We define the state and action of the environment as $s_t \in S, a_t \in A$, respectively. Generally, the agent takes action a_t from the current state s_t to a new state $s_{t+1} \in S$ and obtains an immediate reward $\mathrm{r}_t \in R$.

In this paper, we mainly adopt policy-based reinforcement learning. The method is considered to learn a parameterized policy rather than selecting actions by consulting value functions. Value functions are mainly used for policy parameter learning, not for action selection [23]. In the process, the goal of the agent is to choose a policy to maximize the expected reward. The policy is defined as $\pi_\theta(\mathrm{a}\,|\mathrm{s}) = \mathrm{P}(a_t = \mathrm{a}\,|s_t = \mathrm{s}, \theta_t = \theta) \approx \pi(\mathrm{a}\,|\mathrm{s})$. In the decision-making epoch, data are generated through the interaction between the agent and the environment to optimize the policy. Generally, the long-term expected reward is expressed as:

$$\mathrm{G}_t = \mathrm{r}_{t+1} + \Upsilon \mathrm{r}_{t+2} + \cdots = \sum_{\mathrm{k}=0}^{\infty} \Upsilon^{\mathrm{k}} \mathrm{r}_{t+\mathrm{k}+1}, \quad (\mathrm{r} \in \mathrm{R}) \tag{7}$$

where $\gamma \in [0, 1]$ denotes the discount factor. In the policy gradient method, the optimization goal is defined as follows:

$$J(\theta) = \mathrm{V}_{\pi_\theta}(s) = \mathbb{E}_{\pi_\theta}(G_t) \tag{8}$$

where $\mathrm{V}_{\pi_\theta}(s)$ is the value function of π_θ, the policy determined by θ. Here, the goal is to maximize the reward under this distribution $J(\theta)$:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} [(\sum_{t=0}^{T} log\pi_\theta(a_{i,t}\,|s_t))(\sum_{t=0}^{T} r(s_t, a_t))] \tag{9}$$

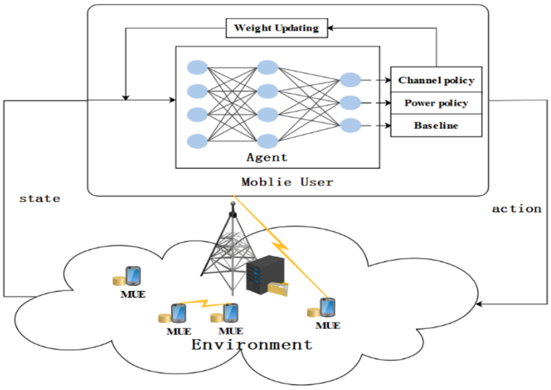

FIGURE 3. The policy gradient learning for D2D-enabled MEC communication networks.

Furthermore, the agent learns an optimal policy π^*, which is denoted as:

$$\pi^* = argmax\mathbb{E}_{\tau\sim\pi_\theta}(G_t) \tag{10}$$

where the τ represents a trace obtained by using a policy interaction. Generally, the sample approximation to the gradient is given by

$$\nabla_\theta J(\theta)=\frac{1}{N}\sum\nolimits_{i=1}^{N}[(\sum_{t=0}^{T}\nabla_\theta log\pi_\theta(a_{i,t}\,|s_t\,))(\sum\nolimits_{t=0}^{T}r(s_t,a_t))] \tag{11}$$

where $\nabla_\theta log\pi_\theta(a_{i,t}\,|s_t\,)$ is the score function. The gradient is a partial derivative of $J(\theta)$ about θ. The above equation provides us with an unbiased gradient calculation formula. However, it may have a large difference, so we employ the gradient with a baseline as follows:

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum\nolimits_{i=1}^{N}[(\sum_{t=0}^{T}\nabla_\theta log\pi_\theta(a_{i,t}\,|s_t\,)) \times (\sum\nolimits_{t=0}^{T} r(s_t,a_t) - b_t)] \tag{12}$$

where the b_t is a baseline. The b_t is varied in the environment state during the learning process. We use a network to estimate its value. The learning rule of RL is also known as the reinforce rule [24], and it can adjust the parameters of the agent to reinforce the action with high cumulative reward [25]. Therefore, there is a high baseline to acquire higher valued actions under the reinforce rule. Conversely, the baselines of low-value actions are low [26].

B. RESOURCE ALLOCATION AND POWER CONTROL METHOD

The DRL framework of the D2D-enabled MEC system is illustrated in Fig. 2. There are many MUE and D2D users in one cell. During the interaction between agents and the environment, the D2D transmitter takes action, including the select channel and power level. Next, the state, action space, reward function and update rule of channel allocation and power control problem are described in detail.

Agent: Here, each active D2D link is designed as an agent. The agent learns and makes decisions by interacting with the environment.

State: The system states mainly include three components: the communication mode of MUE $\mathcal{M}_{m,i}$, the channel state $\mathcal{C}_{c,i}$, the power level $\mathcal{P}_{p,i}$, and i refers to ith subchannel. Therefore, the system state is defined as a matrix:

$$\mathcal{S}\,(\mathrm{t}) = \{M_{m,i}\,(t)\,, \mathcal{C}_{c,i}\,(t)\,, \mathcal{P}_{p,i}\,(t)\} \tag{13}$$

where the vectors $\mathcal{M}_{m,i}$, $\mathcal{C}_{c,i}$, $\mathcal{P}_{p,i}$ are explained in detail as follows. $\mathcal{M}_m$ is defined as $\mathcal{M}_m(t) = [\mathcal{M}_D(t), \mathcal{M}_C(t)]$, and $\mathcal{M}_D \in \{0, 1\}$, $\mathcal{M}_C(\mathrm{t}) \in \{0, 1\}$. If the MUE selects the D2D mode, $\mathcal{M}_D\,(\mathrm{t}) = 1$, $\mathcal{M}_C\,(\mathrm{t}) = 0$, otherwise $\mathcal{M}_D\,(\mathrm{t}) = 0$, $\mathcal{M}_C\,(\mathrm{t}) = 1$. $\mathcal{C}_{c,i}(t)$ is defined as:

$$\mathcal{C}_{c,i}\,(t) = \begin{bmatrix} \mathcal{C}_{1,1}\,(t) & \cdots & \mathcal{C}_{1,K}\,(t) \\ \vdots & \ddots & \vdots \\ \mathcal{C}_{M,1}\,(t) & \cdots & \mathcal{C}_{M,K}\,(t) \end{bmatrix} \tag{14}$$

Here, $\mathcal{C}_{c,i}\,(t)$ indicates whether the channel is used by MUE. If yes $\mathcal{C}_{c,i}\,(t) = 1$; otherwise, $\mathcal{C}_{c,i}\,(t) = 0$. In addition, $\mathcal{P}_{p,i}\,(t)\,\epsilon[0, 24\text{ dB}]$ represents the power level in the ith subchannel, which is a continuous variable. It is defined as:

$$\mathcal{P}_{p,i}\,(t) = \begin{bmatrix} \mathcal{P}_{1,1}\,(t) & \cdots & \mathcal{P}_{1,K}\,(t) \\ \vdots & \ddots & \vdots \\ \mathcal{P}_{M,1}\,(t) & \cdots & \mathcal{P}_{M,K}\,(t) \end{bmatrix} \tag{15}$$

Action: In each learning process, there are two actions that are defined as:

$$\mathrm{A}(t) = \{\mathcal{A}_1(\mathrm{t}), \mathcal{A}_2(\mathrm{t})\} \tag{16}$$

where $\mathcal{A}_1(\mathrm{t})$ selects a channel, and $\mathcal{A}_2(\mathrm{t})$ represents selecting the power level. The actions depend on the interaction with the environment. More specifically, in our learning model, the action is two-objective. The channel selection uses a softmax output. In addition, the action of power selection is chosen stochastically from a distribution parameterized at time t by the network. Here, we adopt a Gaussian distribution.

Reward Function: Generally, the agent receives an immediate reward r_{t+1} and a new environment state s_{t+1}. In our work, there are D2D communication modes and cellular communication modes. However, in these communication modes, the reward function can be given as:

$$\mathrm{r}_t = \begin{cases} 1, & \textit{if the constraints are satisfied} \\ 0, & \textit{otherwise.} \end{cases} \tag{17}$$

Here, the proposed approach is based on different cached content types $\mathcal{C}$ of users to guarantee their communication requests and meet the QoS demands of cellular users. We define the constraints as follows:

$$\begin{cases} \gamma_i > \gamma^*, \\ \mathrm{W}_{c,i} \geq \mathrm{W}_{c,s}, \\ \mathrm{W}_{D,i} \geq \mathrm{W}_{D,s}, \end{cases} \tag{18}$$

where γ_i is the SINR of the cellular user, andγ^* is the threshold of SINR. To ensure the communication quality of the cellular link, we consider the impact on the cellular user SINR when a D2D user reuses the spectrum resource. When the SINR is greater than a threshold γ^*, the maximum power at this time is set to the transmit power of the D2D user. The W_c represents the transmission rate requirements of the different cached contents. It is defined as:

$$\mathrm{W}_{c,i} = b_i \cdot log_2(1+\gamma_i^m) \tag{19}$$

In addition, the transmission rate of a D2D user is given by

$$\mathrm{W}_{D,i} = b_i \cdot log_2(1+\gamma_i^n) \tag{20}$$

The demand for data rate is different when the requirement arrives at each time. Therefore, the agent will learn how many subchannels and how much power should be allocated to the D2D user. Our approach not only ensures the normal communication of cellular users but also maximizes the reuse of channel resources and optimizes system capacity. When the above conditions are met, the reward is 1; otherwise, a penalty is given.

C. TRAINING ALGORITHM

We adopt a policy gradient algorithm to learn resource allocation and power control. In the policy gradient algorithm, the policy parameters are updated sequentially. The deep neural network is used to train data. In the D2D-enabled MEC system, the D2D transmitter is set as an agent. The agent interacts with the environment and then takes action. During the learning process, the agent continuously updates the policy according to the policy gradient algorithm until the optimal strategy is learned. Our approach first determines the communication mode and then combines the channel selection and power selection where the agent has two different actions to achieve a goal. Training the core network is illustrated in Fig. 3. We define the number of hidden layers as 2 and the number of neurons as 256. The state $\mathcal{S}$ (t) is the input of the network, and the output is the probability distribution over all possible actions of channel selection and power selection. The mode selection is a two-label classification problem according to different contents. When the content cache $x_{k,f} = 1$, the content is cached in the MUE where the agent selects the D2D mode. In addition, the $x_{k,f} = 0$ denotes that the content c_f is cached in the local MEC server, and the agent selects the cellular mode. In each episode, the main goal of the agent is to learn the policies of the channel selection and power selection. In the training process, the optimal actions are unknown, and the good or bad learning result is provided via the reward. Furthermore, there are three loss functions, and the loss of the baseline is given by

$$\mathcal{L}_b = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=0}^{T_n}[b_t^i - r_t^i]^2 \tag{21}$$

where b_t^i is the estimated value of the reward, and r_t^i represents the one-step reward. N denotes the number of samples, T represents the length of the trace, and i denotes the ith trace.

The loss function of channel selection is

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T_n}\pi_{1,t}^{i}[a_t^i \,\Big|\, s_t^i]r_t^i \tag{22}$$

where $\pi_{1,t}^i$ denotes the probability of selecting a channel, and r_t^i is the one-step reward. The loss function of the power selection is

$$\mathcal{L}_p = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T_n}\pi_{2,t}^{i}[a_t^i \,\Big|\, s_t^i]r_t^i \tag{23}$$

where $\pi_{2,t}^i$ denotes the probability of selecting a power. The total loss is given by

$$\mathcal{L} = \mathcal{L}_b + \mathcal{L}_c + \mathcal{L}_p \tag{24}$$

We optimize the cross-entropy loss to train the action network and backpropagate the gradients through the core network. The update rules are shown in Algorithm 1 and Algorithm 2. Algorithm 1 shows the procedure of resource allocation and power control. Algorithm 2 mainly describes the update steps of the policy gradient. In Algorithm 1, we run Algorithm 2 to learn channel selection and power control policies. The D2D user's method of selecting the channel and power can ensure the edged cache of the MUE and avoid the interference of the D2D-enabled MEC system.

V. EXPERIMENT AND EVALUATION

In this section, we present experiments to evaluate our proposed joint channel selection and power control method. The experiments are conducted in an Ubuntu operating system (CPU Intel core i7-4790 3.6 GHz; memory 16GB, GPU NVIDIA Quadro K2200, which contains 640 CUDA computing core units and 4GB graphics memory).

As illustrated in Fig. 2, we consider a cell where the MUEs are deployed based on the spatial Poisson process. The D2D mode and cellular mode are selected among the active mobile devices, and each MUE can construct one D2D link or cellular link. In addition, we adopt the Manhattan case detailed in 3GPP TR 36.885 to set the simulation [27]. In one time-slot (0.5 ms), the radio resource is organized in a number of downlink RBs with 180 kHz per RB. In addition, we set

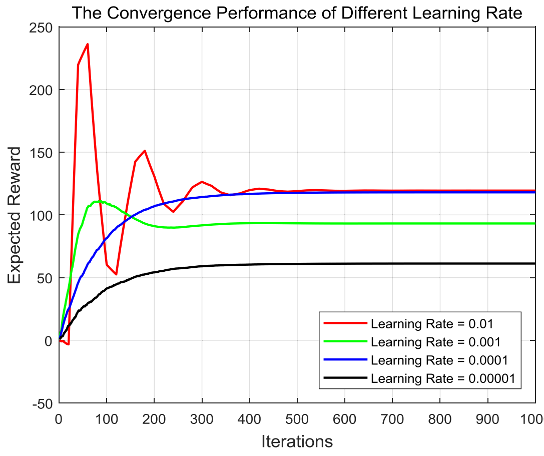

FIGURE 4. The convergence performance of different learning rate.

Algorithm 1 Channel Allocation and Power Control Method

```
begin
    Initialization:
        For $t = 0, t = (t_1, ..., t_N)$
            Randomly create a state matrix: $S(t)$
            Create an action matrix: $A(t) = 0$
        Initialize D2D-enabled MEC system model parameter
            Determine the communication mode of MUE according to cached content
                D2D mode (if $x_{k,f} = 1$)
                Cellular mode (if $x_{k,f} = 0$)
        D2D user randomly select a first channel and power level
        End for
    Processing:
    Loop:
        For t in $T$, do
    (1) Selected channel $C$ and power $P$
    (2) Calculate:
            $\gamma_{\ell}^{i}$ of the $i$th channel of the cellular user
            $\gamma_{\jmath}^{i}$ of the $i$th D2D pair
            System capacity
    (3) Check SINR to guarantee QoS of users according to Constraints
    (4) Run Algorithm 2, learning channel and power selection policy
    (5) If the D2D transmission restarts in this time slot
            End if
        End for
        Set $t = t + 1$
        Create a new potential state matrix: $S(t + 1)$
        End loop
end
```

the number of PRBs to 10. Generally, there are two types of D2D communication, namely, in-band or out-band communication. In all simulations, we set the D2D communication distance to 50m. Hence, the type of D2D communication is in-band communication. In addition, the D2D communication connections are supported through cellular (Uu) and sidelink radio interfaces, respectively. In this experiment, the deep neural network for each agent consists of 2 hidden layers, whose number of neurons is 256. In our D2D-enabled MEC system, the MEC mainly performs content caching and content forwarding. The MEC server is mainly deployed in the base station and provides various functions through the mobile edge computing application. The MEC server here is mainly a multi-user single server because only one MEC edge server is arranged after each cell base station. The main simulation parameters are presented in Table 1. We evaluate our approach on the above parameter settings.

Algorithm 2 Learning Algorithm of the Policy Gradient

```
begin
    Initialization:
        $t = 0$, the network parameter $\theta$
        $A_t$ is the action of D2D user, $S_t$ is the environment state
    for $i$ in $N$ do
        Observe $S_t$, and initialize D2D transmitter power
        For $t = 0, \ldots, T_n$ do
            Select channel according $\pi_{1,t}^{i}[a_t^i | s_t^i]$
            Select power according $\pi_{2,t}^{i}[a_t^i | s_t^i]$
        Obtain the trace $r_t^i$ and observe state $s_{t+1}^{i}$ according to system capacity
        Repeat this process for next state $s_{t+1}$
        End
    end
        Calculate loss $\mathcal{L}_b, \mathcal{L}_c, \mathcal{L}_p$
        Calculate total loss $\mathcal{L}$
        Use a gradient descent to update parameter $\theta$ and
        minimize loss $\mathcal{L}$
end
```

TABLE 1. **The parameter of the simulation [28].**

Parameter	Value
Cell radius	500m
D2D communication distance	50m
D2D transmit power	[0-23dB]
Resource block bandwidth	180kHz
Transmission power	24dB
Noise power/RB	-116dB
path loss model between BS and users	15.3+37.6log(d(km))(dB)
path loss model between BS and users	28+40log10(d(km))(dB)
Macro BS antenna gain	17dBi
User antenna gain	4dBi
Learning rate	0.2
Discount factor	0.99
Server location of MEC server	BSs
Distance of MEC server to end users	Small (< 500m)
System management of MEC	Distributed
Storage space of MEC server	5%*5

First, we carry out numerical experiments under various settings of learning rates to validate the proposed work. We set that the number of MUEs to 5. According to the cache requirement, one MUE is selected to become a D2D user, which reuses the channel of one cellular user. We assume that the power level is in the range of (0, 24) (dB). As shown in Fig. 4, we study the convergence performance of the

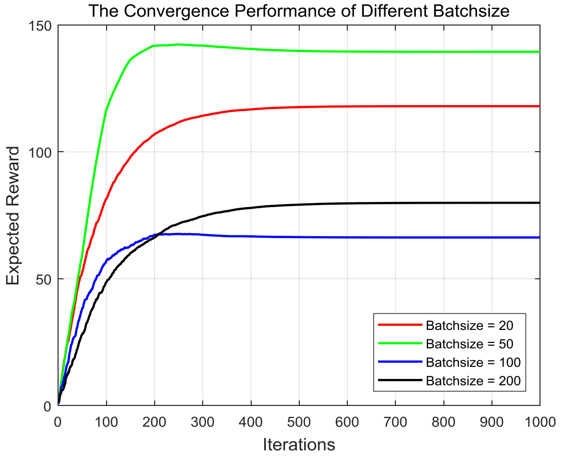

FIGURE 5. The convergence performance of different batch size.

proposed algorithm at different learning rates. The learning rates are 0.01, 0.001, 0.0001, and 0.0001. It can be seen that there is a similar trend in Fig. 4. When the MUE is in the D2D communication mode, the figure shows that the learning rate causes the algorithm to converge to an optimum. The convergence time is different in different learning rates. As seen from the figure, when the learning rate is 0.0001, the convergence performance is the best. Hence, in the following simulations, we set the learning rate to 0.0001 because its convergence performance is better than the others. In this figure, the initially expected reward is low because the agent explores the optimal strategy, and then all curves gradually rise and tend to stabilize.

As shown in Fig. 5, we compare the expected rewards of users in four batch sizes. We set the learning rate to 0.0001, and the batch sizes to 20, 50, 100, and 200, respectively. It is shown in the figure that the expected reward on different batch sizes is increased. However, the small or large batch size does not regularly affect the expected reward. Additionally, under these conditions, the convergence time is different. Since a different batch size requires different training duration and convergence speed, we adopt the batch size (= 50) in the following experiment because under this batch, the expected reward is the largest, and it consumes less time.

Fig. 6 depicts the cross-entropy loss function $\mathcal{L}$ of our policy gradient network. Here, we set the learning rate is 0.0001 and the batch size is 50. In the figure, we enable to observe the simulated variations in the loss function defined as in (24), which reveals that the convergence of our proposed algorithm can be ensured. When the learning network first started training, the value of the loss was relatively large, and the network was in the update phase. As the number of training processes increases, the value of loss gradually decreases. Specifically, the training loss $\mathcal{L}$ gradually decreases and stabilizes after training 200 interactions, whose fluctuation is mainly due to the random sampling of training data. It means that our algorithm automatically updates its decision policy and converge to the new optimal value. The figure shows

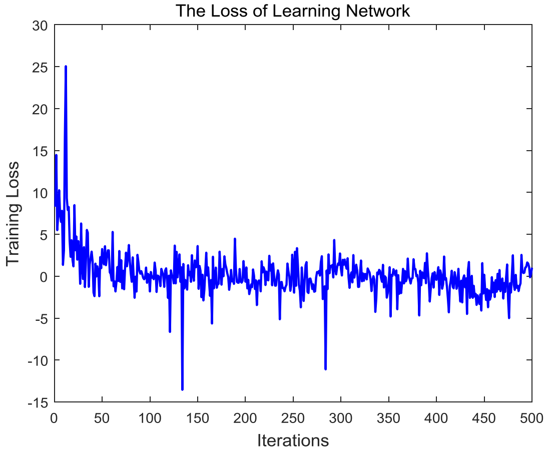

FIGURE 6. The training loss of the policy gradient network.

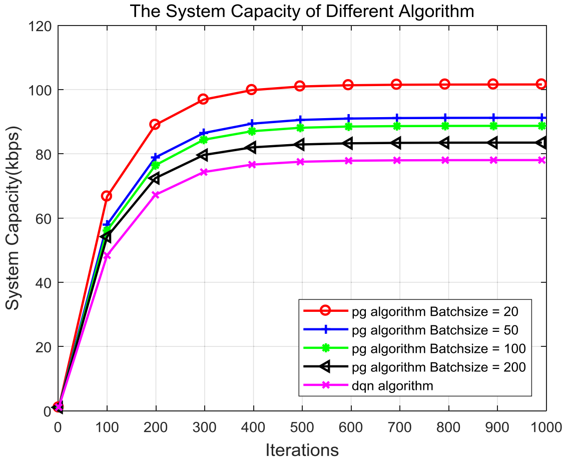

FIGURE 7. The system capacity of different algorithms.

that the policy gradient method has good convergence in joint resource allocation and power selection, and the convergence time is short.

In Fig. 7, we further show the system capacity in different resource allocation algorithms s at each step in the same episode. Two algorithms are also simulated for comparison. They are the proposed policy gradient algorithm and the deep Q-network (DQN) algorithm. Here, we set the learning rate to 0.0001. In our proposed algorithm, the batch sizes are 20, 50, 100, 200. By making a comparison between our proposed algorithm and the DQN algorithm in the same conditions, it can be found that the performance of former outperforms the latter. Our method shows its effectiveness on maximizing the system capacity faced with dynamic and complex wireless environments because our power selection method is a decision made in a continuous state space, but DQN is a choice made on discrete power. Hence, our method can learn more power control strategies. We observe that even though the proposed method has a batch size of 200, resulting in the lowest system capacity, it achieves better performance than DQN. It demonstrates that our approach allows to significantly reinforce policy learning when the agent interacts actively with the environment. Furthermore, the Gaussian distribution is used as the parameterized policy to generate stochastic actions of power selection, and softmax is used to perform channel selection. In continuous value states and action variables, we use a policy-gradient approach to learn the optimal policy through interacting with the environment.

The experiments prove that cellular communication and D2D communication can coexist and share RBs for their data transmissions. The proposed joint resource allocation and power selection method can maximize system capacity while avoiding interference. During the learning process, the agent continuously updates the strategy to learn how to allocate resources and select power. Based on the simulation results, each agent can learn how to meet the cellular communication constraints while avoiding interference with D2D-enabled MEC communications and maximizing the total system capacity.

VI. CONCLUSION

Information-centric wireless networking (ICWN) has become one of the most important networking paradigms in future 5G wireless networks. In the recent ICWN paradigm, D2D-enabled MECs can collaboratively cache popular content on a variety of nearby devices, which helps to improve spectral efficiency and reduce traffic congestion. This paper introduced a novel resource allocation and power control method with the policy gradient in a comprehensive D2D-enabled MEC system of IWCN. Specifically, we have modeled this problem as model-free reinforcement learning. In addition, due to the unknown channel environment and ever-changing transmission power, we updated the parameters with the regular policy gradient method. The Gaussian distribution was used as the parameterized policy to generate stochastic actions of power selection, and softmax was used to perform channel selection. Numerical results show that the method has good convergence.

REFERENCES

[1] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[2] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.

[3] J. Guo, B. Song, F. R. Yu, Y. Chi, and C. Yuen, "Fast video frame correlation analysis for vehicular networks by using CVS–CNN," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6286–6292, Jul. 2019. doi: 10.1109/TVT.2019.2916726.

[4] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11339–11351, Dec. 2017.

[5] D. Kutscher and S. Eum, *Information-Centric Networking (ICN) Research Challenges*, document RFC 7927, Jul. 2016.

[6] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May 2015.

[7] A. A. Ateya, A. Muthanna, and A. Koucheryavy, "5G framework based on multi-level edge computing with D2D enabled communication," in *Proc. IEEE 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, p. 1.

[8] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10190–10203, Nov. 2018.

[9] S. Wang, M. Zafer, and K. K. Leung, "Online placement of multi-component applications in edge computing environments," *IEEE Access*, vol. 5, pp. 2514–2533, 2017.

[10] R. Chai, J. Lin, M. Chen, and Q. Chen, "Task execution cost minimization-based joint computation offloading and resource allocation for cellular D2D MEC systems," *IEEE Syst. J.*, to be published.

[11] L. Song, D. Niyato, Z. Han, and E. Hossain, "Game-theoretic resource allocation methods for device-to-device communication," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 136–144, Jun. 2014.

[12] P. TalebiFard, "An information centric networking approach towards contextualized edge service," in *Proc. 12th Annu. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2015, pp. 250–255.

[13] Y. He, C. Liang, F. R. Yu, and V. C. M. Leung, "Integrated computing, caching, and communication for trust-based social networks: A big data DRL approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[14] H. Xing, L. Liu, J. Xu, and A. Nallanathan, "Joint task assignment and resource allocation for D2D-enabled mobile-edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4193–4207, Jun. 2019.

[15] X. Du, M. Zhang, K. E. Nygard, S. Guizani, and H.-H. Chen, "Self-healing sensor networks with distributed decision making," *Int. J. Sensor Netw.*, vol. 2, nos. 5–6, pp. 289–298, 2007.

[16] X. Du, Y. Xiao, S. Ci, M. Guizani, and H.-H. Chen, "A routing-driven key management scheme for heterogeneous sensor networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Glasgow, U.K., Jun. 2007, pp. 3407–3412.

[17] R. Zhang, L. Song, Z. Han, X. Cheng, and B. Jiao, "Distributed resource allocation for device-to-device communications underlaying cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 1889–1893.

[18] Y. Qiu, Z. Ji, Y. Zhu, G. Meng, and G. Xie, "Joint mode selection and power adaptation for D2D communication with reinforcement learning," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Lisbon, Portugal, Aug. 2018, pp. 1–6.

[19] M. Zhao, Y. Wei, M. Song, and G. Da, "Power Control for D2D communication using multi-agent reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Beijing, China, Aug. 2018, pp. 563–567.

[20] T. Peng, Q. Lu, H. Wang, S. Xu, and W. Wang, "Interference avoidance mechanisms in the hybrid cellular and device-to-device systems," in *Proc. IEEE 20th Int. Symp. Pers., Indoor Mobile Radio Commun.*, Tokyo, Japan, Sep. 2009, pp. 617–621.

[21] P. Abbeel and J. Schulman, "Deep reinforcement learning through policy optimization," in *Proc. NIPS*, 2016.

[22] T. Hou, G. Feng, S. Qin, and W. Jiang, "Proactive content caching by exploiting transfer learning for mobile edge computing," in *Proc. IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–6.

[23] R. S. Sutton and G. A. Barto, *Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1998.

[24] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.

[25] Y. Zhang, B. Song, S. Gao, X. Du, and M. Guizani, "Monopolistic models for resource allocation: A probabilistic reinforcement learning approach," *IEEE Access*, vol. 6, pp. 49721–49731, 2018.

[26] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[27] *3rd Generation Partnership Project: Technical Specification Group Radio Access Network: Study LTE-Based V2X Services: (Release 14)*, document 3GPP TR 36.885 V2.0.0, Jun. 2016.

[28] S. Nie, Z. Fan, M. Zhao, X. Gu, and L. Zhang, "Q-learning based power control algorithm for D2D communication," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.