IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 25, 2021, accepted March 17, 2021, date of publication March 22, 2021, date of current version March 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3067662

Wireless Access Control in Edge-Aided Disaster Response: A Deep Reinforcement Learning-Based Approach

HANG ZHOU[1], **XIAOYAN WANG**[1], **(Senior Member, IEEE)**, **MASAHIRO UMEHIRA**[1], **(Member, IEEE)**, **XIANFU CHEN**[2], **(Member, IEEE)**, **CELIMUGE WU**[3], **(Senior Member, IEEE)**, **AND YUSHENG JI**[4], **(Senior Member, IEEE)**

[1]Graduate School of Science and Engineering, Ibaraki University, Hitachi 316-8511, Japan
[2]VTT Technical Research Centre of Finland, 90570 Oulu, Finland
[3]Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu 182-8585, Japan
[4]Information Systems Architecture Research Division, National Institute of Informatics, Tokyo 101-8430, Japan

Corresponding author: Xiaoyan Wang (xiaoyan.wang.shawn@vc.ibaraki.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research(C) under Grant 20K11764, in part by the Telecommunications Advancement Foundation, and in part by the Research Organization of Information and Systems (ROIS) National Institute of Informatics (NII) Open Collaborative Research under Grant 2020-20FA01.

ABSTRACT The communication infrastructure is most likely to be damaged after a major disaster occurred, which would lead to further chaos in the disaster stricken area. Modern rescue activities heavily rely on the wireless communications, such as safety status report, disrupted area monitoring, evacuation instruction, rescue coordination, etc. Large amount of data generated from victims, sensors and responders must be delivered and processed in a fast and reliable way, even when the normal communication infrastructure is degraded or destroyed. To this end, reconstructing the post-disaster network by deploying MDRU (Movable and Deployable Resource Unit) and relay unit at edge is a very promising solution. However, the optimal wireless access control in this heterogeneous hastily formed network is extremely challenging, due to the frequent varying environment and the lack of statistics information in advance in post-disaster scenarios. In this paper, we propose a learning based wireless access control approach for edge-aided disaster response network. More specifically, we model the wireless access control procedure as a discrete-time single agent Markov decision process, and solve the problem by exploiting deep reinforcement learning technique. By extensive simulation results, we show that the proposed mechanism significantly outperforms the baseline schemes in terms of delay and packet drop rate.

INDEX TERMS Disaster response network, deep reinforcement learning, wireless access control, network edge.

I. INTRODUCTION

There is a shocking increase in the number of disasters during the past few decades [1], humans have had to deal with dire consequences of many disasters frequently. Timely and smooth information exchange among victims, responders and outside world after the occurrence of disasters, is crucial for effective crisis mitigation and life rescue. Most of the information exchange and delivery, e.g., safety status report, disrupted area monitoring, evacuation instruction and rescue coordination, are carried out by wireless communications [2]. However, major disaster would damage large parts of the wireless communication infrastructure, which inevitably leads to further chaos in the affected area. For instance, ruinous damages on communication infrastructure happened in the 2011 Great East Japan Earthquake, that approximately 29000 base stations were damaged after the earthquake. Although the mobile carriers have expended a great deal of effort to recover the network infrastructure immediately after the earthquake occurred, it still took a month and a half [3]. It is well known that the first 72 hours after the disaster are critical to mitigate the damage and save lives.

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

Therefore, waiting the restoration of the normal network infrastructure is too late for disaster response, and fast and efficient post-disaster network reconstruction is urgently needed.

To provide communications during disaster relief, the researches on hastily formed networks [4] have attracted lots of attentions. It has the potential to be rapidly deployed in an ad-hoc way when normal communications infrastructure has been degraded or destroyed after a disaster. Specifically, the NTT (Nippon Telegraph and Telephone Corporation) research teams have proposed a network recovery approach by using the MDRU (Movable and Deployable Resource Unit) [5] to reconstruct the network in disaster area on demand. As a vehicle-type resource unit, MDRU is capable of establishing the network connection via satellite communication or optical fiber to provide Internet service for users. It can work for more than 5 days powered by the battery and be easily deployed in the disaster affected area [6]. Furthermore, since modern disaster response often requires to process various collected information such as image, voice and video, and correspond accordingly, the MDRU can also function as an intelligent edge unit, which has the capability to handle the collected information from users or devices in the area. By providing this computing capabilities at edge, MDRU can significantly reduce the resource occupation at central cloud server and traffic volume in the whole network.

However, considering the extremely large disaster affected area, and the limited service range of MDRU (i.e., typically 500 meters radius), it is impractical to use MDRUs only to cover the whole affected area. Moreover, due to the comparatively high cost and tight time limitations, immediately deploying a great number of MDRUs after the disaster occurred would be extraordinary difficult. Therefore, for the purpose of recovering the communications in a cost-efficient way, a heterogeneous disaster response network architecture which combines MDRU and multiple relay units has been investigated [7]–[10]. The relay units are capable of extending the network coverage from MDRU by exploiting multi-hop transmission techniques.

Under this network architecture, various kinds of User Equipments (UE) can exchange their messages or report their monitored information. For instance, victims can use their wireless devices to request help from responders or obtain instructions from the government agency; and rescue drones or robots with wireless transceivers can upload images or videos to facilitate the analytic tasks [11] in the MDRU for the detection of survivors or assessment of damage status. Without the loss of generality, it is reasonable to assume that UEs may move continuously and generate data either periodically or on-demand which depends on the applications, and they have the energy harvesting capability to prolong the operating time [12] due to no reliable power supply in disaster area. In this type of edge-aided disaster response network scenario, the optimal wireless access control of UE is challenging, due to the environment dynamics in terms of the channel, movement, packet and energy status. Moreover, the survived communication infrastructure and traffic volume in post-disaster scenario would significantly differ from that in normal time, therefore the conventional optimization based methods [13] are difficult to be applied since there is no statistical information in advance. Motivated by the aforementioned research issues, in this work, we propose a DRL based wireless access control mechanism for edge-aided disaster response networks. The wireless access control of UE in the edge-aided network is modeled as an MDP, and the proposed mechanism learns an access policy to optimize its performance via the interaction with the wireless environment. The proposed mechanism could adapt to the dynamic post-disaster environment without a priori knowledge. Our main contributions in this paper are summarized as follows.

- We formulate the wireless access problem in an edge-aided disaster response network as a single-agent MDP. We propose a DRL based wireless access control approach by letting agent UE interact with and learn from the unknown post-disaster environment.
- To balance the delay and drop rate, we apply a buffer in UE which enables the UE works in a delay tolerant way. The UE does not need to upload the monitored data to MDRU immediately, it can learn the optimal time for transmitting by taking into consideration the channel and energy status.
- Compared with other baseline schemes, we validate the practicability of the proposed approach by extensive simulations. The results show that the proposed approach significantly outperforms the baseline schemes in terms of packet delay and drop rate, in any network settings.

The remainder of this paper is structured as follows. In Section II, we provide background on DRL and an overview of related work. we In Section III, we introduce the considered edge-aided disaster response network and the system model used in this paper. In Section IV, we formally formulate the problem as a single-agent MDP and discuss its general solution. In Section V, we present the proposed DRL based wireless access control approach in details. In Section VI, we present the simulation results under various network settings to compare the performance of the proposed approach against other baseline schemes. Finally, we draw the conclusions and discuss the future work in Section VII.

II. RELATED WORK

In dynamic network scenario whose statistic information is hard to obtain in advance, the reinforcement learning technique [14] has been widely investigated to optimize its wireless resource allocation. Reinforcement learning technique is one of the three main paradigms in machine learning, besides supervised learning [15] and unsupervised learning [16]. In [17], Xu *et al.* proposed an efficient reinforcement learning-based resource management algorithm, which learns on-the-fly the optimal policy of dynamic workload offloading and autoscaling at network edge. Ortiz *et al.* modeled an energy harvesting point-to-point communication scenario as a MDP, and found the policy that maximizes the throughput by applying Q-learning [18]. As stated in [17]–[21], reinforcement learning based approaches could adapt to dynamic network environment, however, the convergence of the learning results become extremely slow when the environmental state space is large.

To deal with the state space explosion problem in complex scenario, DRL [22] based mechanism have been investigated. DRL was originally designed to learn from high-dimensional input (e.g., raw images) to formulate control policy. Specifically, DRL and its variants can remove the influence of enormous state space by using neural network, and work in conjunction with the unknown environmental statistics [22], [23]. Chen *et al.* proposed a double deep Q-network-based strategic computation offloading algorithm to learn the optimal policy without knowing a priori knowledge of network dynamics [24]. In [25], Cao *et al.* proposed a UE-driven DRL based scheme to let each UE be able to access a proper base station intelligently to enhance the long-term system throughput and avoid frequent handovers. In [26], Sun *et al.* proposed a dynamic resource reservation and DRL-based autonomous virtual resource slicing framework for the next generation radio access network. Additionally, regarding the wireless access control problems, in [27]–[29], the DRL based routing selection approaches have been proposed, and in [30], [31], the DRL based dynamic power allocation approaches have been investigated. In our previous work [32], we also proposed a DRL-based radio access control mechanism for disaster response network, but without the consideration of UE's mobility and buffering capacity. These previous researches have shown that DRL based schemes can learn the features of complex and changing wireless access networks to enhance the performance. However, none of them could be applied in a dynamic wireless access network to solve the combined optimization problem of transmission timing, routing and energy allocation.

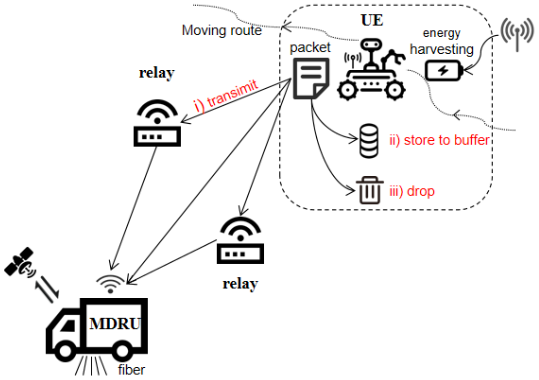


FIGURE 1. An illustration of edge-aided disaster response networks.

III. SYSTEM MODEL

As depicted in Fig. 1, we consider an edge-aided disaster response network consisting of single MDRU, a set of relay units $\mathcal{N} = \{1, \cdots, N\}$, and multiple UEs. The MDRU can either work as a gateway since it is connected to the Internet infrastructure via fiber cable or satellite communications; or work as an edge unit since it has the computation capability to directly process the information collected from the UEs. Multiple relay units are deployed around MDRU to extend its network coverage area. The relay node forwards the packets from UEs to MDRU by using orthogonal channels. Specifically, the network has $N + 1$ available channels with the same bandwidth B, i.e., N channels for relay connections and 1 channel for the direct MDRU connections. Both the MDRU and relays are assumed to be fixed in the disaster area. There are multiple mobile UEs exist in this disaster response network, and they intend to transmit the collected environmental information to MDRU. As illustrated by Fig. 1, the generated packets could be processed in three possible ways. i) The packet is transmitted to MDRU, which could be either transmitted to MDRU directly, or forwarded through one of the relays. ii) The packet is stored to buffer and might be transmitted later. iii) The packet is dropped directly. Since we consider a post-disaster scenario, we assume that the UE has energy harvesting capability to extend its lifetime, and the energy used for transmission could be dynamically allocated. Therefore, in this network model, we consider a combined wireless access control problem for UE, which consists of transmission timing control (i.e., transmit, drop or store to buffer), routing selection and energy allocation. In the following part, we present the model by focusing on a representative UE. All the notations used in the article are summarized in Table 1.

The time is discretized into time slots with the same duration τ, and indexed by a positive integer k. The location of UE at time slot k is denoted by $L^k \in \mathcal{L}$. The number of harvested energy units at time slot $k - 1$ is denoted as E_e^{k-1}, which is i.i.d. and obeys a Poisson distribution with average harvesting rate $\lambda_e \in [0, 1]$. The energy harvested from the environment is stored in an energy storage, and the total energy unit $H^k \in \mathcal{H}$ in the energy storage at time slot k evolves as

$$H^k = \min\{H^{k-1} - E^{k-1} + E_e^{k-1}, \bar{H}\}, \tag{1}$$

where H^{k-1} denotes the residual energy unit at time slot $k-1$, E^{k-1} denotes the allocated energy for transmitting at time slot $k - 1$, and $\bar{H}$ is the upper bound of UE's energy storage.

The packets generated at UE could be either the status reports or environmental monitoring results, which depend on the UE's type and application. The generated packet is assumed to be a Bernoulli random variables with parameter $\lambda_p \in [0, 1]$. Specifically, the generation of packet at the beginning of time slot k is denoted as $P^k \in \mathcal{P}$. $P^k = 1$ denotes a packet is generated, $P^k = 0$ otherwise. Therefore, the probability of packet generation at time slot k is expressed by

$$Pr\{P^k = 1\} = 1 - Pr\{P^k = 0\} = \lambda_p, \tag{2}$$

TABLE 1. Major notations used in the paper.

N/$\mathcal{N}$	number/set of relay
τ	time duration of one scheduling slot
L^k	The location of UE at times slot k
$\mathcal{L}$	The location set of UE
H^k	residual energy unit at time slot k
$\mathcal{H}$	set of energy status
$\bar{H}/\bar{J}$	the upper bound of UE's energy storage/buffer size
E^k	allocated energy for transmitting at time slot k
E_e^k	harvested energy units at time slot k
λ_e	average harvesting rate
λ_p	packet generated rate
λ_p	packet generated rate
P^k	the generation of packet at time slot k
$\mathcal{P}$	the generation of packet at time slot k
τ_n	average number of arrived packets prior at relay n
J^k	the number of packets in the buffer at time slot k
$\mathcal{J}$	set of number of packets in the buffer
T^k	the transmission policy of UE at time slot k
$t_{(u),(n)}^k/t_{(u),(m)}^k$	transmission delay from UE to relay/MDRU at time slot k
R^k	data rate at time slot k
μ	the size of a packet
I^{-1}	0interference and additive background noise
g^k	channel state space at time slot k
$\mathcal{G}$	channel state space
λ	wavelength
B	bandwidth of a frequency band
d^k	the distance between transceiver in time slot k
w	the random variation
σ_n^2	zero mean Gaussian random variable
m_n	the number of arrived packets earlier than UE in relay n
$\mathcal{S}$	state space
s^k	the state space in time slot k
ϕ, ϕ^*	the control/optimal policy
ϕ_T, ϕ_E	the transmission/energy policy
r^k	cost function
ρ	weight factors of dropping packet
ξ	weight factors of saving packet
$\hat{s}$	initial state space
$V(s, \phi)$	value function
γ	a time discount factor
α	learning rate
θ^k	Q network parameters in time slot k
$\tilde{\theta}$	target network parameters

where $Pr\{\cdot\}$ denotes an event's probability. As we mentioned, the generated packet at UE would be transmitted to MDRU either by direct MDRU uplink or relayed by one of the relay units $n \in \mathcal{N}$. Intuitively, the routing selection depends on the UE's position, channel states, and traffic volume. We assume that the MDRU and relay units handle the packets from different UEs one by one based on FIFO (First-In First-Out) rule. We model the packet arrival in MDRU and relay units as a Poisson process, and the average number of arrived packets prior to agent UE is τ_n.

We apply a buffer at UE to enable the UE store its packets when there is no energy or the channel state is poor. The packets stored in the buffer and new generated packets could be bundled and processed. The buffer size of UE is denoted by $\bar{J}$. We use $J^k \in \mathcal{J}$ to represent the number of packets in the buffer in the beginning of k-th time slot, where $J^k \leq \bar{J}$. Notice that the packet would be lost when the stored packet J^k is already reaches its upper bound $\bar{J}$.

We define $T^k \in \{-2, -1\} \bigcup \{0\} \bigcup \mathcal{N}$ to express the transmission policy of UE at time slot k. To be specific, $T^k = -1$ denotes that the generated packet is dropped, $T^k = -2$ denotes that the generated packet is stored in buffer, $T^k = 0$ denotes that the packet is directly transmitted to MDRU, $T^k = n$ denotes that the packet is forwarded to relay unit n. At the beginning of each time slot k, the agent UE makes a combined wireless access control decision which includes transmission policy T^k and energy allocation E^k.

The transmission delay of the packet could be calculated as follows by considering two cases separately. Regarding the case of packet is relayed by relay unit n, i.e., $T^k = n$, the total delay consists of the transmission delay from UE to relay unit n, and that from relay unit n to MDRU. The transmission delay from UE to relay unit n in time slot k, $t_{(u),(n)}^k$, can be derived by

$$R_{(u),(n)}^k t_{(u),(n)}^k = (J^k + P^k)\mu, \tag{3}$$

where $R_{(u),(n)}^k$ is the data rate from UE to relay unit n, and μ is the size of a packet [24]. The data rate $R_{(u),(n)}^k$ could be obtained by

$$R_{(u),(n)}^k = B \log_2 \left(1 + I^{-1} g_{(u),(n)}^k \frac{E^k}{t_{(u),(n)}^k} \right), \tag{4}$$

where I^{-1} is the received average power of interference and additive background noise, $g_{(u),(n)}^k \in \mathcal{G}$ is the channel gain between UE and relay unit n during time slot k. We assume that $g_{(u),(n)}^k$ keeps constant in one time slot, but changes from slot to slot. The channel gain in dB is modeled by $20 \log_{10} \left(\frac{\lambda}{4\pi d^k} \right) + w$, where λ is the wavelength, d^k is the distance between transceiver in time slot k which varies with UE's location L^k, and w accounts for the random variation to show the effect of fading which is set as a zero mean Gaussian random variable with variance σ_n^2.

During time slot k, relay unit n must handle the packets that arrived earlier than agent UE's first. By taking into consideration the number of arrived packets earlier than that of agent UE at relay unit n during time slot k, i.e., m_n^k, the total time $t_{(n),(m)}^k$ for UE's packet waiting and forwarding from relay unit n to MDRU is derived by

$$R_{(n),(m)}^k t_{(n),(m)}^k = (2m_n^k + J^k + P^k)\mu, \tag{5}$$

$$R_{(n),(m)}^k = B \log_2 \left(1 + I^{-1} g_{(n),(m)}^k \frac{E'}{t_{(n),(m)}^k} \right), \tag{6}$$

where $R_{(n),(m)}^k$ is the data rate of relay unit n to MDRU. For simplicity, we assume that relay unit allocates same energy E' for all the packets, and the channel state $g_{(n),(m)}^k \in \mathcal{G}$ between relay unit n and MDRU could be calculated similar as $g_{(u),(n)}^k$. Finally, the total delay in this relay case is the summation of $t_{(u),(n)}^k$ and $t_{(n),(m)}^k$.

Regarding the case that packet is directly transmitted to MDRU, i.e., $T^k = 0$, the transmission delay $t_{(u),(m)}^k$ from UE to MDRU could be derived by

$$R_{(u),(m)}^k t_{(u),(m)}^k = (m_m^k + J^k + P^k)\mu, \tag{7}$$

$$R_{(u),(m)}^k = B \log_2 \left(1 + I^{-1} g_{(u),(m)}^k \frac{E^k}{t_{(u),(m)}^k} \right), \tag{8}$$

where m_m^k is the number of arrived packets earlier than that of agent UE at MDRU during time slot k. Notice that the transmission delay calculated previously does not include the waiting time in the buffer, which could be easily obtained by multiply the time slot duration τ with the number of time slots that the packets stayed in the buffer.

IV. PROBLEM FORMULATION

In this section, we formulate the considered wireless access control problem at the edge-aided disaster response networks as a discrete-time single-agent MDP. To formulate a MDP problem, we define the state s^k in state space $\mathcal{S}$, the control policy ϕ, the state transition function from s^k to s^{k+1}, and the reward function r^k in the following parts.

A. STATE SPACE

The current state s^k in time slot k belongs to $\mathcal{S} = \mathcal{G}\times\mathcal{L}\times\mathcal{H}\times\mathcal{P}\times\mathcal{J}$, which contain the location L^k of agent UE, the channel gain g^k, the energy unit H^k, the packet arrivals P^k, and the number of packet J^k in the buffer. Here, the channel gain's state g^k equals $\{g^k_{(u),(1)}, g^k_{(1),(m)} \cdots g^k_{(n),(m)}, g^k_{(u),(m)}\}$ for each time slot k, whose size depends on the number of relay units n.

Given a stationary control policy ϕ and based on the UE's mobility, packet arrivals and energy harvesting models, the state sequence $\{s^k : k \in \mathbb{N}\}$ is a controlled Markov chain with the state transition probability as

$$\begin{aligned} &Pr\{s^{k+1}|s^k, \phi(s^k)\} \\ &= \left(\prod_{n=1}^{N+1} Pr\{g^{k+1}_{(u),(n)}|g^k_{(u),(n)}\}\right) \\ &\quad \cdot Pr\{L^{k+1}|L^k\}\cdot Pr\{H^{k+1}|H^k, \phi(s^k)\} \\ &\quad \cdot Pr\{P^{k+1}|P^k\}\cdot Pr\{J^{k+1}|J^k, \phi(s^k)\}. \end{aligned} \tag{9}$$

B. ACTION SPACE

The control policy $\phi(s^k) = (\phi_T(s^k), \phi_E(s^k)) = (T^k, E^k)$ is a joint wireless access control action applied in agent UE, where $\phi = (\phi_T, \phi_E)$ is a stationary wireless access policy including stationary transmission policy and energy allocation policy. Formally, the action space $a^k = \{T^k, E^k\}$ consists of transmission control and energy allocation. The transmission control T^k is stated in Section III, and the energy allocation E^k depend on the total energy unit H^k at time slot k.

C. REWARD FUNCTION

The goal of our research is to minimize the summation of total delay and packet drop rate. We define an immediate cost function to represent r^k, which quantifies the wireless access control experience of the agent UE in time slot k that is formulated by

$$r^k = t^k + \rho\cdot p^k + \xi\cdot J^k, \tag{10}$$

where t^k is the total delay for transmission at time slot k, $\rho\cdot p^k$ and $\xi\cdot J^k$ represent the penalties of dropping packet and storing packets in buffer respectively, where $p^k = 1$ if the packet is dropped and $p^k = 0$ otherwise, ρ and ξ are non-negative weight factors. Based on the edge-aided disaster response network model addressed previously, the total transmission delay t^k at time slot k could be calculated by

$$t^k = \begin{cases} 0, & T_k = -1, -2 \\ t^k_{(u),(m)}, & T_k = 0 \\ t^k_{(u),(n)} + t^k_{(n),(m)}, & T_k = n \end{cases} \tag{11}$$

In this network, wireless access control approach should be optimized to minimize the whole long-term cost expectations, which requires a priori knowledge of network dynamics in disaster environment.

D. OPTIMIZATION PROBLEM

Given the states sequence observation $\{s^t : t \in \mathbb{N}\}$, and the stationary wireless access policy ϕ, the expected long-term cost conditioned on an initial state $\hat{s}$ can be expressed by taking the expectations with the immediate cost utility as follows.

$$V(s,\phi) = \mathbf{E}_\phi\left[(1-\gamma)\sum_{k=1}^{\infty}\gamma^{k-1}r^k|\hat{s}=s\right]. \tag{12}$$

where $s \in \mathcal{S}$ is a network state, $\gamma \in [0, 1)$ is a time discount factor, γ^{k-1} denotes the discount factor to the $(k-1)-$th power. Our goal is to find an optimal wireless access policy ϕ^* to minimize the expected long-term cost function $V(s,\phi)$ as

$$\phi^* = \underset{\phi}{\operatorname{argmin}}\, V(s,\phi),\ \ \forall s \in \mathcal{S}. \tag{13}$$

V. PROPOSED LEARNING-BASED WIRELESS ACCESS CONTROL APPROACH

A. GENERAL SOLUTION: Q-LEARNING

The formulated problem in Eqn. (13) is in general a single-agent infinite-horizon MDP with the discounted cost criterion. The solution of Eqn. (13) is equivalent to the following Bellman equation solution [33].

$$\begin{aligned} V(s^k) = \min_{\phi(s^k)}\Bigg[&(1-\gamma)r^k \\ &+\gamma\cdot\sum_{s^{k+1}\in\mathcal{S}} Pr\{s^{k+1}|s^k,\phi(s^k)\}\cdot V(s^{k+1})\Bigg]. \end{aligned} \tag{14}$$

This equation could be solved by iteration based method with the complete knowledge of s^k involvement in terms of moving location, channel state, energy queue states, and packet arrival. Since obtaining these knowledge in advance is extremely hard, if not impossible, in the post-disaster scenario, we rely on the reinforcement learning mechanism, i.e., Q-learning [14], to solve this problem without the priori knowledge of the network state transition statistics.

At each time slot k, based on the current observation of the network states s^k, the agent UE takes a combined action (T^k, E^k), and receives a corresponding reward r^{k+1} at the end of time slot k. Based the state and actions, Q-learning can build a lookup Q-table, the cell of which is denoted as Q-value $Q(s^k, (T^k, E^k))$ and represents the value of this state-action pair. The optimal action is chosen based on it at the end of time slot k, and the Q-value $Q(s^k, (T^k, E^k))$ will be updated by

$$\begin{aligned} Q(s^k,(T^k,E^k)) \leftarrow\ & Q(s^k,(T^k,E^k)) \\ & + \alpha^k \Big(r^k + \gamma \min Q(s^{k+1},(T^{k+1},E^{k+1})) \\ & - Q(s^k,(T^k,E^k))\Big), \qquad (15) \end{aligned}$$

where $\alpha^k \in [0, 1)$ is the learning rate which generally decreases with time. In this work, the reward is the cost function r^k that given by Eqn. (10). It has been proven that if the following three conditions are satisfied, the convergence of the learning process could be ensured [5]. 1) the network state transition probability under the optimal stationary control policy is stationary, 2) $\sum_{k=1}^{\infty} \alpha^k$ is infinite and $\sum_{k=1}^{\infty} (\alpha^k)^2$ is finite, and 3) all state-action pairs are visited infinitely often.

However, when the network state space and action space become huge, the reinforcement learning process suffers from extremely low convergence speed. For instance, we consider a comparatively simple UE model which the agent UE has energy storage limit $\bar{H} = 6$ and buffer size $\bar{J} = 2$. When we consider a small post-disaster scenario that the agent UE could move in a 500 m × 500 m area and the number of APs $N = 6$, there will be approximate 19×10^8 network states.

To this end, in order to find the solution of Eqn. (13) within limited time slots, we present the proposed DRL based wireless access control approach in next subsection.

B. DRL-BASED WIRELESS ACCESS CONTROL

To deal with the massive network state space $\mathcal{S}$, deep reinforcement learning [22] has been proposed recently, which utilizes a convolutional neural network, i.e., Q network, to approximate the Q-Table. The Q network learns to map state-action pairs to Q values, rather than uses a lookup table, by which significantly improves the convergence of the training process. In this work, we propose to utilize DRL technique to efficiently find the optimal wireless access policy for a mobile agent UE in the considered edge-aided disaster response networks.

DRL employs Q network as a function approximator to estimate the Q-function, such as $Q(s^k, (T^k, E^k)) \approx Q(s^k, (T^k, E^k); \theta^k)$, where θ^k denotes a weight vector associated with Q network. Similar to most of DRL based mechanism [24]–[26], [32], we assume that the agent UE is equipped with a replay memory [35]. The replay memory can store parts of the historical experience in terms of state-action pairs, reward and the state transitions. The agent UE will randomly sample a mini-batch of historical experience from the replay memory to train the Q network online. Furthermore, to make training more stable, the agent maintains a target Q network, by which it keeps a copy of the Q network and uses it in the Bellman equation.

The learning process of DRL is very similar to the Q-value updating process of Q learning given in Eqn. (15). The difference is that a weight vector θ^k in the neural network is attached to state-action pair. The Q-value update process in DRL could be expressed by

$$\begin{aligned} & Q(s^k,(T^k,E^k);\theta^k) \\ & \leftarrow Q(s^k,(T^k,E^k);\theta^k) + \alpha^k\Big[(1-\gamma)r^k \\ & \quad + \gamma Q\big(s^{k+1}, \underset{(T^{k+1},E^{k+1})}{\operatorname{argmin}} Q(s^{k+1},(T^{k+1},E^{k+1});\tilde{\theta});\theta^k\big) \\ & \quad - Q(s^k,(T^k,E^k);\theta^k)\Big]. \qquad (16) \end{aligned}$$

where $\tilde{\theta}$ is the weight of target network.

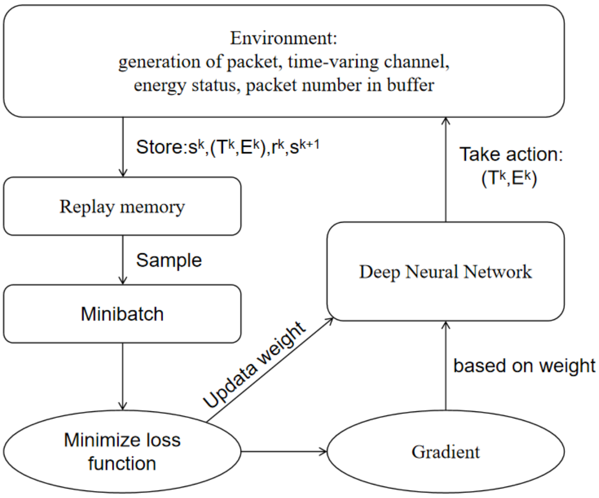


FIGURE 2. Learning process of the proposed approach.

Fig. 2 illustrates the learning process of our approach. The DRL uses a replay memory to store the historical transitions $\mathcal{M}^k = \{(s^{k-M+1}, (T^{k-M+1}, E^{k-M+1}), r^{k-M+1}, s^{k-M+2}), \cdots, (s^k, (T^k, E^k), r^k, s^{k+1})\}$, where M is the replay memory size. Then the UE randomly samples a mini-batch of transitions, i.e., $\tilde{\mathcal{M}}^k \subseteq \mathcal{M}^k$ to train θ^k in the direction of minimizing the loss function given by Eqn. (17). By differentiating the loss function with θ^k, we have the gradient shown in Eqn. (18). And we summarize the training process of agent UE in Algorithm 1.

$$\begin{aligned} & L(\theta^{k+1}) \\ & = E_{(s^k,(T^k,E^k),r^k,s^{k+1})\in\tilde{\mathcal{M}}^k}\Big[\Big((1-\gamma)r^k \\ & \quad + \gamma Q\big(s^{k+1}, \underset{(T^{k+1},E^{k+1})}{\operatorname{argmin}} Q(s^{k+1},(T^{k+1},E^{k+1});\tilde{\theta}^k);\theta^k\big) \\ & \quad - Q(s^k,(T^k,E^k);\theta^{k+1})\Big)^2\Big]. \qquad (17) \end{aligned}$$

Algorithm 1 DRL Based Wireless Access Control Approach

```
Initialize:
Initialize replay memory with size $M$
Initialize the Q network parameters with weight $\theta$
Initialize the target network parameters with weight $\tilde{\theta}$
repeat
    At the beginning of time slot $k$, observe the current state $s^k \in \mathcal{S}$
    Select an action $(T^k, E^k)$ randomly with probability $\epsilon$, or $\arg\min_{(T^k,E^k)} Q(s^k, (T^k, E^k); \theta^k)$ with probability $1-\epsilon$
    Obtain the reward $r^k$ based on Eqn. (10)
    Observe the state transitions $s^{k+1}$
    Store $\mathcal{M}^k = \{(s^k, (T^k, E^k), r^k, s^{k+1})\}$ into replay memory
    if time slot $k$ < mini-batch size then
        mini-batch transitions $\tilde{\mathcal{M}}^k = \mathcal{M}^k$
    else
        randomly samples mini-batch transitions $\tilde{\mathcal{M}}^k \subseteq \mathcal{M}^k$
    end if
    Compute the gradient by Eqn. (18) to update the weight $\theta^{k+1}$
    Every $\delta$ time slot $\tilde{\theta} = \theta$
until the loss function converges
```

$$\begin{aligned}
&\nabla_{\theta^{k+1}} L(\theta^{k+1}) \\
&\quad = E_{(s^k,(T^k,E^k),r^k,s^{k+1})\in\tilde{\mathcal{M}}^k}\bigg[\Big((1-\gamma)r^k \\
&\qquad + \gamma Q\big(s^{k+1}, \operatorname*{argmin}_{(T^{k+1},E^{k+1})} Q(s^{k+1}, (T^{k+1}, E^{k+1}); \tilde{\theta}^k); \theta^k\big) \\
&\qquad - Q(s^k, (T^k, E^k); \theta^{k+1})\Big)\nabla_{\theta^{k+1}} Q(s^k, (T^k, E^k); \theta^{k+1})\bigg].
\end{aligned} \tag{18}$$

VI. EXPERIMENTS

In this section, we demonstrate extensive simulation results based on TensorFlow [11] to show the performance of the proposed learning-based approach, and validate its superiority by comparing it with other baseline schemes.

A. GENERAL SETUPS

As illustrated in Fig. 1, we consider an edge-aided disaster response network with a $500m \times 500m$ area. We assume that there are a single MDRU acts as the gateway and edge unit, and 6 relay units to expand the network coverage. The locations of them are fixed, and the coordinates of MDRU and relay units are $(0, 0)$ and $(200, 0)$, $(0, 200)$, $(200, 200)$, $(300, 100)$, $(300, 300)$, $(100, 300)$, respectively.

The agent UE moves based on Manhattan grid mobility model with a constant speed of 10m/s. The variance of channel fading σ_n^2 is set to 6dB [34] and the generated channels gains are quantized to 32 possible values finally. We consider two representative post-disaster scenarios, i.e., uniform distributed scenario and clumped distribution scenario. Specifically, in uniform distributed scenario, UEs are uniformly distributed in the post-disaster area, and in clumped distributed scenario, UEs forms clusters and are concentrated in some particular places such as shelters. Accordingly, the average numbers of arrived packets earlier than the agent UE's, i.e., τ_n, at relay unit n are set to $\{1.25, 0.75, 1, 0.95, 0.95, 0.85\}$ and $\{1.25, 0.75, 1, 10, 10, 0.85\}$ respectively. And the average number of arrived packets at MDRU is fixed at 2. For the setting of the proposed DRL-based approach, we set the replay memory size as $M = 10000$ and mini-batch size as 512.

We consider two types of UE models that may depend on different applications. One is the UE without buffer, i.e., $\bar{J} = \{0\}$, which is applied to real time or delay sensitive applications. Another is the UE with buffer, i.e., $\bar{J} = \{1, 2\}$, which is applied to delay tolerant or deliver rate first applications. The weight factors in the cost function Eqn. (10) are set as $\rho = 1$ and $\xi = 0$ when $\bar{J} = 0$, and $\rho = 5$ and $\xi = 0.6$ otherwise. Other main simulation parameters are summarized in Table 2.

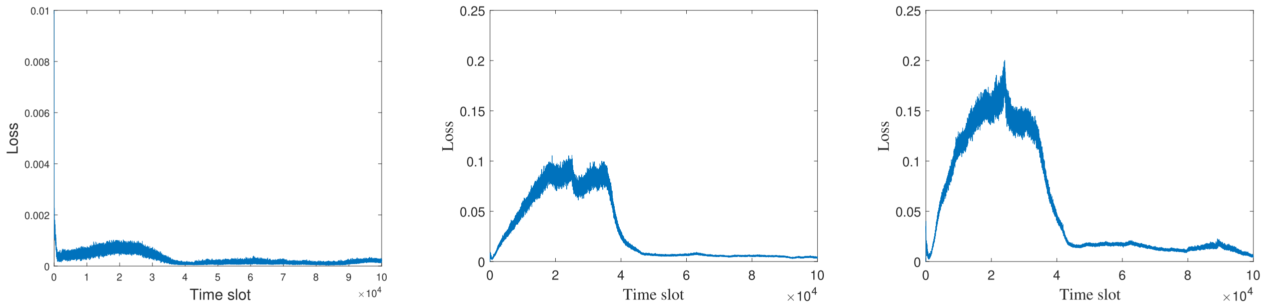
TABLE 2. Simulation parameters.

Parameter	Value	Unit
Packet size	3×10^5	bit
Maximum energy	6	unit
Energy unit	5×10^{-3}	J
Initial energy	2	unit
Bandwidth	10^6	Hz
Background noise	10^{-11}	W
Total time slots	10×10^4	
Time slot duration	1	s

B. SIMULATION RESULTS

1) CONVERGENCE WITH DIFFERENT BUFFER SIZES

Firstly, we validate the convergence of the proposed DRL-based approach by showing the results of loss function in Eqn. (17) varying with time. Here we show a result with the average packet generation rate $\lambda_{(p)} = 0.3$ and the average energy harvesting rate $\lambda_{(e)} = 0.4$, and we confirm that the convergence results with other $\lambda_{(p)}$ and $\lambda_{(e)}$ are similar. In this result, the neural network has 1 hidden layer with 512 neurons. According to the result illustrated in Fig. 3, the convergence of the proposed DRL-based approach could be confirmed from approximately 40000 time slot. And as shown in Figs. 3(b) and 3(c), for the UE models with buffer, there is a significant rise before the convergence for the loss value. Since the UE has to learn whether to transmit the packet immediately or carry it to find a better opportunity. As long as the UE continues moving inside this area, an optimal and stable performance could be achieved after the convergence. Regarding the cases that the UE moves outside the area or the locations of MDRU and relay changes, a relearning process is required. Notice that all the results in the remainder are obtained after the loss function converges, i.e., the last 1×10^4 time slots.



(a) $\bar{J} = 0$ (b) $\bar{J} = 1$ (c) $\bar{J} = 2$

FIGURE 3. The loss function with $\bar{J} = \{0, 1, 2\}$.

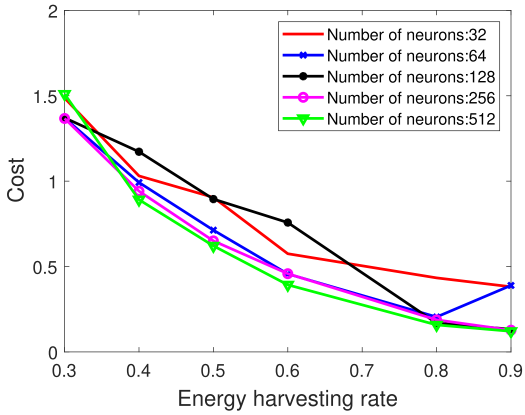


FIGURE 4. Cost with different number of neurons.

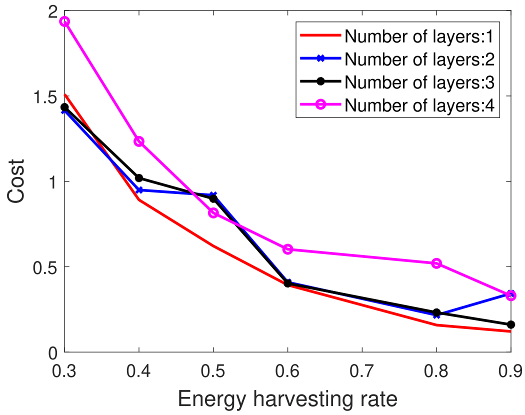


FIGURE 5. Cost with different number of layers.

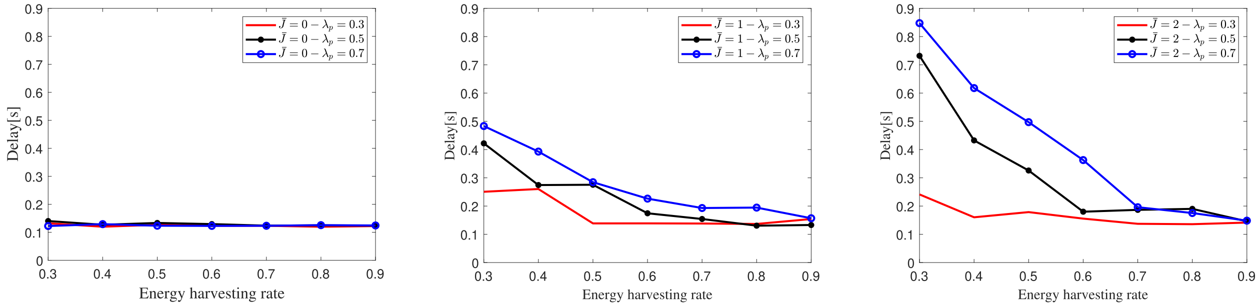
2) PERFORMANCE UNDER VARIOUS NEURAL NETWORK STRUCTURES

Next, we demonstrate the impacts of neural network structure on the proposed DRL-based approach with buffer size $\bar{J} = 2$. Specifically, we show the variance of cost given in Eqn. (10) with different average energy harvesting rate. For the result illustrated in Fig. 4, we fix the number of hidden layer at 1 and change the number of neurons from 32 to 512. And for the result illustrated in Fig. 5, we change the number of hidden layers from 1 to 4, and keep the number of neurons for each layer at 512. First of all, as expected, we can confirm that the costs decrease as the energy harvesting rate increases, since the cost value represents a combination of delay and packet drop ratio. Next, we find that the setting of 1 hidden layer with 512 neurons results in the lowest and stablest cost, which is used in the following simulations. The reason is that over a limited time horizon, adding more hidden layers to the DQN may lead to higher training errors [36], and a wider (not deeper) DQN can better approximate the Q-function in our considered scenario.

3) IMPACTS OF BUFFER SIZE UNDER DIFFERENT $\lambda_{(p)}$ AND $\lambda_{(e)}$

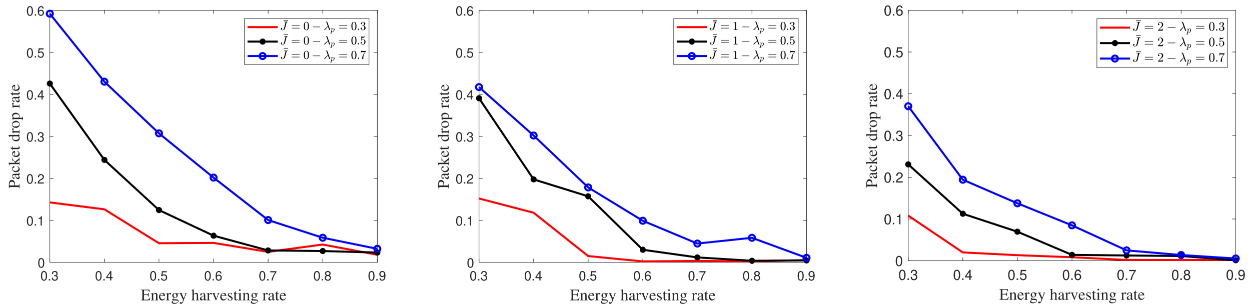
For the proposed approach with different buffer size, we evaluate its average delay and packet drop rate, at different packet generation rates $\lambda_{(p)}$ and energy harvesting rates $\lambda_{(e)}$. The average delay consists of the transmission delay and the waiting time in buffer. We plot the results varying with the energy harvesting rate $\lambda_{(e)}$, at packet generation rates $\lambda_{(p)} = \{0.3, 0.5, 0.7\}$. In Fig. 6, we illustrate the average delay for different buffer sizes. First of all, from Fig. 6(a), if there is no buffer at UE, we can observe that packet generation rate does not affect the delay, and high energy harvesting rate only improve the delay very slightly. This is because the delay in this no buffer UE case can only be affected by the routing selection, since the UE cannot choose the timing for the transmission. For the cases of UE with buffer that shown in Figs. 6(b) and 6(c), it is obvious that the delay reduces when the energy harvesting rate increases. And as expected, higher packet generation rate leads to higher delay. And generally, larger buffer size results in larger delay, since the probability of storing the packet and transmitting it later increases. But when the energy harvesting rate increases, i.e., higher than 0.7, there is no much difference between $\bar{J} = 1$ and $\bar{J} = 2$. This is because when there is enough energy, the agent UE prefers to transmit the packet directly rather than to store it into the buffer.

In Fig. 7, we illustrate the packet drop rate for different buffer sizes. Regardless of different buffer sizes, the packet drop rate reduces when the energy harvesting rate increases.



(a) $\bar{J}=0$ (b) $\bar{J}=1$ (c) $\bar{J}=2$

FIGURE 6. The average delay for different packet generation rates with $\bar{J}=\{0,1,2\}$.



(a) $\bar{J}=0$ (b) $\bar{J}=1$ (c) $\bar{J}=2$

FIGURE 7. The average packet drop rate for different packet generation rates with $\bar{J}=\{0,1,2\}$.

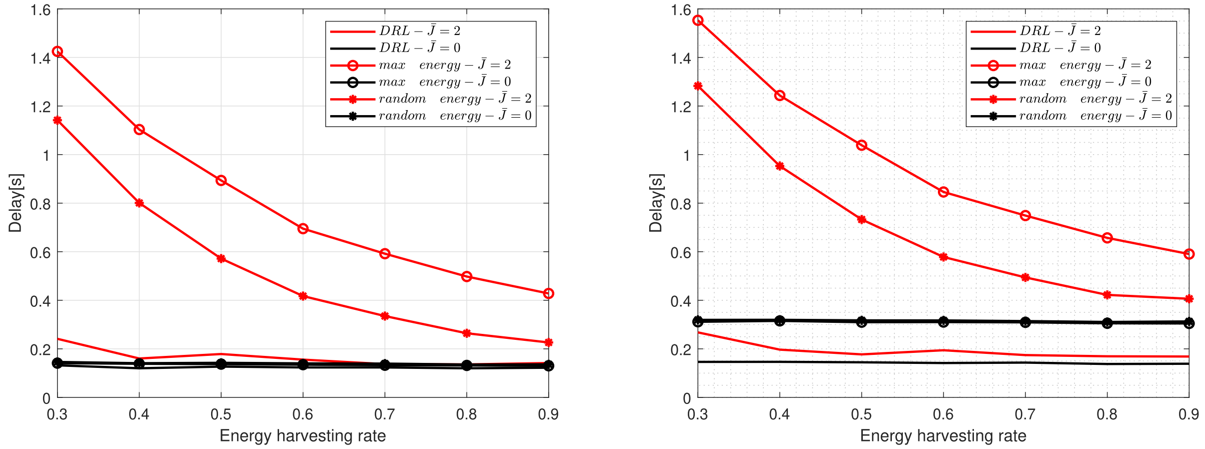
For the case of UE without buffer, there is a significant change in packet drop rate that reaches to approximately $10\sim50\%$, when the packet generation rate varies. And by applying buffer in the agent UE, we can confirm that the larger buffer size leads to lower packet drop rate, which is as expected. Specifically, when $\lambda_{(e)}$ is low and $\lambda_{(p)}$ is high, a great number of packets still have to be dropped even with buffer. However, as the energy harvesting rate increases and when the packet generation rate is not so high, applying the buffer in UE has obvious advantages in reducing the packet drop rate.

4) COMPARISON WITH BASELINE SCHEMES IN DIFFERENT ENVIRONMENT

In this work, we considered a combined wireless access control problem consisting of transmitting time control, routing selection and energy allocation. No existing work can be applied directly to solve this issue. Therefore, in this subsection, we compare the performance of the proposed DRL based approach with two baseline schemes at packet generation rate $\lambda_{(p)}=0.7$. Both the uniform distributed and clumped distributed post-disaster scenarios are considered, and to represent different types of UEs, buffer size $\bar{J}=\{0,2\}$ are applied. We consider two baseline schemes, i.e., random energy and max energy schemes. In random energy scheme, UE sends the packet by randomly allocating the transmitting energy, and in the max energy scheme, UE always uses the maximal energy in its energy storage to transmit. And in both schemes, UE sends the packet to the nearest MDRU or relay unit.

First, in Fig. 8 we compare the average delay of the proposed approach with the baseline schemes in two typical post-disaster scenarios. From Fig. 8(a), in the uniform distribution scenario, we can observe that the proposed approach achieves the similar delay when $\bar{J}=\{0\}$. However, when $\bar{J}=\{2\}$, the proposed approach can significantly reduce the delay compared with two baseline schemes, especially when the energy harvesting rate is low. The reason is that the proposed approach could learn an optimal transmitting energy allocation policy to significantly reduce the time of the packet waiting in the buffer. Regarding the clumped distribution scenario, as shown in Fig. 8(b), the proposed approach could substantially reduce the delay for both $\bar{J}=\{0\}$ and $\bar{J}=\{2\}$ cases. The reason of the improvement on the no-buffer case is that the proposed approach can find the optimal route to forward the packet which avoids the relay unit with heavy traffic volumes.

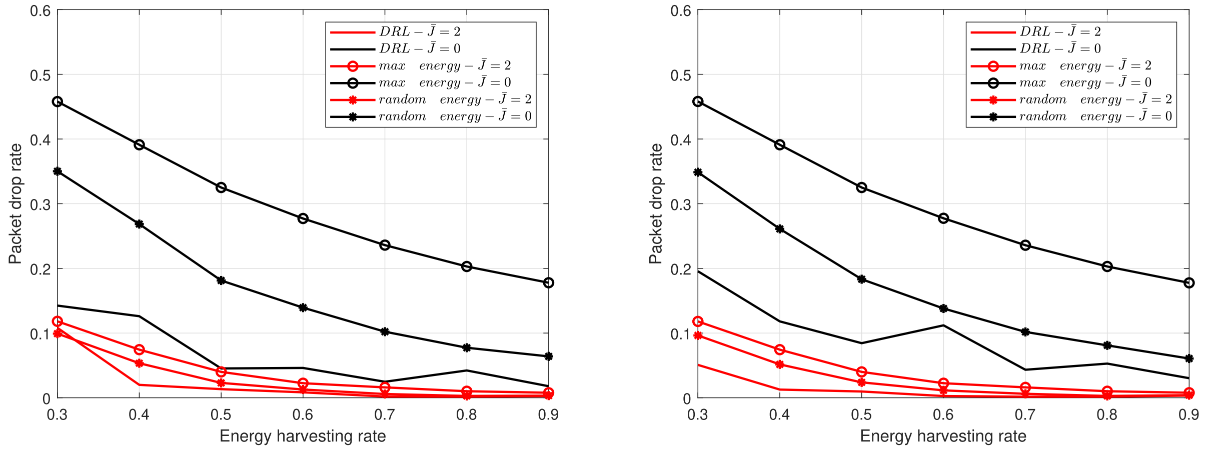
Then, we show the comparison of packet drop rate varying with the energy harvesting rate $\lambda_{(e)}$ under two post-disaster scenarios in Fig. 9. Firstly, we can observe that contrary to the result of delay shown in Fig. 8, applying the buffer at UE could significantly reduce the packet drop rate for all three schemes, since the buffer can store the packet when there is no energy is harvested. In the uniform distribution scenario, as illustrated in Fig. 9(a), the proposed approach



(a) Uniform distribution scenario

(b) Clumped distribution scenario

FIGURE 8. Comparisons on average delay.



(a) Uniform distribution scenario

(b) Clumped distribution scenario

FIGURE 9. Comparisons on packet drop rate.

could significantly reduce the packet drop rate for both $\bar{J} = \{0, 2\}$ cases. For instance, when the energy harvesting rate is higher than 0.5, the proposed approach could keep the packet drop rate lower than 5% and 1% for no-buffer and with-buffer cases, respectively. Regarding the results in clumped distributed scenario shown in 9(b), the similar improvement of the proposed approach could be confirmed. But the performance of the proposed approach degrades when $\bar{J} = \{0\}$ compared with that in uniform distribution scenario, since forwarding the packet to the relay unit with low traffic volume but far away may result in the packet loss due to the poor channel conditions.

For the conclusion, except the delay is almost same for the no-buffer case in uniform-distribution scenario, the proposed approach outperforms the baseline schemes in all other scenarios. By interacting from the surrounding wireless environment, the agent UE could learn the optimal wireless access control policy to reduce the delay and packet drop rate and adapt to different post-disaster scenarios.

VII. CONCLUSION AND FUTURE WORK

In this paper, we investigated the wireless access control problem in an edge-aided disaster response network, and proposed a DRL-based approach to control the transmit timing, routing and energy allocation without the knowledge of the network statistics in priori. The proposed approach takes into account the network dynamics of UE mobility, channel states, energy harvesting and packet generation. By interacting with the unknown post-disaster wireless environment, the proposed approach learns the optimal wireless access control policy to minimize both the delay and packet drop rate. The performance of the proposed approach was validated by comparing with baseline schemes in two typical post-disaster environments. It was confirmed that the proposed approach outperforms the baseline schemes significantly in terms of delay and packet drop rate.

For the future work, we consider to improve the proposed scheme in the following two aspects. First, the current proposed scheme only considers the local optimal wireless access control. In the future, we intend to propose a global optimal wireless access control scheme for multiple UEs by applying federated learning technique. Second, we assume an orthogonal channel access model in this work. In the future, we intend to consider more complex and realistic network model, in which the energy control, channel assignment, routing selection could be optimized together.

REFERENCES

[1] K. Eshghi and R. C. Larson, ''Disasters: Lessons from the past 105 years,'' *Disaster Prevention Manage., Int. J.*, vol. 17, no. 1, pp. 62–82, Feb. 2008.

[2] H. Nishiyama, K. Suto, and H. Kuribayashi, ''Cyber physical systems for intelligent disaster response networks: Conceptual proposal and field experiment,'' *IEEE Netw.*, vol. 31, no. 4, pp. 120–128, Jul./Aug. 2017.

[3] X. Wang, F. Jiang, L. Zhong, Y. Ji, S. Yamada, K. Takano, and G. Xue, ''Intelligent post-disaster networking by exploiting crowd big data,'' *IEEE Netw.*, vol. 34, no. 4, pp. 49–55, Jul. 2020.

[4] C. B. Nelson, B. D. Steckler, and J. A. Stamberger, ''The evolution of hastily formed networks for disaster response: Technologies, case studies, and future trends,'' in *Proc. IEEE Global Humanitarian Technol. Conf.*, Seattle, WA, USA, Oct. 2011, pp. 467–475, doi: 10.1109/GHTC.2011.98.

[5] T. Sakano, Z. M. Fadlullah, T. Ngo, H. Nishiyama, M. Nakazawa, F. Adachi, N. Kato, A. Takahara, T. Kumagai, H. Kasahara, and S. Kurihara, ''Disaster-resilient networking: A new vision based on movable and deployable resource units,'' *IEEE Netw.*, vol. 27, no. 4, pp. 40–46, Aug. 2013.

[6] T. Sakano, S. Kotabe, T. Komukai, T. Kumagai, Y. Shimizu, A. Takahara, T. Ngo, Z. M. Fadlullah, H. Nishiyama, and N. Kato, ''Bringing movable and deployable networks to disaster areas: Development and field test of MDRU,'' *IEEE Netw.*, vol. 30, no. 1, pp. 86–91, Jan./Feb. 2016.

[7] Q. T. Minh, K. Nguyen, C. Borcea, and S. Yamada, ''On-the-fly establishment of multihop wireless access networks for disaster recovery,'' *IEEE Commun. Mag.*, vol. 52, no. 10, pp. 60–66, Oct. 2014.

[8] X. Wang, H. Zhou, L. Zhong, Y. Ji, K. Takano, S. Yamada, and G. Xue, ''Capacity-aware cost-efficient network reconstruction for post-disaster scenario,'' in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.

[9] L. Zhong, Y. Ji, X. Wang, S. Yamada, K. Takano, and G. Xue, ''Population-aware relay placement for wireless multi-hop based network disaster recovery,'' in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–6.

[10] H. Nishiyama, M. Ito, and N. Kato, ''Relay-by-smartphone: Realizing multihop device-to-device communications,'' *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 56–65, Apr. 2014.

[11] N. Chaudhuri and I. Bose, ''Application of image analytics for disaster response in smart cities,'' in *Proc. 52nd Hawaii Int. Conf. Syst. Sci. (HICSS)*, 2019, pp. 1–10.

[12] J. Yang and S. Ulukus, ''Optimal task scheduling in an energy harvesting communication system,'' *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, Jan. 2012.

[13] H. Zhou, Y. Ji, X. Wang, and S. Yamada, ''eICIC configuration algorithm with service scalability in heterogeneous cellular networks,'' *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 520–535, Feb. 2017.

[14] C. J. C. H. Watkins and P. Dayan, ''Q-learning,'' *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992.

[15] H. Kwon and S.-H. Kim, ''Improving mobile device classification using security events for preventing wireless intrusion,'' *Int. J. Secur. Appl.*, vol. 7, no. 6, pp. 181–190, 2013, doi: 10.14257/ijsia.2013.7.6.19.

[16] G. Tuysuzoglu, D. Birant, and A. Pala, ''Majority voting based multi-task clustering of air quality monitoring network in Turkey,'' *Appl. Sci.*, vol. 9, no. 8, p. 1610, 2019.

[17] J. Xu, L. Chen, and S. Ren, ''Online learning for offloading and autoscaling in energy harvesting mobile edge computing,'' *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 3, pp. 361–373, Sep. 2017.

[18] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, ''Reinforcement learning for energy harvesting point-to-point communications,'' in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[19] K.-L.-A. Yau, P. Komisarczuk, and P. D. Teal, ''Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues,'' *J. Netw. Comput. Appl.*, vol. 35, no. 1, pp. 253–267, Jan. 2012.

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

[21] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. M. Hirsch, Ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2010.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, ''Human-level control through deep reinforcement learning,'' *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[23] H. van Hasselt, A. Guez, and D. Silver, ''Deep reinforcement learning with double Q-learning,'' in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 2094–2100.

[24] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis, ''Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning,'' *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4005–4018, Jun. 2019.

[25] Y. Cao, S.-Y. Lien, and Y.-C. Liang, ''Deep reinforcement learning for multi-user access control in non-terrestrial networks,'' *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1605–1619, Mar. 2021.

[26] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, ''Dynamic reservation and deep reinforcement learning based autonomous resource slicing for virtualized radio access networks,'' *IEEE Access*, vol. 7, pp. 45758–45772, 2019.

[27] X. Chen, C. Wu, T. Chen, H. Zhang, Z. Liu, Y. Zhang, and M. Bennis, ''Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective,'' *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2268–2281, Apr. 2020.

[28] R. Yin, Z. Wu, S. Liu, C. Wu, J. Yuan, and X. Chen, ''Decentralized radio resource adaptation in D2D-U networks,'' *IEEE Internet Things J.*, early access, Aug. 12, 2020, doi: 10.1109/JIOT.2020.3016019.

[29] H. Zhang, D. Zhan, C. J. Zhang, K. Wu, Y. Liu, and S. Luo, ''Deep Reinforcement Learning-Based Access Control for Buffer-Aided Relaying Systems With Energy Harvesting,'' *IEEE Access*, vol. 8, pp. 145006–145017, Aug. 2020.

[30] Y. S. Nasir and D. Guo, ''Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks,'' *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[31] R. Ding, F. Gao, and X. S. Shen, ''3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach,'' *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796–7809, Dec. 2020, doi: 10.1109/TWC.2020.3016024.

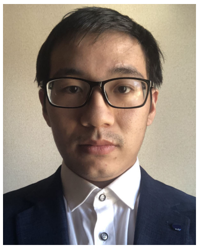
[32] H. Zhou, X. Wang, M. Umehira, X. Chen, C. Wu, and Y. Ji, ''Deep reinforcement learning based access control for disaster response networks,'' in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[33] R. Bellman, *Dynamic Programming*. New York, NY, USA: Dover, 2003.

[34] M. Viswanathan, *Wireless Communication Systems in MATLAB*, 2nd ed. Jun. 2020.

[35] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, ''Prioritized experience replay,'' 2015, *arXiv:1511.05952*. [Online]. Available: http://arxiv.org/abs/1511.05952

[36] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.



HANG ZHOU received the B.E. degree from the Wuhan University of Science and Technology, China, in 2017. He is currently pursuing the master's degree with Ibaraki University, Japan. His research interest includes machine learning based wireless networking.

XIAOYAN WANG (Senior Member, IEEE) received the B.E. degree from Beihang University, China, and the M.E. and Ph.D. degrees from the University of Tsukuba, Japan. From 2013 to 2016, he worked as an Assistant Professor (by special appointment) with the National Institute of Informatics (NII), Japan. He is currently working as an Associate Professor with the Graduate School of Science and Engineering, Ibaraki University, Japan. His research interests include intelligent networking, wireless communications, cloud computing, big data systems, and security and privacy.

MASAHIRO UMEHIRA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1978, 1980, and 2000, respectively. Since 1980, he has been with Nippon Telegraph and Telephone Corporation (NTT), where he has been involving in the research and development of modem and TDMA equipment for satellite communications, TDMA satellite communication systems, broadband wireless access systems for mobile multimedia services, and ubiquitous wireless systems. From 1987 to 1988, he was with the Communications Research Center Canada, Department of Communications, Canada, as a Visiting Scientist. Since 2006, he has also been a Professor with Ibaraki University, Japan. His research interests include broadband wireless access technologies, wireless networking, cognitive radio, future satellite communication systems, and wireless-based ubiquitous systems. He received the Young Engineer Award and the Achievement Award from IEICE, in 1987 and 1999, respectively. He also received the Education, Culture, Sports, Science and Technology Minister Award, in 2001, and the TELECOM System Technology Award from the Telecommunications Advancement Foundation, in 2003.

XIANFU CHEN (Member, IEEE) received the Ph.D. degree (Hons.) in signal and information processing from the Department of Information Science and Electronic Engineering (ISEE), Zhejiang University, Hangzhou, China, in March 2012. Since April 2012, he has been with the VTT Technical Research Centre of Finland, Oulu, Finland, where he is currently a Senior Scientist. His research interests include various aspects of wireless communications and networking, with emphasis on human-level and artificial intelligence for resource awareness in next-generation communication networks. He received the Exemplary Reviewer Certificate of IEEE TRANSACTIONS ON COMMUNICATIONS, in 2021. He was a recipient of the Best Paper awards from the 5th IEEE International Conference on Cloud and Big Data Computing (CBDCom 2019) and the 10th EAI International Conference on Mobile Networks and Management (EAI MONAMI 2020). He has served and also serving as a Track Co-Chair and a TPC Member for a number of IEEE ComSoc flagship conferences. He is the Vice Chair of the IEEE Special Interest Group on Big Data with Computational Intelligence and the IEEE Special Interest Group on AI Empowered Internet of Vehicles. He has served as a Guest Editor for several international journals, including *IEEE Wireless Communications* Magazine, and an Editorial Board Member for the First Editorial Board of *Journal of Communications and Information Networks*. He serves as an Editor for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, *Wireless Communications and Mobile Computing*, and *China Communications*.

CELIMUGE WU (Senior Member, IEEE) received the M.E. degree from the Beijing Institute of Technology, China, in 2006, and the Ph.D. degree from The University of Electro-Communications, Japan, in 2010. He is currently an Associate Professor with the Graduate School of Informatics and Engineering, The University of Electro-Communications. His current research interests include vehicular networks, intelligent transport systems, the IoT, and mobile edge computing. He is also serving as an Associate Editor for IEEE ACCESS, *IEICE Transactions on Communications*, *International Journal of Distributed Sensor Networks*, and *Sensors* (MDPI).

YUSHENG JI (Senior Member, IEEE) is currently a Professor with the National Institute of Informatics (NII), Japan, and the Graduate University for Advanced Studies, (SOKENDAI). Her research interests include network architecture, resource management, and quality of service provisioning in wired and wireless communication networks. She is/has been an Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, an Associate Editor of IEICE transactions and IPSJ journal, a Guest Editor-in-Chief, a Guest Editor, and a Guest Associate Editor of Special Issues of IEICE transactions and IPSJ journal.