Cooperative Communications with Relay Selection based on Deep Reinforcement Learning in Wireless Sensor Networks

Yuhan Su, Xiaozhen Lu, Yifeng Zhao, Lianfen Huang, and Xiaojiang Du, *Senior Member, IEEE*

***Abstract*—Cooperative communication technology has become a research hotspot in wireless sensor networks (WSNs) in recent years, and will become one of the key technologies for improving spectrum utilization in wireless communication systems in the future. It leverages cooperation among multiple relay nodes in the wireless network to realize path transmission sharing, thereby improving the system throughput. In this paper, we model the process of cooperative communications with relay selection in WSNs as a Markov decision process and propose DQ-RSS, a deep-reinforcement-learning-based relay selection scheme, in WSNs. In DQ-RSS, a deep-Q-network (DQN) is trained according to the outage probability and mutual information, and the optimal relay is selected from a plurality of relay nodes without the need for a network model or prior data. More specifically, we use DQN to process high-dimensional state spaces and accelerate the learning rate. We compare DQ-RSS with the Q-learning-based relay selection scheme and evaluate the network performance on the basis of three aspects: outage probability, system capacity, and energy consumption. Simulation results indicate that DQ-RSS can achieve better performance on these elements and save the convergence time compared with existing schemes.**

***Index Terms*—Wireless sensor networks, cooperative communications, relay selection; outage probability, deep reinforcement learning.**

I. Introduction

WIRELESS sensor networks (WSNs) have been widely applied in different fields, such as military, political, and medical settings [1]. However, in some applications, due to the large number and small size of sensor nodes deployed in some applications, the system throughput of nodes is insufficient, and energy is scarce. Furthermore, a poor network environment renders node replacement challenging. Therefore, the issue of how to improve and prolong the network life cycle is a pertinent issue in current WSNs. Cooperative communication technology is considered a key technology that can improve sensor network performance [2], which utilizes the broadcast characteristics of wireless systems, can increase the system's data rate and expand node coverage by deploying relay nodes in WSNs. Relay technology has been used frequently in modern wireless communications and adopted in LTE Release 10 [3]. The relay mode is mainly divided into amplify-and-forward (AF) and decode-and-forward (DF) parts according to the manner in which the relay node processes the received signal [4]. In the AF mode, the relay node amplifies the received signal directly and forwards it to the user; in the DF mode, the relay node decodes the received signal and forwards it to the user after re-encoding. The AF mode is simple to implement but carries a disadvantage related to noise transmission; signal amplification leads to noise amplification. The DF mode has no such drawback due to decoding operations at the relay node. To enable accurate decoding, however, the source node and relay node each require a better channel environment. In response to the above problems, various cooperation protocols with multiple relays have been proposed in recent years [5].

For multi-relay scenarios, the most intuitive collaboration strategy is for each relay to coordinate data transmission between the source node and destination node. [6] analyzed the bit error rate of this cooperative strategy and found that the strategy can obtain the full set gain (i.e., the number of relays plus the first-order diversity gain). However, to avoid mutual interference between relay nodes, orthogonal channels (orthogonal frequency bands or time slots) must be allocated for all relay nodes.

Forwarding information consumes extensive time or frequency resources, greatly reducing the spectrum efficiency of relay cooperation. To address this problem, in [7], the authors proposed the idea of "opportunistic relay", which forwards source node information by selecting the "best relay" selected according to the channel state. The results of [7] show that the cooperative strategy based on the "best relay" can also obtain the full set gain and only requires two orthogonal channels, compared with the strategy in which all relays participate in cooperation. The cooperative strategy greatly reduces system overhead and improves spectrum efficiency.

In a scenario such as a WSN or ad-hoc network, the system overhead involved in acquiring global channel state information (CSI) is large due to the large number of network nodes. A network node can only obtain the CSI of the local one-hop link via channel feedback between nodes. Therefore, the opportunistic relay cooperation strategy based on "best relay" is difficult to implement in such wireless network scenarios. Furthermore, because network state information in dynamic WSNs is inherently inaccurate and subject to change, using CSI as the sole relay selection criterion is not sufficient for dynamic WSNs. The working state of relay nodes and the system state at the current time exert significant impact on the quality of cooperative communications; therefore, these factors should be considered in the relay selection problem. Some

This work was supported in part by the 2018 National Natural Science Foundation of China under Grant 61871339, and in part by the Key Laboratory of Digital Fujian on IoT Communication, Architecture and Safety Technology under Grant 2010499. *(Corresponding author: Yifeng Zhao.)*

researches have applied Q-learning to solve the relay selection problem, e.g., [8] proposed a Q-learning-based relay selection algorithm for wireless cooperative networks, named QL-RSA. The Q-learning algorithm is a model-free reinforcement learning algorithm, modeled by the Markov decision process (MDP), which uses an iterative approach to approximate the optimal solution. A source node with learning capabilities can determine the optimal relay to participate in collaborative communication based on the observed state of previous system performance and the quality function describing the reward.

However, the Q-learning-based relay selection scheme uses a Q-table to store the Q-value and can only deal with the problem within a small state space. In the context of a large state space, because the storage capacity of a Q-table is limited, it can not cover the complete state space; thus, the traditional reinforcement learning algorithm cannot deal with such problems. To rectify this problem, Google DeepMind introduced deep learning in reinforcement learning [9], [10] and proposed Deep-Q-Net (DQN). This approach utilizes the perceptual and acquisition ability of deep learning to enable the reinforcement learning algorithm to extract environmental features and address the pitfall of the Q-table being unable to cover the entire state space.

In this paper, we propose a DQN-based relay selection scheme in WSNs, named DQ-RSS. Our scheme combines deep learning with Q-learning to accelerate learning by selecting the optimal relay among a plurality of relay candidates according to outage probability and channel information. The main contributions of this paper are as follows:

- We investigate the use of cooperative communications for WSNs, and the MDP model is developed to describe the relay selection problem.
- We propose DQ-RSS, a deep-Q-learning-based relay selection scheme for WSNs. In DQ-RSS, an optimal relay is selected from multiple relay candidates according to outage likelihood and channel efficiency.
- We introduce a Q-learning-based relay selection scheme to our system model, named Q-RSS, and then compare it with DQ-RSS. The results show that DQ-RSS achieves a lower outage probability in the system and higher utility compared with Q-RSS. In addition, the results also show that DQ-RSS can save the convergence time compared with Q-RSS, suitable for industrial implementation.

The remainder of this paper is organized as follows. We review related work in Section II. The cooperative communications system model and MDP problem are respectively presented in Sections III and IV. A DQN-based cooperative communications scheme for WSNs is designed in Section V. We provide simulation results in Section VI and conclude in Section VII.

II. Related Work

Many studies have examined the relay selection problem. [6] proposed a relay-selection-based collaboration protocol wherein the protocol indicates that the optimal relay is the node with the largest instantaneous scaling harmonic averaging function of its source and relay destination channel gains. The authors in [7] proposed an optimal relay selection scheme based on local measurements of the instantaneous channel conditions, and [11] further confirmed the outage-optimality of this scheme. A dual-relay cooperation scheme based on the average link signal-to-noise ratio (SNR) was proposed in [12], which is suitable for wireless ad-hoc networks.

Extensive research has also investigated cooperation or competition scenarios in WSNs. In [13], a relay selection scheme based on feedback and adaptive forwarding in WSNs was proposed. The authors in [14] put forth a novel cooperative communication scheme, energy-efficient cooperative communication, to improve data transmission performance for WSNs. In this scheme, a cooperative reply is performed at each hop by the best-suited node selected from those that have successfully overheard the transmitted packet. Outage probability was used to evaluate system performance in [15], and the authors presented a relay selection scheme with optimal power allocation in WSNs.

In addition, machine learning provides a highly appealing mathematical tool for designing algorithms for cooperative communications. [16] proposed a relay selection scheme based on multi-agent reinforcement learning for resource-constrained WSNs, which can adaptively select an optimal relay according to the quality of service of cooperative communication. [17] proposed a distributed-Q-Learning-based medium access control protocol in WSNs, which enables the system can schedule the radio resources adaptively according to the network traffic load. [18] used unmanned aerial vehicles to relay the message of an onboard unit and improved the bit error rate of vehicular ad-hoc networks (VANETs) against smart jammers; a form of reinforcement learning, called hotbooting policy hill climbing, was also employed to help the VANET resist jamming under an unknown VANET model and jamming model. The authors in [8] used a reinforcement learning technique to search the optimal relay; they proposed a Q-learning-based relay selection scheme for wireless networks, which maximized total-network capacity. However, neither the [16] nor [8] considered the convergence time and computational complexity of the proposed scheme, and they also did not exploit other machine-learning-based schemes to compare with the reinforcement-learning-based scheme. Our work is based on the adopted scheme in [8] and improve the convergence speed with DQN. Moreover, we also compare the other performance of our proposed scheme with the scheme proposed in [8].

III. System Model

A. Network Model

As shown in Fig. 1, we consider a WSN system consisting of a source node S, a destination node D, and M half-duplex relays (R_m for $m = 1, ..., M$). A direct link exists between nodes S and D, and relays are located between them.

Let $\mathbf{d}^{(k)} = [d_1^{(k)}, d_2^{(k)}, d_3^{(k)}]$ denote the topology vector of the network at time slot k, where $d_1^{(k)}$ is the distance between node S and relay R_m; $d_2^{(k)}$ is the distance between relay R_m and node D; and $d_3^{(k)}$ corresponds to the distance in the $S-D$ link. The distances $d_2^{(k)}$ and $d_3^{(k)}$ change as node D moves.

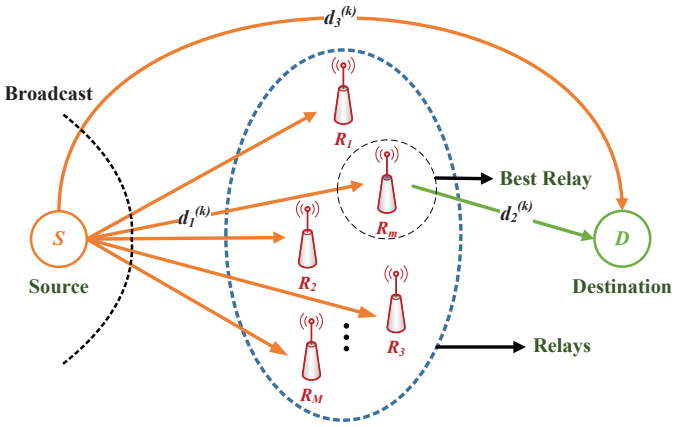


Fig. 1: Illustrative of the relay selection scheme in WSNs.

Node S broadcasts signals to node D and relays M with a fixed transmit power P_S, and M relays amplify their signals and forwards them to the node D with a fixed transmit power P_R. Each relay node normalizes the received signal from node S and forwards it to node D.

B. Channel Model

We denote the system channel gain vector by $\mathbf{g}^{(k)} = [g_1^{(k)}, g_2^{(k)}, g_3^{(k)}]$ where $g_1^{(k)}$, $g_2^{(k)}$, and $g_3^{(k)}$ are the channel gain of the $S - R_m$ link, $R_m - D$ link, and $S - D$ link at time slot k, respectively. Similar to [19], the channel gain is modeled as

$$g_i^{(k)} = \frac{|h_i^{(k)}|^2}{{d_i^{(k)}}^{-\gamma}}, i = 1, 2, 3. \quad (1)$$

where $\mathbf{h}^{(k)} = [h_1^{(k)}, h_2^{(k)}, h_3^{(k)}]$ consists of channel coefficients of the $S - R_m$ link $h_1^{(k)}$, the $R_m - D$ link $h_2^{(k)}$, the $S - D$ link $h_3^{(k)}$ at time slot k, respectively. Channel coefficients include the effect of path-loss, fading, and shadowing. The path-loss exponent γ is set according to [20]. System channels are modeled as complex additive white Gaussian noise with zero mean and variance σ^2. The SNR of signals received by the node D sent from the node S and relay R_m at time slot k are denoted by $\rho_1^{(k)}$ and $\rho_2^{(k)}$, respectively.

$$\rho_1^{(k)} = \frac{P_S g_3^{(k)}}{\sigma^2}; \quad (2)$$

$$\rho_2^{(k)} = \frac{P_R g_2^{(k)}}{\sigma^2}. \quad (3)$$

C. Outage Analysis

As noted in the previous section, if the opportunity relay selection policy in [6] is used, it is sufficient to select only one relay for $R_m - D$ transmission from the candidate relays. Therefore, we use an optimal relay for transmission, as in (5) and (6). In this case, the entire communication process includes two time slots: the first time slot node S broadcasts information to all nodes while relay nodes and node D receive information simultaneously; the selected relay node R_m forwards information in the second time slot, and node D receives and processes the information in the two time slots via maximum ratio combining.

The maximum mutual information, in time slot k, between nodes S and D for direct transmission is given by

$$I_{S,D}{}^{(k)} = \log_2(1 + \frac{P_S g_3^{(k)}}{\sigma^2}). \quad (4)$$

The mutual information between node S and relay R_m at time slot k can be written as

$$I_{S,R_m}{}^{(k)} = \frac{1}{2}\log_2(1 + \frac{P_S g_1^{(k)}}{\sigma^2}). \quad (5)$$

Without a loss of generality, we assume all noise variances are identical and all relays use the AF protocol to participate in cooperation in this model. When relay R_m is assigned for cooperation, the mutual information between nodes S and D becomes

$$I_{S,D}{}^{(k)}(R_m) = \frac{1}{2}\log_2(1 + \frac{P_S g_3^{(k)}}{\sigma^2} + \frac{P_S P_R g_1^{(k)} g_2^{(k)}}{\sigma^2(P_S g_1^{(k)} + P_R g_2^{(k)}) + \sigma^4}). \quad (6)$$

An outage event occurs when the link capacity does not meet the required rate. The probability of this event is determined by the average SNR of the link and its channel fading distribution model. When the required rate is R, we define the outage probability of cooperative communications such that R_m is selected as the optimal relay for cooperative communications as $p_{S,D}^{(k)}(R_m)$, and the outage probability of the direct transmission between node S and node D is $p_{S,D}^{(k)}$. Then, the outage probability can be expressed as

$$p_{S,D}^{(k)} = Pr(I_{S,D}^{(k)} < R); \quad (7)$$

$$p_{S,D}^{(k)}(R_m) = Pr(I_{S,D}^{(k)}(R_m) < R). \quad (8)$$

According to [21], [22], $p_{S,D}^{(k)}$ and $p_{S,D}^{(k)}(R_m)$ can be calculated as shown in (4) and (5), respectively.

$$p_{S,D}^{(k)} = 1 - \exp[-\frac{\sigma^2(2^R - 1)}{P_S g_3^{(k)}}]; \quad (9)$$

$$p_{S,D}^{(k)}(R_m) = 1 - \exp[-\frac{\sigma^2(2^R - 1)}{P_S g_3^{(k)} + P_R g_2^{(k)}}]. \quad (10)$$

The SNR can be obtained from node D; thus, $p_{S,D}^{(k)}$ and $p_{S,D}^{(k)}(R_m)$ can be calculated at node D and then sent back to node S through feedback to evaluate the quality of the relay selection.

IV. MDP Analysis of Relay Selection

Given the description of the system model, the collaborative communication process in this paper is analogous to the process of state transition. The system selects and implements an action (i.e., either selects the relay or not) in the current state to obtain the next system state. The system state in the next time slot is only related to the current state and action. Therefore, we model the relay selection problem as MDP [23]. MDP is an optimal decision process for stochastic

dynamic systems based on Markov process (MP), which is an important class of stochastic processes, its original model is Markov chain. MDP is the main research area of sequential decision making, which is the product of the combination of MP and deterministic dynamic programming, so it is also called Markov-type stochastic dynamic programming.

We quantize the channel gain to N levels with $g_i^{(k)} \in \{G_a\}_{1\le a\le N}, i = 2, 3$, which are modeled as a Markov chain with N states. As shown in Fig. 2, the transition probability of the channel gain $g_i^{(k)}$ from G_m to G_n during time slot k is represented by $p_{i,m,n}^{(k)}$:

$$p_{i,m,n}^{(k)} = \text{Prob}(g_i^{(k)} = G_n | g_i^{(k-1)} = G_m). \tag{11}$$

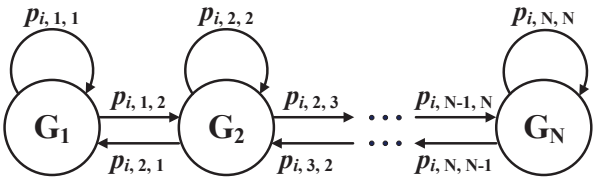


Fig. 2: Markov chain based channel model of the system with N states.

A. Actions

According to the current channel state and system state, we must select an action to execute. We define all actions as

$$\mathbf{A} = \{a^{(k)}\}, k = 1, ..., K, \tag{12}$$

where $a^{(k)} \in \mathbf{A} = \{0, 1, 2, ..., m, ..., M\}$. If $a^{(k)} = 0$, then the system chooses direct transmission at time slot k, if $a^{(k)} = m$, then relay R_m is selected to participate in cooperative communication at time slot k.

B. States

We combine the channel state with the system state and define the set of states as follows:

$$\mathbf{s}^{(k)} = [\rho_1^{(k)}, \rho_2^{(k)}, I^{(k)}], \tag{13}$$

where $\rho_1^{(k)}$ and $\rho_2^{(k)}$ represent the channel states of the $S - D$ link and $R_m - D$ link at time slot k, respectively; and $I^{(k)}$ is the mutual information of the system at time slot k, representing the current system state.

C. Utility

Defining a utility function is highly important in DQ-RSS. The utility function for this system can be defined as

$$u^{(k)}(\mathbf{s}, a) = \ln(I^{(k)}) - c, \tag{14}$$

where c is the power consumption factor, representing the degree of power consumption corresponding to different actions, and $\ln(.)$ is the natural logarithm function. The concavity of the logarithmic function captures the system utility well in terms of mutual information: as mutual information increases, the utility grows more rapidly when mutual information is low than when it is high.

V. Learning-based relay selection scheme

The system states and actions are discrete, and the system state changes discontinuously when an action is performed. Therefore, to solve the MDP problem, we use reinforcement learning to select nodes [24]–[27]. As a type of machine learning method, reinforcement learning adopts the "attempt and failure" mechanism in in interaction with the environment, and uses evaluative feedback signals to achieve decision-making optimization. Because reinforcement learning can find the optimal policy without knowledge of the environment model during the learning process, it has broad application prospects in solving the complex optimization decision problems. With reinforcement learning, the agent selects and performs an action a under state $\mathbf{s}$, after the environment accepts the action, it changes to the next state $\mathbf{s}'$, and feeds a reward signal r to the agent. The agent then selects the follow-up action according to the reward signal. The basic framework for reinforcement learning is shown in Fig. 3.

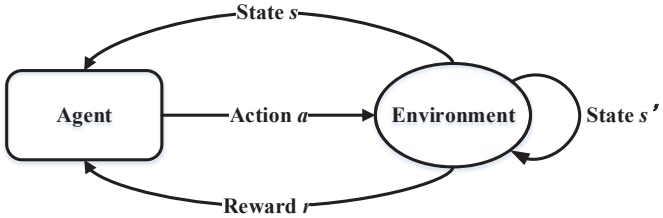


Fig. 3: Basic framework of reinforcement learning.

The Q-learning algorithm as a type of reinforcement learning achieves the optimal actions based on the environment state. However, the Q-learning algorithm will suffer from a low learning speed in a large state space. The generalization and function approximation abilities of deep neural networks can compensate for this limitation. The combination of reinforcement learning and deep learning has demonstrated good performance in many areas [28]–[30]; as such, we decided to use this method to map between states and actions [31]–[33].

A. Q-learning-based relay selection scheme

Here, we introduce a Q-learning-based relay selection scheme into the cooperative communication system in WSNs (Q-RSS). In this scheme, node S acts as an agent to interact continuously with the environment in a trial and error manner to determine the best action. The environment will generate a reward (utility) based on the agent's action feedback. $Q(\mathbf{s}, a)$ is the expectation of gaining action $a \in \mathbf{A}$ in state $\mathbf{s}$ at a certain time. It is updated in each time slot according to the iterative Bellman equation as follows:

$$Q(\mathbf{s}, a) \leftarrow (1 - \alpha)Q(\mathbf{s}, a) + \alpha[u(\mathbf{s}, a) + \delta V(\mathbf{s}')], \tag{15}$$

where $\alpha \in (0, 1]$ represents the learning rate factor and $\delta \in [0, 1]$ is the discount factor, indicating that the value of future rewards is lower than the direct reward. $\mathbf{s}'$ is the next state if node S selects an action a, the value function $V(\mathbf{s})$ maximizes $Q(\mathbf{s}, a)$ over action set $\mathbf{A}$ as given by

$$V(\mathbf{s}) \leftarrow \max_{a\in\mathbf{A}}[Q(\mathbf{s}, a)]. \tag{16}$$

Algorithm 1. Q-learning-based relay selection scheme.

1: Initialize α, δ, $\mathbf{s}^{(0)}$, $\mathbf{A}$, $Q=0$, $V(\mathbf{s})=0$,$\pi=\pi^*$.
2: **for** $k=1,2,3,...,K$ **do**
3: Broadcast the message to all nodes.
4: Choose $a^{(k)}\in\mathbf{A}$ via $\epsilon-greedy$ policy;
5: **if** $a^{(k)}=m$ **then**
6: Send the message to the destination node via the $m-th$ relay node;
7: **if** $a^{(k)}=0$ **then**
8: Send the message to the destination node directly.
9: **end if**
10: Receive the SNR ρ_1, ρ_2, and the mutual information I of the system from destination node.
11: Obtain utility $u^{(k)}$ via (14);
12: Update $Q\left(\mathbf{s}^{(k)},a^{(k)}\right)$ via (15);
13: Update $V\left(\mathbf{s}^{(k)}\right)$ via (16);
14: Update $\pi\left(\mathbf{s}^{(k)},a^{(k)}\right)$ via (17);
15: $\mathbf{s}^{(k+1)}=\left[\rho_1^{(k)},\rho_2^{(k)},I^{(k)}\right]$.
16: **end for**

Node S can apply Q-learning to derive an optimal relay selection policy without knowing the system model. When a relay node is selected as the optimal relay to transmit the data packet to node D at time slot k, the Q-value corresponding to the selected relay is updated as in (15).

The Q-value obtained by the policy π can be represented by $Q^{\pi}(\mathbf{s},a)$, which is defined as the sum of the discount rewards (utility) obtained when action a is taken in the state $\mathbf{s}$ and then follows the policy π. $Q^{\pi^*}(\mathbf{s},a)$ is the Q-value obtained by following the optimal policy π^*. As such, the optimal policy π^* can be derived from

$$\pi^*(x)=\arg\max_{a\in\mathbf{A}}Q^{\pi^*}(\mathbf{s},a). \tag{17}$$

Node S avoids staying at the local optimal by the $\varepsilon-greedy$ algorithm, as summarized in Algorithm 1. The convergence time of Algorithm 1 depends on the size of the state space; therefore, if the dimension of the state is large, the convergence time of Algorithm 1 will be longer.

B. DQN-based relay selection scheme

As mentioned in the previous section, Algorithm 1 is only valid when the state space of the system is small. But as the state space becomes larger, the effect of Algorithm 1 is poor. To overcome this challenge, we apply DQN to accelerate the convergence of Algorithm 1. DQN is a new algorithm that combines deep learning with reinforcement learning to achieve end-to-end learning from perception to action. Put simply, the algorithm shares human characteristics in that it inputs perceptual information (e.g., vision) and then directly outputting motion through deep neural networks without manual work in between. After the introduction of DQN, deep reinforcement learning has received extensive attention, and people began to study it further and apply it to practical applications. In recent years, the results of deep reinforcement learning have emerged in an endless stream. The most representative ones are Google DeepMind's consecutive papers on deep reinforcement learning published in Nature in 2015 and 2016 [9], [10], which marks a new stage in the research and application of deep reinforcement learning. Subsequent deep reinforcement learning is used to handle a range of challenging tasks such as robot control, speech recognition, natural language processing, video analysis, etc.

In DQN, the deep neural network replaces the Q-table; we do not need Q-table lookup and instead only need to calculate $Q(\mathbf{s},a)$ through the neural network. DQN inputs the current state and action, and then uses the deep neural network to obtain an estimate of $Q(\mathbf{s},a)$, given by definition as

$$Q(\mathbf{s},a)=\mathbb{E}[r+\delta\max_{a\in\mathbf{A}}Q(\mathbf{s}',a')]. \tag{18}$$

As shown in Fig. 4, the agent uses two independent deep convolutional neural networks (CNNs) as Q-network approximator: one is the action-value function approximator $Q(\mathbf{s},a;\theta)$, and the other is the target action-value function approximator $Q(\mathbf{s},a;\theta^-)$, where θ and θ^- represent the current and previous parameters, respectively. In each time slot k, the agent (node S) stores its interactive experience tuple $\mathbf{e}^{(k)}=(\mathbf{s}^{(k)},a^{(k)},u^{(k)},\mathbf{s}^{(k+1)})$ in a replay memory $D=\mathbf{e}^{(1)},...,\mathbf{e}^{(k)}$. It then randomly samples from display memory D to update the CNN parameter $\theta^{(k)}$. For the loss function selected from [9], the mean square-error of the target value is minimized in mini-batches. This step is repeated every J time slots.

$$L(\theta^{(k)})=\mathbb{E}[(u+\delta\max_{a\in\mathbf{A}}Q(\mathbf{s}',a';\theta^{(k-1)}-Q(\mathbf{s},a;\theta^{(k)}))^2]; \tag{19}$$

thus

$$\begin{aligned}\nabla_{\theta^{(k)}}L(\theta^{(k)})=&-\mathbb{E}[(u+\delta\max_{a\in\mathbf{A}}Q(\mathbf{s}',a';\theta^{(k-1)}\\&-Q(\mathbf{s},a;\theta^{(k)}))\nabla_{\theta^{(k)}}Q(\mathbf{s},a;\theta^{(k)})].\end{aligned} \tag{20}$$

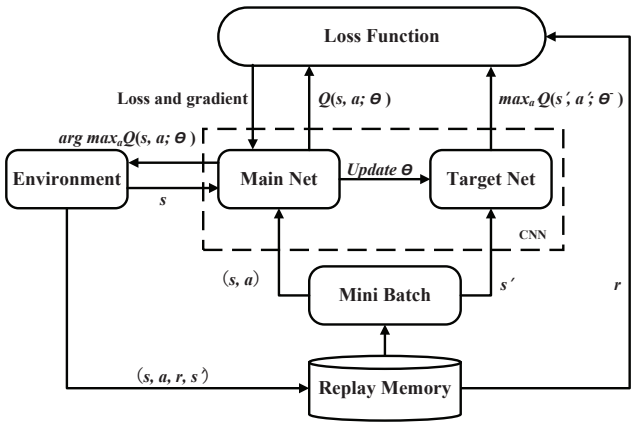


Fig. 4: Illustration of the DQN-based relay selection scheme.

As shown in Fig. 5, the network consists of two convolutional (Conv) layers and two fully connected (Fc) layers to estimate the Q-value in (18) of each action. The first convolutional layer consists of 20 filters, each of which has

TABLE I: Parameters of the CNN

Layer	Input	Filter size	Number of filters	Output
Conv 1	1×9×9	4×4	20	5×5×20
Conv 2	5×5×20	3×3	40	3×3×40
Fc 1	360	\	180	180
Fc 2	180	\	$\|\mathbf{A}\|$	$\|\mathbf{A}\|$

a size of 4×4 and uses the rectified linear unit as an activation function. The second convolutional layer consists of 40 filters, each of which measures 3×3 and uses the same non-linear rectifier. The first Fc layer uses 360 rectified linear units, whereas the second uses 180 units. The length of the status sequence is $W = 13$; that is, the input of CNN is $\phi^{(k)} = (\mathbf{s}^{(k-W)}, \mathbf{s}^{(k-W+1)}, ..., \mathbf{s}^{(k)})$, and the output is the sequence of a given system state. Estimated Q-values of the actions are reshaped into an 8×8 matrix, as summarized in Algorithm 2. The CNN parameters are listed in TABLE I.

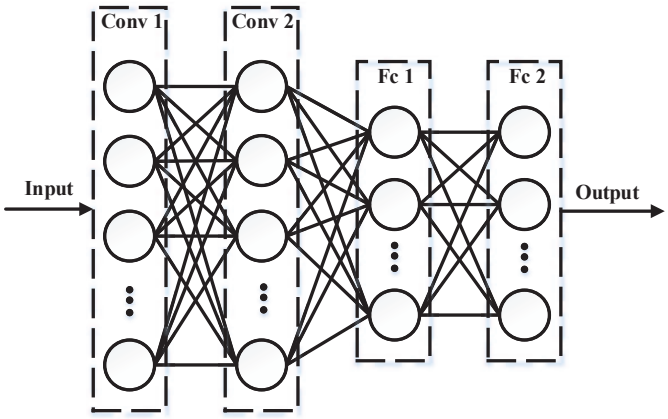


Fig. 5: Illustration of the CNN neural network for approximating Q-function.

Algorithm 2. DQN-based relay selection scheme.

Initialize α, δ, $\mathbf{A}$, D, W, J.
Initialize the Q-network with random weights θ.
Initialize the target Q-network parameters with θ^{-}=θ.
Initialize the beginning state $\mathbf{s}$.
for $k = 1, 2, 3, ..., K$ **do**
  The source broadcast the message to all nodes;
  if $k \leq W$ **then**
    Choose $a^{(k)} \in \mathbf{A}$ at random;
  else
    input $\phi^{(k)} = (\mathbf{s}^{(k-W)}, \mathbf{s}^{(k-W+1)}, ..., \mathbf{s}^{(k)})$ to the CNN as shown in Fig. 5, with weights θ.
    Obtain the CNN output.
    Choose $a^{(k)} \in \mathbf{A}$ via $\epsilon - greedy$;
  end if
  Receive the SNR ρ_1, ρ_2, and the mutual information I of the system from destination node.
  Obtain utility $u^{(k)}$;
  $D \leftarrow D + \{\mathbf{s}^{(k)}, a^{(k)}, u^{(k)}, \mathbf{s}^{(k+1)}\}$
  for $j = 1, 2, 3, ..., J$ **do**
    Select $(\mathbf{s}^{(j)}, a^{(j)}, u^{(j)}, \mathbf{s}^{(j+1)})$ from D at random;
    $y^{(j)} \leftarrow u^{(j)} + \delta \max_{a'} Q(\mathbf{s}^{(j+1)}, a'; \theta)$
  end for
  Calculate $\theta^{(k)}$ via (20).
  Update the CNN weights with $\theta^{(k)}$.
end for

VI. Simulations

A. Simulation settings

We simulate a WSN system where two relay nodes are dropped randomly in a $200m \times 200m$ area with uniform distribution, such that $\mathbf{A} = \{0, 1, 2\}$; we quantize the state to 7 levels, thus the total number of states is $7^3 = 343$; the power consumption corresponding to different actions is $c \in \{0, 0.02, 0.01\}$. A source node is placed at the center. One destination node moves over time in the simulation area, and the mobility model is a random waypoint model. The path-loss coefficient of the channels is taken as $\gamma = 2.4$. The transmit power of the source node is assumed to be 0.1W, and the transmit power of the relay nodes is chosen randomly from the {0.04W, 0.05W, 0.06W} set. The white Gaussian noise power is -10dB. The required outage rate is set at $R = 1$. According to the IEEE 802.15.4 standard protocol, we assume that the radio of all nodes operates in the 2.4 GHz industrial scientific medical band. The learning rate α and discount factor δ are set to 0.7 and 0.5, respectively.

Simulations were performed to evaluate the performance of the learning-based relay selection scheme (DQ-RSS and Q-RSS) and random relay selection scheme (Random); the Random scheme indicates that the relay selection policy is random.

B. Simulation results

As shown in Fig. 6 and Fig. 7, the Q-RSS achieves an optimal policy in the cooperative communication system after convergence. The DQ-RSS achieves an optimal policy even faster and takes much less time to converge. Essentially, the DQ-RSS outperforms the Q-RSS, which in turn exceeds the random relay selection scheme with a higher utility and lower outage probability.

Fig. 8 presents the energy consumption convergence of the proposed scheme and other comparison schemes. All three schemes have the same power allocation. We use set c to indicate the energy consumption of the source node when selecting different actions. From this perspective, we can observe the existing solution. User choices that employ power distribution techniques perform better than other solutions because we can consider the energy consumption factor in the reward, and the DQ-RSS can converge quickly to the optimal policy.

Fig. 9 shows the average utility comparison for the DQ-RSS with the Q-RSS and the random relay selection scheme. The DQ-RSS achieves a much higher average utility as compared with the Q-RSS and the random relay selection scheme.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2925719, IEEE Sensors Journal

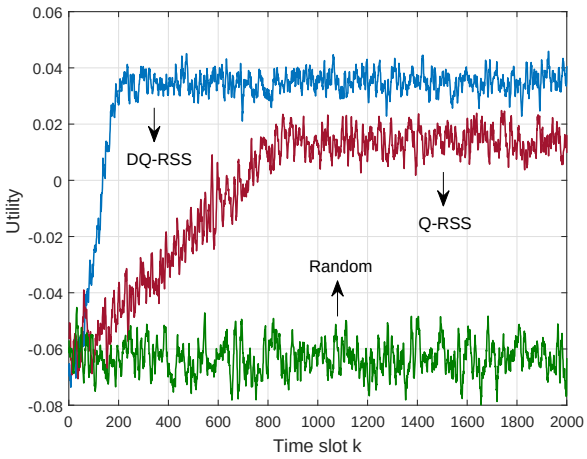


Fig. 6: Performance convergence of the utility.

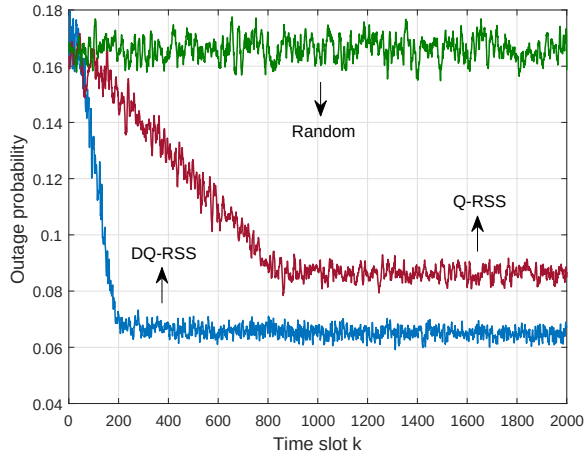


Fig. 7: Performance convergence of the outage probability.

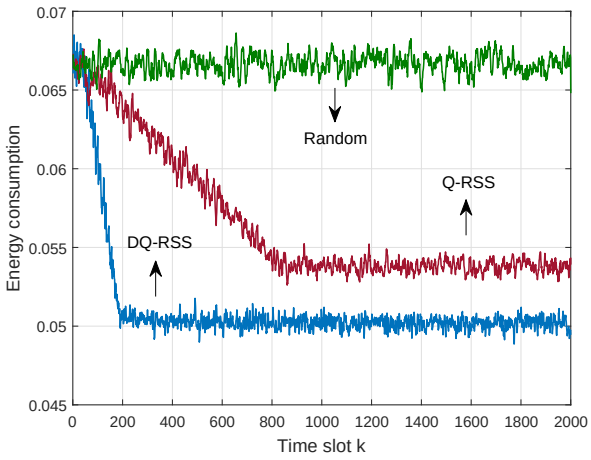


Fig. 8: Performance convergence of the energy consumption.

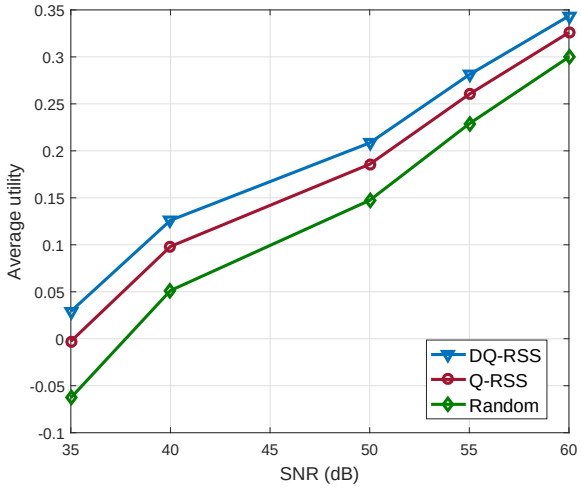


Fig. 9: Performance comparison of the average utility.

Fig. 10 illustrates the average outage probability comparison of the three relay schemes. Clearly, the DQ-RSS realizes a remarkable performance gain. For example, when the SNR is 40dB, the outage probability of the DQ-RSS is roughly two times less than the random relay selection scheme and approximately 30% lower than the Q-RSS. This shows that DQ-RSS can effectively improve the reliability of WSNs. In addition, we can see that with high SNR value, the average outage probability is still high, which is because the Q-RSS and DQ-RSS have high outage probability when they have not converged in the learning process.

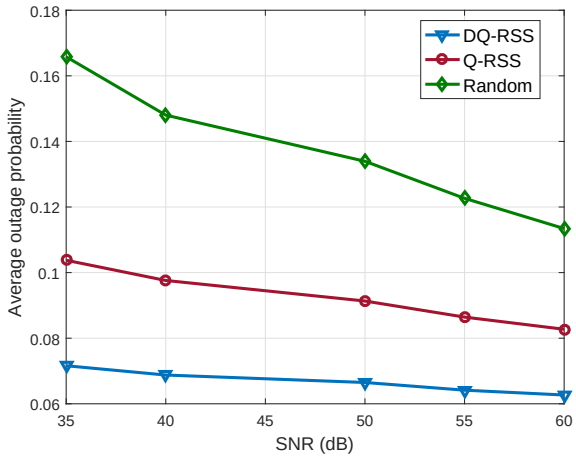


Fig. 10: Performance comparison of the average outage probability.

Fig. 11 displays the average energy consumption of DQ-RSS and the compared schemes. The other two schemes have equal power allocation; the DQ-RSS performs better than these schemes due to effective power allocation.

DQ-RSS can achieve such good performance because it can judge whether relay forwarding is needed in the current system state through feedback regarding system utility. If necessary,

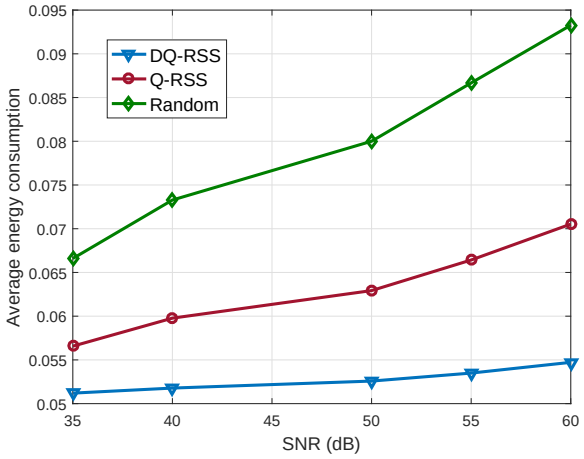


Fig. 11: Performance comparison of the average energy consumption.

TABLE II: Computation overhead of three schemes

Scheme	Memory (MB)	Relay selection time (ms)	Convergence time (time slot)
DQ-RSS	144	0.068	240
Q-RSS	48.4	0.061	1000
Random	24.6	0.059	\

DQ-RSS can adaptively select the optimal relay cooperation to ensure excellent performance.

To verify the DQ-RSS's costs of computer resources, We compared the three relay selection schemes in the aspects of memory, relay selection time and convergence time with python 3.5 on Ubuntu 14. The results are shown in TABLE II. The results show that our proposed relay selection scheme saves the time to converge to the optimal relay selection policy. For example, DQ-RSS takes about 240 time slots to achieve the optimal policy, while Q-RSS requires 1000 time slots to implement. Although DQ-RSS occupies a lot of system memories, it greatly reduces the iteration number and accelerates the convergence, and the learning modules can be deployed on the server (e.g., intelligent sensor motes) independently [34], thus the nodes only need to perform actions according to the policy, reducing the computational overhead and the managing cost. Moreover, most industrial computers have calculation ability to support it.

VII. Conclusions

In this paper, we investigate the use of cooperative communications with adaptive relay selection for WSNs, and propose DQ-RSS, a DQN-based optimal relay selection scheme. We have exploited recent advances and formulated the optimization of opportunistic relaying as a MDP model. A source node collects the CSI from the environment and then sends the integral system state to the DQN to derive the optimal policy for relay selection. Simulation results show that the proposed relay selection scheme exceeds the Q-learning based relay selection scheme and the random relay selection scheme in terms of higher utility, lower outage probability, lower energy consumption, as well as saves the convergence time in the WSN. In the future work, we will consider the mobility of sensor nodes, more complex channel models to study their impacts on actual WSNs.

References

[1] Y. Xiao, V. K. Rayi, B. Sun, X. Du, F. Hu, and J. M. Galloway, "A survey of key management schemes in wireless sensor networks," *Computer Communications*, vol. 30, pp. 2314–2341, 2007.

[2] X. Du, M. Guizani, Y. Xiao, and H.-H. Chen, "Transactions papers a routing-driven elliptic curve cryptography based key management scheme for heterogeneous sensor networks," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 1223–1229, 2009.

[3] A. Nosratinia, T. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 74–80, 2004.

[4] D. W. G. Laneman, J.N.; Tse, "Cooperative diversity wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 3062–3080, 2004.

[5] K. Liu, A. Sadek, W. Su, and A. Kwasinski, "Cooperative communications and networking," *Cambridge University Press: Cambridge, US*, 2008.

[6] A. Ibrahim, A. Sadek, W. Su, and K. Liu, "Cooperative communications with relay-selection: when to cooperate and whom to cooperate with?," *IEEE Trans. Wirel. Commun.*, vol. 7, no. 7, pp. 2814–2827, 2008.

[7] A. Bletsas, A. Khisti, D. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, 2006.

[8] M. A. Jadoon and S. Kim, "Relay selection algorithm for wireless cooperative networks: a learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 11, no. 7, pp. 1061–1066, 2017.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. 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," *IEEE Transactions on Industrial Electronics*, no. 518, pp. 529–533, 2015.

[10] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, D. S. Lanctot, M., N. J. Grewe, D., N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, no. 529, pp. 484–489, 2016.

[11] A. Bletsas, H. Shin, and M. Win, "Cooperative communications with outage-optimal opportunistic relaying," *IEEE Trans. Wirel. Commun.*, vol. 6, no. 9, pp. 3450–3460, 2007.

[12] Y. Lin, J. Song, and V. Wong, "Cooperative protocols design for wireless ad-hoc networks with multi-hop routing," *Mobile Networks and Applications*, vol. 14, no. 2, pp. 143–153, 2009.

[13] L. Sun, T. Zhang, L. Lu, and H. Niu, "Cooperative communications with relay selection in wireless sensor networks," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 2, pp. 513–517, 2009.

[14] W. Fang, F. Liu, F. Yang, L. Shu, and S. Nishio, "Energy-efficient cooperative communication for data transmission in wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 56, no. 4, pp. 2185–2192, 2010.

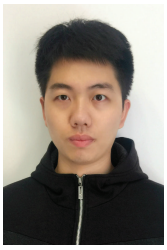
[15] L. Li, K. Namuduri, and S. Fu, "Cooperative communication based on random beamforming strategy in wireless sensor networks," *2012 IEEE Global Communications Conference (GLOBECOM)*, Anaheim, CA, USA. 3-7 Dec. 2012.

[16] X. Liang, I. Balasingham, and V. C. M. Leung, "Cooperative communications with relay selection for qos provisioning in wireless sensor networks," *2009 IEEE Global Telecommunications Conference (GLOBECOM)*, Honolulu, HI. 30 Nov.-4 Dec. 2009.

[17] S. Galzarano, A. Liotta, and G. Fortino, "Ql-mac: A q-learning based mac for wireless sensor networks," *International Conference on Algorithms & Architectures for Parallel Processing*, New York, Inc. 2013.

[18] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "Uav relay in vanets against smart jamming with reinforcement learning," *IEEE Trans. Vehic. Tech.*, vol. 67, no. 5, pp. 4087–4097, 2018.

[19] X. Lu, D. Xu, L. Xiao, L. Wang, and W. Zhuang, "Anti-jamming communication game for uav-aided vanets," *2017 IEEE Global Communications Conference (GLOBECOM)*, Singapore. 4-8 Dec. 2017.
[20] V. Erceg, L. Greenstein, S. Tjandra, S. Parkoff, A. Gupta, B. Kulic, A. Julius, and R. Bianchi, "An empirically based path loss model for wireless channels in suburban environments," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 7, pp. 1205–1211, 1999.
[21] A. Ibrahim, Z. Han, and K. Liu, "Distributed energy-efficient cooperative routing in wireless networks," *IEEE Trans. Wirel. Commun.*, vol. 7, no. 10, pp. 3930–3941, 2008.
[22] X. Du, Y. Xiao, M. Guizani, and H.-H. Chen, "An effective key management scheme for heterogeneous sensor networks," *Ad Hoc Networks*, vol. 5, pp. 24–34, 2007.
[23] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.
[24] H. M., "W-learning: competition among selfish q-learners," Cambridge: University of Cambridge, 1995.
[25] L. D. S. G. Wei, Q., "A novel dual iterative, q-learning method for optimal battery management in smart residential environments," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 2509–2518, 2015.
[26] J. Ni, M. Liu, L. Ren, and S. Yang, "A multiagent q-learning-based optimal allocation approach for urban water resource management system," *IEEE Transactions on Automation Science & Engineering*, vol. 91, no. 2, pp. 331–356, 2003.
[27] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
[28] M. Spencer, J. Eickholt, and J. Cheng, "A deep learning network approach to abinitio protein secondary structure prediction," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 12, no. 1, pp. 103–112, 2015.
[29] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.
[30] Y. Su, L. Huang, and C. Feng, "Qred: A q-learning-based active queue management scheme," *Journal of Internet Technology*, vol. 19, no. 4, pp. 1169–1178, 2018.
[31] T. Kobayashi, T. Shibuya, and M. Morita, "Q-learning in continuous state-action space by using a selective desensitization neural network," *Ieice Technical Report Neurocomputing*, vol. 111, pp. 119–123, 2011.
[32] Y. Su, X. Du, L. Huang, Z. Gao, and M. Guizani, "Lte-u and wi-fi coexistence algorithm based on q-learning in multi-channel," *IEEE Access*, vol. 6, pp. 13644–13652, 2018.
[33] T. Teng and A. Tan, "Reinforcement learning under uncertainties with self-organizing neural networks," *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Singapore. 6-9 Dec. 2015.
[34] C. Savaglio, P. Pace, G. Aloi, A. Liotta, and G. Fortino, "Lightweight reinforcement learning for energy efficient communications in wireless sensor networks," *IEEE Access*, vol. 7, pp. 29355–29364, 2019.



Yuhan Su received the B.S. degree from Huaqiao University, Xiamen, China, in 2015. He is currently pursuing the Ph.D. degree in communication and information system with Xiamen University, Xiamen, China. His research interests include wireless communications, congestion control, radio resource management and network coding.

Xiaozhen Lu received the B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2017. She is currently pursuing the PhD.degree with the Department of Communication Engineering, Xiamen University, Xiamen, China. Her research interests include network security and wireless communications.

Yifeng Zhao received his B.S. degree in Communication Engineering in 2002, M.S degree in electronic circuit system in 2005 and Ph.D. degree in Communication Engineering in 2014 from Xiamen University. He is an assistant professor of Communication Engineering, Xiamen University, Xiamen, Fujian, China. His current research interests include Mmwave communication, Massive MIMO and Machine learning applied in wireless communications.

Lianfen Huang received the B.S. degree in radio physics and the Ph.D. degree in communication engineering from Xiamen University, Xiamen, China, in 1984 and 2008, respectively. She was a Visiting Scholar with Tsinghua University, Beijing, China, in 1997, and with The Chinese University of Hong Kong, Hong Kong, in 2012. She is currently a Professor of communication engineering with Xiamen University. Her research interests include wireless communication, wireless network, and signal process.

Xiaojiang Du received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park in 2002 and 2003, respectively. He is currently a Professor with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. He received over U.S. 5 million research grants from the U.S. National Science Foundation, the Army Research Office, Air Force, NASA, the State of Pennsylvania, and Amazon. His research interests include security, wireless networks, and systems. He has authored over 320 journals and conference papers in these areas, as well as a book published by Springer. He is a Life Member of the ACM. He received the Best Paper Award at the IEEE GLOBECOM 2014 and the Best Poster Runner-Up Award at the ACM MobiHoc 2014. He served as the Lead Chair of the Communication and Information Security Symposium of the IEEE International Communication Conference 2015 and a Co-Chair of the Mobile and Wireless Networks Track of the IEEE Wireless Communications and Networking Conference 2015.