Computers, Materials & Continua
DOI:10.32604/cmc.2020.014484
Article

Tech Science Press

Cooperative Channel Assignment for VANETs Based on Dual Reinforcement Learning

Xuting Duan[1,2], Yuanhao Zhao[1,2], Kunxian Zheng[1,2,*], Daxin Tian[1,2], Jianshan Zhou[1,2,3] and Jian Gao[4]

[1]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, 100191, China
[2]School of Transportation Science and Engineering, Beihang University, Beijing, 100191, China
[3]Department of Engineering and Design, University of Sussex, Brighton, BN1 9RH, UK
[4]Research Institute of Highway Ministry of Transport, Beijing, 100088, China
[*]Corresponding Author: Kunxian Zheng. Email: zhengkunxian@buaa.edu.cn

Received: 23 September 2020; Accepted: 06 October 2020

Abstract: Dynamic channel assignment (DCA) is significant for extending vehicular *ad hoc* network (VANET) capacity and mitigating congestion. However, the un-known global state information and the lack of centralized control make channel assignment performances a challenging task in a distributed vehicular direct communication scenario. In our preliminary field test for communication under V2X scenario, we find that the existing DCA technology cannot fully meet the communication performance requirements of VANET. In order to improve the communication performance, we firstly demonstrate the feasibility and potential of reinforcement learning (RL) method in joint channel selection decision and access fallback adaptation design in this paper. Besides, a dual reinforcement learning (DRL)-based cooperative DCA (DRL-CDCA) mechanism is proposed. Specifically, DRL-CDCA jointly optimizes the decision-making behaviors of both the channel selection and back-off adaptation based on a multi-agent dual reinforcement learning framework. Besides, nodes locally share and incorporate their individual rewards after each communication to achieve regional consistency optimization. Simulation results show that the proposed DRL-CDCA can better reduce the one-hop packet delay, improve the packet delivery ratio on average when compared with two other existing mechanisms.

Keywords: Vehicular *ad hoc* networks; reinforcement learning; dynamic channel assignment

1 Introduction

VANET is a specific application of MANET (Mobile Ad-hoc Network) in vehicle to vehicle/vehicle to infrastructure communication scenario. As a research hotspot of intelligent transportation, VANETs lay a crucial foundation for various intelligent vehicular functions. Traditional MANETs adopt the single-channel communication mode where all nodes can only access one common channel for data transmission. With the widespread use of connected vehicles in the future, the quantity of VANETs nodes increase continuously, leading to progressively fierce competition in wireless resources. The single-channel communication mode tends to cause severe resources conflicts when large numbers of nodes access the channel concurrently. Therefore, the capacity of the wireless network using the traditional single-channel communication mode is seriously limited by the quantity of channels.

As a widely used VANETs wireless communication protocol standard, WAVE (Wireless Access in Vehicular Environment) provides a 75 MHz bandwidth in the 5.9 GHz frequency band for vehicle to vehicle/vehicle to infrastructure communication and divides the 75 MHz bandwidth into seven channels. These channels enable nodes to transmit data packets simultaneously under diverse channels. CH178 is the control channel (CCH) that can only be used to transmit control and public security information. CH174, CH176, CH180 and CH182 are service channels (SCHs), used to transmit both public security and private service information. CH184 and CH172 are reserved for future use. IEEE 1609.4 is used for multi-channel operations, such as channel synchronization, coordination, and switching of WAVE. WAVE providers broadcast WAVE Service Advertisement (WSA) packets containing the offered services information and the network parameters necessary to join the advertised Basic Service Set (BSS) [1]. After receiving the WSA, the WAVE users interested in the service access the corresponding SCH in the SCH Interval (SCHI) to obtain service data. However, we find that the quality of service (QoS) of VANET both in line-of-sight (LOS) and none-line-of-sight (NLOS) scenarios are not ideal in the preliminary field experiment. This channel coordination mode is not suitable for a vehicular direct communication scenario. For example, in a unicast multi-hop routing scenario, the data transmitting node needs to transmit data on a specific SCH with the next hop node selected by the routing protocol. The process of the optimal SCH selection and the channel coordination therein is not clearly defined in IEEE 1609.4.

The limited communication range, the change of network topology, and the distributed execution of channel assignments between the nodes make the global network state of VANETs unknown to the nodes. Therefore, the channel assignment of local vehicular direct communication is actually an optimization problem under an unknown state model. In addition, the lack of centralized control makes DCA more challenging. This paper applies RL to the DCA problem in a dynamic environment because of its widespread usage and outstanding performance in the field of optimal decision-making without state models. In order to meet the dual requirements of VANETs for network capacity and latency, we design a dual RL framework to jointly optimizes the decision-making behaviors of both the channel selection and back-off adaptation, and achieve multi-agent collaboration by sharing their individual rewards. Finally, our DRL-CDCA is compared with two other conventional baseline mechanisms under the same simulation scenario. After simulation, it can be found that DRL-CDCA is superior to two other conventional baseline mechanisms in the one-hop packet delay and the packet delivery ratio on average.

The main contributions of this paper are summarized as follows:

- As an important branch of machine learning, the existing RL theory has made some achievements in many fields, but it is mainly based on the interaction between environment and robot and game development. Owing to few people effort the application prospect of RL theory in the Internet of vehicles, its development of RL theory in promoting joint channel selection and medium access control (MAC) layer back-off for vehicle to vehicle (V2V) communication and networking is not mature. We propose the first work to demonstrate the feasibility and potential of RL based method in joint channel selection decision and access fallback adaptation design to enhance V2V communication.
- In addition, we improve the original RL theory to combine RL theory with vehicle field, and design and combine two components to adapt to and improve the decision-making performance of RL agent in V2V communication: (I) A dual Q network structure for joint optimization of channel selection and reverse adaptation; (II) A distributed consensus reward mechanism to promote cooperative decision-making among learners Behavior.

The remainder of this paper is organized as follows: Section 2 mainly introduces related work in channel assignment. Section 3 mainly displays the previous field experiments and the problems found according to the experimental results Section 4 describes the system model and problem formulation. Section 5 describes the details of the mechanism proposed in this paper. The performance of our DRL-CDCA is compared to two other existing mechanisms in Section 6. Finally, concluding remarks are presented in Section 7.

2 Related Work

Q-learning-based DCA proposed in Nie et al. [2] uses RL to solve DCA problem in a cellular mobile communication system. Q-learning-based DCA is a single agent RL (SARL) mechanism, as well as a centralized one. The base station as a centralized node assigns the channel to each communication node pair. However, the channel assignment of VANETs is not done by a central node like the base station of Q-learning-based DCA; Instead, each node assigns the channel independently. Therefore, the centralized channel assignment mechanism like Q-learning-based DCA for cellular mobile communication systems cannot adapt to VANET. A novel deep RL (DRL)-based DCA (DRL-DCA) algorithm is proposed in Liu et al. [3], where the system state is reformulated into an image-like fashion, and then, convolutional neural network is used to extract useful features. DRL-DCA models the multibeam satellite system as the agent and the service event as the environment. From the perspective of VANETs, DRL-DCA is equivalent to the RSU (roadside unit)-based channel assignment mechanism, which is a centralized channel assignment mechanism. Wei et al. [4] combines RL method Q-learning with deep neural network to approximate the value function in complex control application. The RL-CAA mentioned in Ahmed et al. [5], and the RLAM mentioned in Louta et al. [6] are also the centralized channel assignment mechanisms in different application scenarios, and are not suitable for the distributed channel assignment problem of the vehicular direct communication scenario studied in this paper. Therefore, this paper models the Markov decision process of RL for distributed scenarios, and designs the DCA mechanism based on the distributed RL model.

In recent years, some advanced MAC (medium access control) protocols [7–9] have been designed to enhance the communication capabilities of VANET. An adaptive multi-channel assignment and coordination (AMAC) scheme for the IEEE 802.11p/1609.4 is proposed in Almohammedi et al. [10], which exploits channel access scheduling and channel switching in a novel way. However, AMAC's channel selection mechanism is still based on WBSS (WAVE Basic Service Set) service release-subscription, not for vehicular direct communication scenarios. A mechanism called safety communication based adaptive multi-channel assignment is proposed in Chantaraskul et al. [11] to adaptively adjust the channel switching interval. However, there is no mention of the strategy of SCH selection in Chantaraskul et al. [11]. An RSU-coordinated synchronous multi-channel MAC scheme for VANETs is proposed in Li et al. [12], which supports simultaneous transmissions on different SCHs. However, in the scenario where the vehicle-to-vehicle is directly connected, the MAC mechanism without RSU cooperation cannot be realized. In Ribal et al. [13], deep reinforcement learning is applied to VANETs. Specifically, it is used to implement the vehicle to RSU that meets QoS requirements. The RSU can distinguish different system states according to its remaining battery, the quantity of the mobile nodes and the communication requests before assigning suitable SCH to each OBU (on board unit). The method proposed in Ribal et al. [13] still belongs to the RSU-based channel assignment mechanism, which cannot solve the channel assignment problem in a fully distributed scenario.

Furthermore, due to the uneven traffic flow density, VANET node density will also be affected. At the same time, the routing is of multi hop and multi node, which brings severe challenges to the operation and optimization of networking and transmission. Liu et al. [14] gives a novel multi-hop algorithm for wireless network with unevenly distributed nodes. In this algorithm, each unknown node estimates its location using the mapping model by elastic net constructed with anchor nodes. However, as an important feature of VANET, the randomness and uncertainty of the target motion are not discussed in depth. A referential node scheme for VANET is proposed in Wang et al. [15], where further analysis for channel assignment is not implemented.

It can be seen that a large part of the channel assignment mechanism for VANETs is a central channel assignment mechanism, and some of the existing fully distributed channel assignment mechanisms are still basically based on the WBSS service release-subscription, not for the vehicular direct communication scenario.

3 Preliminary Field Experiment

In the preliminary work of this paper, a field experiment for vehicle road collaborative application was carried out in Tong Zhou automobile test field, Beijing. The filed testing environment is shown in Fig. 1. In the experiment, the communication performance of VANET is measured in LOS and NLOS scenarios respectively in different level of vehicle speed. As a representative index to evaluate network performance, the QoS of the VANET was emphatically noticed.

Figure 1: Beijing Tong Zhou automobile test base

During the experiment, a vehicle equipped with on-board unit runs on the test road, while roadside units are deployed in a relatively fixed location. On-board terminal and roadside terminal form network through communication between on-board unit and roadside unit. Topology structure of VANETs' testing is shown in Fig. 2. When the experiment launching, the applications of driving road collaboration are developed by the data exchange and share between roadside and vehicle. The tester monitors the on-board unit and roadside unit with computer in and out of the vehicle respectively. Then QoS indicators are further calculated.

In this field experiment, QoS refers to the selection of delay time, delay variation, package loss rates, and throughput. And the ordinary DCA strategy is employed in vehicular networking. The results in various speed and visibility scenarios are shown in Fig. 3.

The results reveal that although the current strategy has almost successfully controlled the time delay in less than 10 ms, the delay variation is changed frequently. Meanwhile, although packet loss does not appear to be serious, considering that there is only one vehicle participating in the test and there is only one single hop routing in this network, its test results can not reflect the universal situation of multi hop and multi routing. In addition, the throughput is within 2 Mb/s in most case while the rated throughput of the communication equipment is 4 Mb/s. Employing the existing DCA strategy, the throughput is 50% less than the rated throughput in order to meet the communication accessibility.

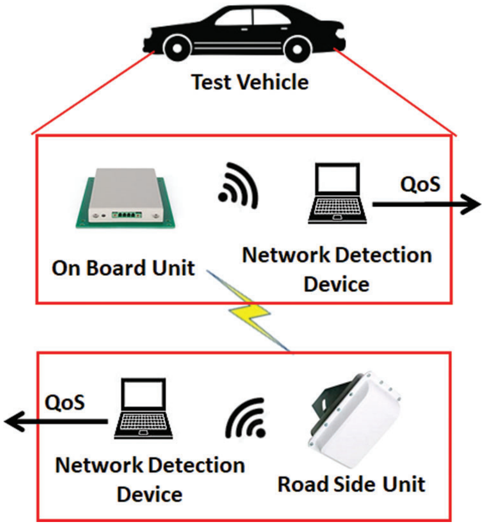


Figure 2: Topology structure of VANET performance testing

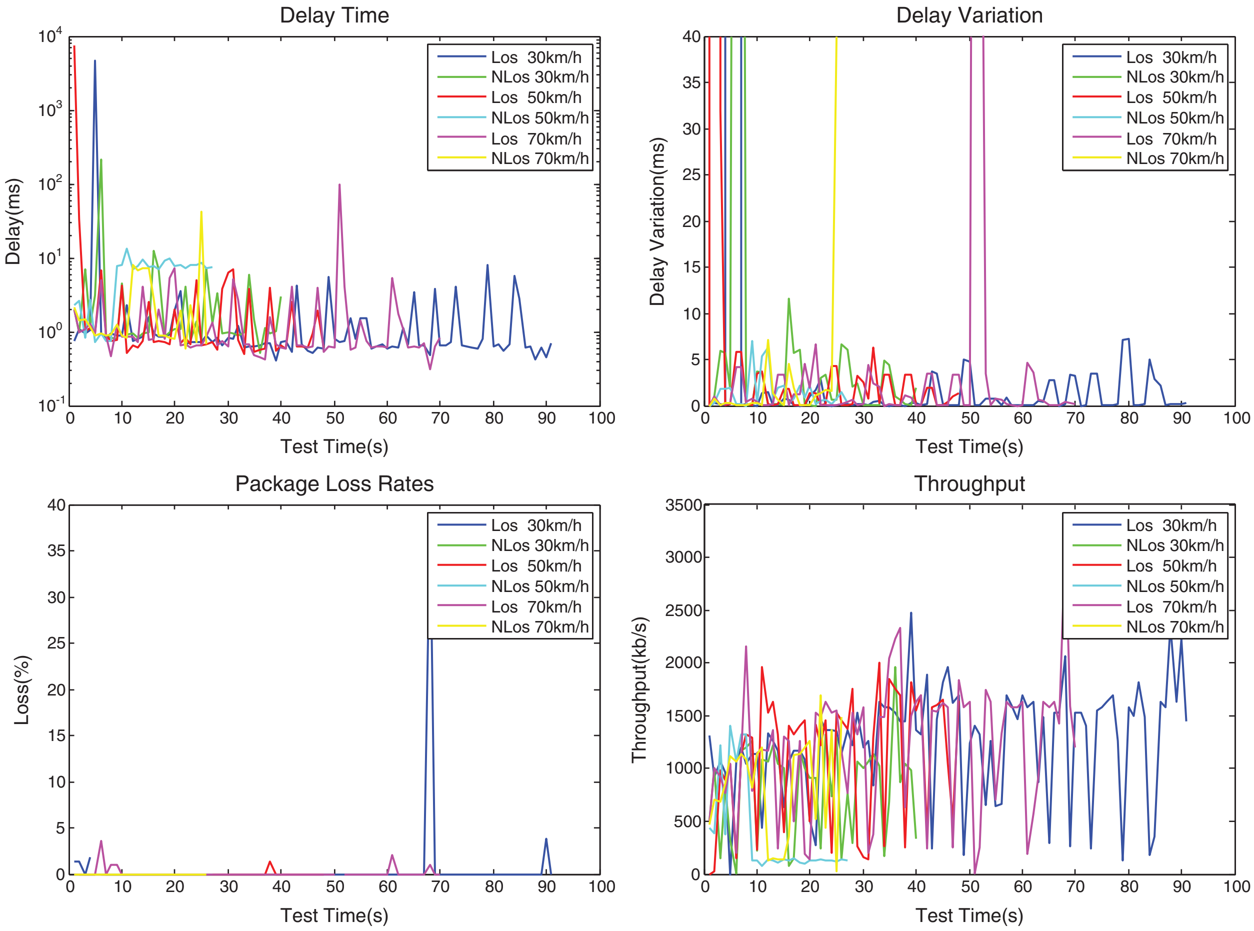


Figure 3: QoS test results in various speed and visibility scenarios

The field experiment results illustrate that the existing DCA is difficult to adapt to VANET. To optimize the channel allocation and networking problems in VANET, this paper proposes DRL-CDCA. Considering the difficulty of deploying the multi hop and multi routing experiment under the condition of multi vehicles in the test field, this paper explores the situation of multi hop and multi routing by simulation.

4 System Model and Problem Formulation

The current state s_{in} observed by RL agent i at the n-th time slot is related to pervious state s_{n-1} and action a_{n-1}, which is a Markov Decision Process (MDP), and can be described by a 5-tuple array $(S_i, A_i, P_i, R_i, \gamma)$:

- S_i: A set of states observed by RL agent i at different time slot, where $s_{in} \in S_i$, s_{in} is the system state observed by the RL agent i at the n-th time slot.
- A_i: A set of actions performed by RL agent i at different time slot, where $a_{in} \in A_i$, a_{in} is the action performed by the RL agent i at the n-th time slot.
- P_i: The state transition probability, which is the probability distribution of the state transition after taking action a in state s. The probability of changing from s_{in} to $s_{i(n+1)}$ after taking action a_{in} can be expressed as $p(s_{i(n+1)}|s_{in}, a_{in})$.
- R_i: The reward function, where $r_{in} \in R_i$, r_{in} is the reward obtained by the RL agent i at the n-th time slot.
- γ: The discount factor, which is used to assign the weight between real-time rewards and long-term rewards. When $\gamma = 0$, the agent only considers real-time rewards, and $\gamma = 1$ means that long-term rewards and real-time rewards are equally important.

In order to jointly optimize channel allocation and back-off adaptation, we apply dual MDP to the DCA problem. The state, action and reward functions of each MDP in dual MDP are described in detail below.

4.1 State

We assume that the quantity of SCH is K. The state of each MDP in dual MDP are described by follow.

- **Channel Selection:** The state of channel selection MDP is $s_{in}^{c} = \{C_{in},\ E_{in}\}$, where $C_{in} = [c_{1in}, \ldots, c_{kin}, \ldots, c_{Kin}]$ denotes the channel busy situation, and c_{kin} is the number of CNPs (communication node pairs) intending to transmit over the k-th SCH. E_{in} denotes the local communication requirement of node i at n-th time slot.
- **Back-off Adaptation:** The state of back-off adaptation MDP is $s_{in}^{b} = \left\{E_{in},\ W_{i(n-1)}\right\}$, where $W_{i(n-1)}$ denotes the back-off window size in the $(n-1)$-th time slot.

4.2 Action

- **Channel Selection:** The action of channel selection MDP is $k_{in} (1 \leq k_{in} \leq K)$, which represents the index of the SCH selected in the n-th time slot by RL agent i.
- **Back-off Adaptation:** The action of back-off adaptation MDP is w_{in}, and $w_{in} \in \{w_{1in}, w_{2in}, w_{3in}\}$, where w_{1in} denotes that agent i maintains the current bock-off window size as $W_{in} = W_{i(n-1)}$, w_{2in} denotes that agent i increases the back-off window size as $W_{in} = 2W_{i(n-1)} + 1$, w_{3in} denotes that agent i reduces the back-off window size as $W_{in} = \dfrac{W_{i(n-1)} - 1}{2}$.

4.3 Reward

We take consideration the dynamics user communication demand and communication performance as indicators of model training. The reward function is designed as follows.

$$r_{in} = \left(\frac{g_{rec}}{g_{tra}}\right)^{\xi} \left(\frac{g_{rec}}{g_{tra} + g_{que}}\right)^{\varrho} \tag{1}$$

where ξ and ϱ are two positive weights, g_{tra} denotes the number of packets transferred by agent i, g_{rec} denotes the number of packets successfully delivered to the receiver, and g_{que} denotes the number of packets waiting in the buffer queue that need to be transferred in the n-th time slot. We use the packet delivery ratio $\left(\frac{g_{rec}}{g_{tra}}\right)$ to denote the communication performance, and the sum of packets transferred and waiting in the buffer to denote the user communication demand at the n-th time slot.

In order to achieve multi-agent collaborative optimization in a distributed manner, we propose a novel reward formulation with a weighted sum strategy, termed consensus reward, which is constructed as:

$$R_{in} = r_{in} + \sum_{r'_{inj} \in R'_{in}} \beta_{inj} r'_{inj} \tag{2}$$

where R'_{in} is the set of local rewards of the neighbor nodes of node i, and r'_{inj} is the local reward of node j, β_{inj} is the weight of r'_{inj}.

The model input S_i, action A_i, and reward R_i of the dual MDP are defined above. To obtain the optimal strategy, the state transition probability P_i and the discount factor γ must be determined. Between them, γ is artificially set. The precondition for determining the state transition probability P_i is the known environment model. However, it is hardly to accurately obtain the system state at the next time slot in VANETs. Consequently, the DCA problem is an unknown environment model problem. In this paper, Q-learning as a widely used RL algorithm is used to achieve the optimal strategy. The base Q-learning strategy update formula is as follows [16]:

$$Q_{n+1}(s_{in}, a_{in}) \leftarrow Q_n(s_{in}, a_{in}) + \alpha\left[R_{in} + \gamma \max_{a_i} Q_n\left(s_{i(n+1)}, a_i\right) - Q_n(s_{in}, a_{in})\right] \tag{3}$$

where $Q_n(s_{in}, a_{in})$ denotes the value function when taking action a_{in} in state s_{in}, and α is the learning rate that denotes the magnitude of the strategy update.

5 Proposed Channel Assignment Mechanism

5.1 Strategy Execution and Update

The model input of DRL-CDCA is a vector of continuous values. Therefore, the state-action value cannot be stored by Q-table. DRL-CDCA uses the neural network to approximate the state-action value. Specifically, we construct two neural networks, i.e., dual neural networks, one of which de noted by $Q_n\left(s^c_{in}, k_{in}; \theta\right)$ is used for channel selection and the other denoted by $Q_n\left(s^b_{in}, w_{in}; \varpi\right)$ for back-off adaptation. θ and ϖ are the weights of the dual neural networks. Fig. 4 shows our methodological framework.

Weights θ and ϖ are updated as following:

$$\begin{cases} Q_{n+1}\left(s^c_{in}, k_{in}; \theta\right) \leftarrow Q_n\left(s^c_{in}, k_{in}; \theta\right) + \alpha\Big[R_{in} \\ \qquad\qquad + \gamma \max_{k_{in}} Q_n\left(s^c_{i(n+1)}, k_{in}; \theta^-\right) - Q_n(s^c_{in}, k_{in}; \theta)\Big] \\ Q_{n+1}\left(s^b_{in}, w_{in}; \varpi\right) \leftarrow Q_n\left(s^b_{in}, w_{in}; \varpi\right) + \alpha\Big[R_{in} \\ \qquad\qquad + \gamma \max_{w_{in}} Q_n\left(s^b_{i(n+1)}, w_{in}; \varpi^-\right) - Q_n(s^b_{in}, w_{in}; \varpi)\Big] \end{cases} \tag{4}$$

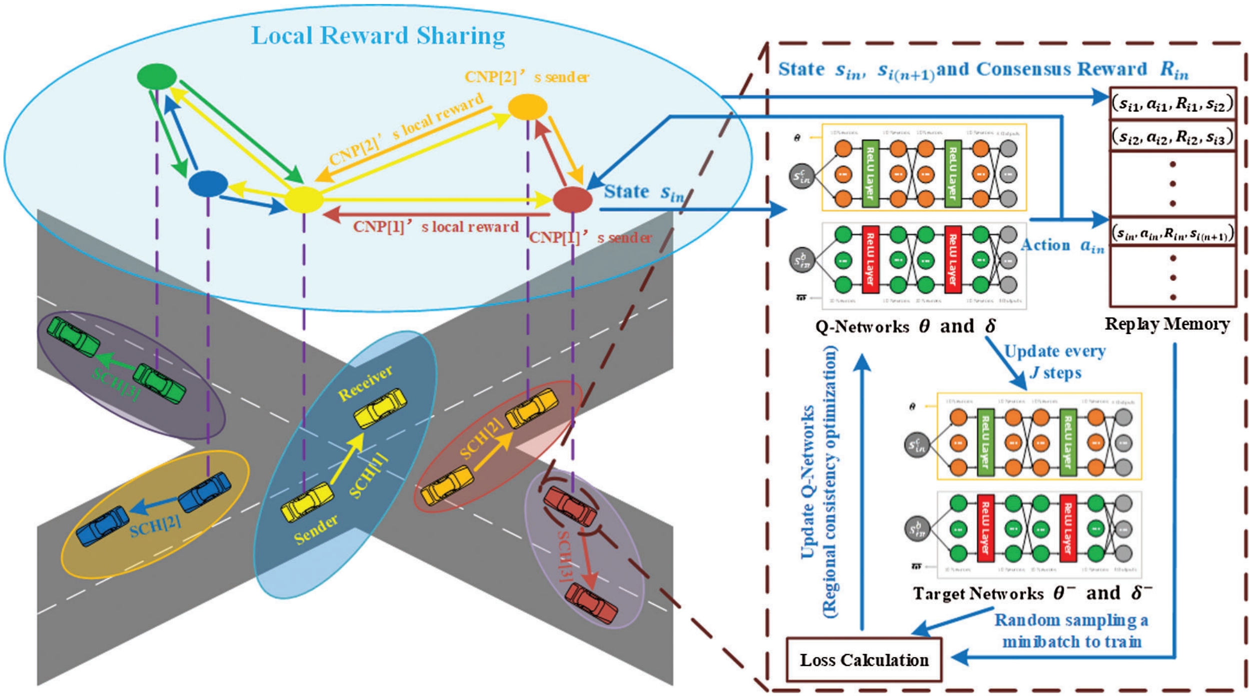

Local Reward Sharing
CNP[2]' s sender
CNP[2]' s local reward
CNP[1]' s local reward
CNP[1]' s sender
State s_{in}
Receiver
Sender
SCH[1]
SCH[2]
SCH[3]
State s_{in}, $s_{i(n+1)}$ and Consensus Reward R_{in}
$(s_{i1}, a_{i1}, R_{i1}, s_{i2})$
$(s_{i2}, a_{i2}, R_{i2}, s_{i3})$
$(s_{in}, a_{in}, R_{in}, s_{i(n+1)})$
Action a_{in}
ReLU Layer
s^c_{in}
s^b_{in}
θ
ϖ
Q-Networks θ and δ
Replay Memory
Update every J steps
Update Q-Networks (Regional consistency optimization)
Target Networks θ^- and δ^-
Random sampling a minibatch to train
Loss Calculation

Figure 4: A scenario of multiple co-existing communication node pairs (CNPs) Driven by multi-agent DRL-CDCA

5.2 Channel Access Process

VANETs nodes *i* and *j* exchange RTS/CTS in the CCH interval for channel coordination, and then nodes *i* and *j* switch to the selected SCH *k* for data transmission in the SCH interval. To better understand channel access, an example shown in Fig. 5 is introduced. There are communication demands between nodes A and B, nodes C and D in the CCH interval. Nodes A and C compete for the access to CCH. Assume that node A first obtains the transmission opportunity. After node A successfully accesses CCH, it sends RTS to node B. When node B receives RTS, it selects SCH channel. Assume that SCH [1] is selected as the transmission channel of nodes A and B in SCH interval. Node B broadcasts CTS containing the information about channel coordination between itself and node A. After receiving CTS, the neighboring nodes of node B update their local state. Then node C attempts to access CCH again. Assume that SCH [3] is selected as the transmission channel of nodes C and D in the SCH interval. In SCH interval, nodes A and C do not immediately send data to the MAC layer. Instead, they randomly evade for a period of time when entering SCHI. The window size W_{in} of the random back-off process is determined by ϖ. Then nodes A and B switch to SCH [1], nodes C and D switch to SCH [3], and transmit data in their own SCH. The receiving node will send feedback ACK as soon as it receives the data packet. The overall algorithm is given in Algorithm 1.

6 Results and Discussion

6.1 Simulation Parameters

The simulation experiment in this paper is based on Veins, which is further based on two simulators: OMNeT++, an event-based network simulator, and SUMO, a road traffic simulator. The neural networks of DRL-CDCA is supported by the third-party machine learning C++ library MLPACK. The simulation scenario shown in Fig. 6 is a part of the scenario based on the city of Erlangen. The parameter settings are given in Tab. 1.

Algorithm 1: Algorithm for Multi-agent DRL-CDCA

```
Algorithm 1 Multi-agent DRL-CDCA
Initialize:
The weights of dual neural networks, training cycle $F$ , update cycle $J$
Run:
For $n = 0 \rightarrow +\infty$ Do
    If CCH Interval Then
        Send RTS/CTS, coordinates the SCH according network $\theta$;
    End If
    If SCH Interval Then
        Source determines the back-off window size $W_{in}$ according network $\varpi$;
        Back-off $t \in [0, W_{in}]$;
        Source observers $S_{i(n+1)}$ and send data packets;
    End If
    If Next CCH Interval Then
        Source calculates local reward;
        Source broadcasts payoff message;
    End If
    If Next SCH Interval Then
        Source calculates the consensus reward;
        Store transition $[s_{in}, a_{in}, R_{in}, s_{i(n+1)}]$ in experience memory;
    End If
    Every $F$ steps, randomly sample a minibatch tuples from the memory;
    Train the dual Q-networks by RMSProp;
    Every $J$ steps, copy weights into target network $\theta^-$ and $\varpi^-$;
End For
```

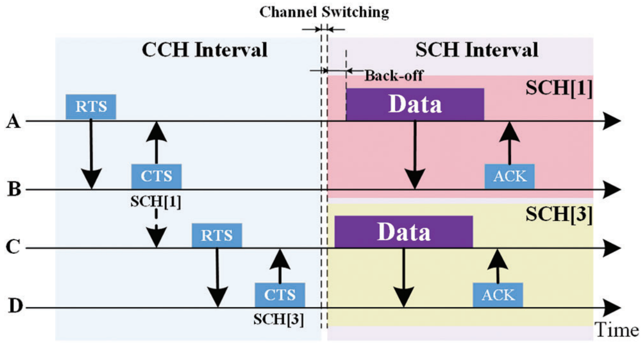
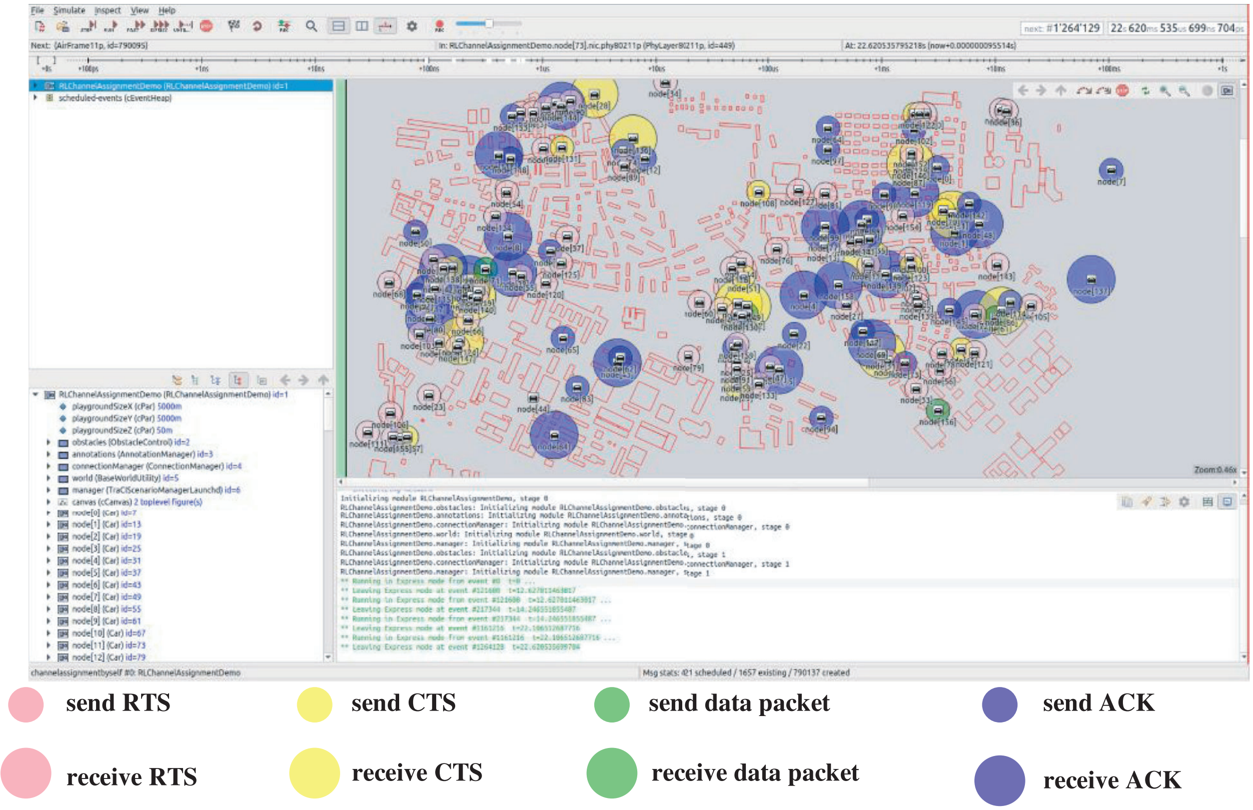


Figure 5: Channel access process

This paper compares the performance of the multi-agent DRL-CDCA with other two existing channel assignment mechanisms as follow:

- **The random assignment mechanism (Random):** Each CNP randomly selects the SCH in each time slot.


File Simulate Inspect View Help
next: #1'264'129
22s 620ms 535us 699ns 704ps
Next: (AirFrame11p, id=790095)
In: RLChannelAssignmentDemo.node[73].nic.phy80211p (PhyLayer80211p, id=449)
At: 22.620535795218s (now+0.000000095514s)
Zoom:0.46x
● send RTS
● send CTS
● send data packet
● send ACK
● receive RTS
● receive CTS
● receive data packet
● receive ACK

```
RLChannelAssignmentDemo (RLChannelAssignmentDemo) id=1
scheduled-events (cEventHeap)
```

```
RLChannelAssignmentDemo (RLChannelAssignmentDemo) id=1
    playgroundSizeX (cPar) 5000m
    playgroundSizeY (cPar) 5000m
    playgroundSizeZ (cPar) 50m
  obstacles (ObstacleControl) id=2
  annotations (AnnotationManager) id=3
  connectionManager (ConnectionManager) id=4
  world (BaseWorldUtility) id=5
  manager (TraCIScenarioManagerLaunchd) id=6
  canvas (cCanvas) 2 toplevel figure(s)
  node[0] (Car) id=7
  node[1] (Car) id=13
  node[2] (Car) id=19
  node[3] (Car) id=25
  node[4] (Car) id=31
  node[5] (Car) id=37
  node[6] (Car) id=43
  node[7] (Car) id=49
  node[8] (Car) id=55
  node[9] (Car) id=61
  node[10] (Car) id=67
  node[11] (Car) id=73
  node[12] (Car) id=79
```

```
Initializing module RLChannelAssignmentDemo, stage 0
RLChannelAssignmentDemo.obstacles: Initializing module RLChannelAssignmentDemo.obstacles, stage 0
RLChannelAssignmentDemo.annotations: Initializing module RLChannelAssignmentDemo.annotations, stage 0
RLChannelAssignmentDemo.connectionManager: Initializing module RLChannelAssignmentDemo.connectionManager, stage 0
RLChannelAssignmentDemo.world: Initializing module RLChannelAssignmentDemo.world, stage 0
RLChannelAssignmentDemo.manager: Initializing module RLChannelAssignmentDemo.manager, stage 0
RLChannelAssignmentDemo.obstacles: Initializing module RLChannelAssignmentDemo.obstacles, stage 1
RLChannelAssignmentDemo.connectionManager: Initializing module RLChannelAssignmentDemo.connectionManager, stage 1
RLChannelAssignmentDemo.manager: Initializing module RLChannelAssignmentDemo.manager, stage 1
** Running in Express mode from event #0  t=0 ...
** Leaving Express mode at event #121600  t=12.627811463817
** Running in Express mode from event #121600  t=12.627811463817 ...
** Leaving Express mode at event #217344  t=14.246551855487
** Running in Express mode from event #217344  t=14.246551855487 ...
** Leaving Express mode at event #1161216  t=22.106512687716
** Running in Express mode from event #1161216  t=22.106512687716 ...
** Leaving Express mode at event #1264128  t=22.620535699704
```

```
channelassignmentbyself #0: RLChannelAssignmentDemo
Msg stats: 421 scheduled / 1657 existing / 790137 created
```

Figure 6: Channel access process

Table 1: Simulation parameters

Parameters	Values
Density ρ	[50, 200, 400] (veh)
Discount Factor γ	0.8
Learning Rate α	0.01
Training Cycle F	100 ms
Update Steps J	20
ξ, ϱ, $\forall\beta_{inj}$	1

- **The greedy selection mechanism (Greedy):** Each CNP selects the SCH that is currently reserved by the least other CNPs.

We use the following metrics to compare the performance of different mechanisms:

- **Packet delivery ratio:** The ratio of the number of ACKs and the number of packets sent in the entire simulation time, representing the adaptability of the channel assignment mechanism to the dynamic network.
- **One hop packet delay:** The average time required for each data sent from the application layer of the source to the application layer of the destination, which is critical for fast data transfer in VANETs.

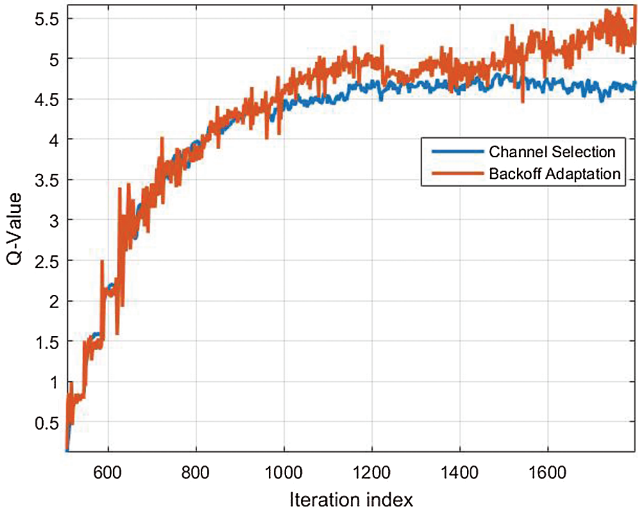
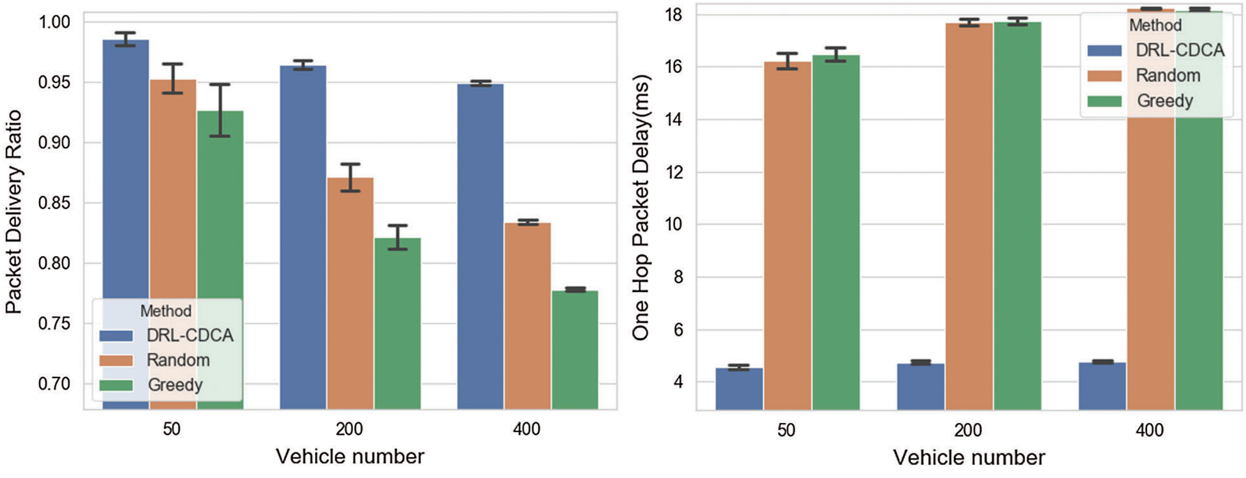


Figure 7: Convergence performance of Multi-agent DRL-CDCA

6.2 Results

Fig. 7 shows the convergence of the multi-agent DRL-CDCA, and it can be seen that the Q-value remains unchanged until 1000 iterations [6–8]. In fact, the channel assignment decisions in VANET are typically highly repetitive, so multi-agent DRL-CDCA can converge quickly.

Fig. 8 evaluates the performance of the DRL-CDCA and other two existing mechanisms with a variable vehicular quantity. It is clear that the performance of all the channel assignment mechanisms become worse as more vehicles appear on the road. Fig. 8(a) plots the change of the packet delivery ratio with different vehicular quantity. In fact, with the increase of the vehicular quantity, the quantity of the data packets also increases, leading to more frequent collisions of data packets. Consequently, the probability of successful transmission gradually decreases. Fig. 8(b) plots the change of the one-hop packet delay with different vehicular quantity. As the communication demand increases, the busy time of the channel also increases, which leads to an increase in the back-off duration of the node, hence the one-hop packet delay



(a) (b)

Figure 8: Simulation results: (a) and (b) compare the performance of different methods in terms of the mean and the standard deviation of packet delivery ratio and one hop packet delay under different vehicle densities. (a) Packet delivery ratio, (b) One hop packet delay

also increases as the number of vehicles increases. Our method can outperform other two conventional baseline mechanisms even in a highly dense situation. For instance, when the vehicle density is set to 400, the packet delivery ratio of the multi-agent DRL-CDCA is 13.83% and 21.98% higher than Random and Greedy, respectively. And the one-hop packet delay of our method is 73.73% and 73.65% lower than that of the other two methods, respectively.

Fig. 9 evaluates the performance of the DRL-CDCA and other two existing mechanisms with a variable vehicle speed. It can be found that the performance of all the channel assignment mechanisms become worse as the vehicular speed increases. Fig. 9(a) plots the change of the packet delivery ratio with different vehicular speeds. As the vehicle speed increases, the network topology of VANET changes rapidly, which easily cause the destination node to leave the signal coverage of the source node or move back off the building. This undoubtedly tends to cause packet reception failures and decrease of the packet delivery ratio. Fig. 9(b) plots the change of the one-hop packet delay with different vehicle speeds. When vehicle speed changes, it has little effect on the "busyness" degree of the channel. Meanwhile, when the node accesses the channel, the random back-off process is almost unaffected by speed change. Therefore, the one-hop packet delay does not change much with changes in speed. The WAVE protocol stack is a wireless communication standard whose design is based on the VANET high-speed mobile environment. Its CCHI and SCHI are both 50 ms. When a neighboring node moves at a speed of 120 km/h, it only moves 1.7 m within 50 ms. Therefore, during the channel coordination and data transmission processes, the VANET topology hardly changes. As a result, the performance of each channel assignment mechanism does not decrease much.

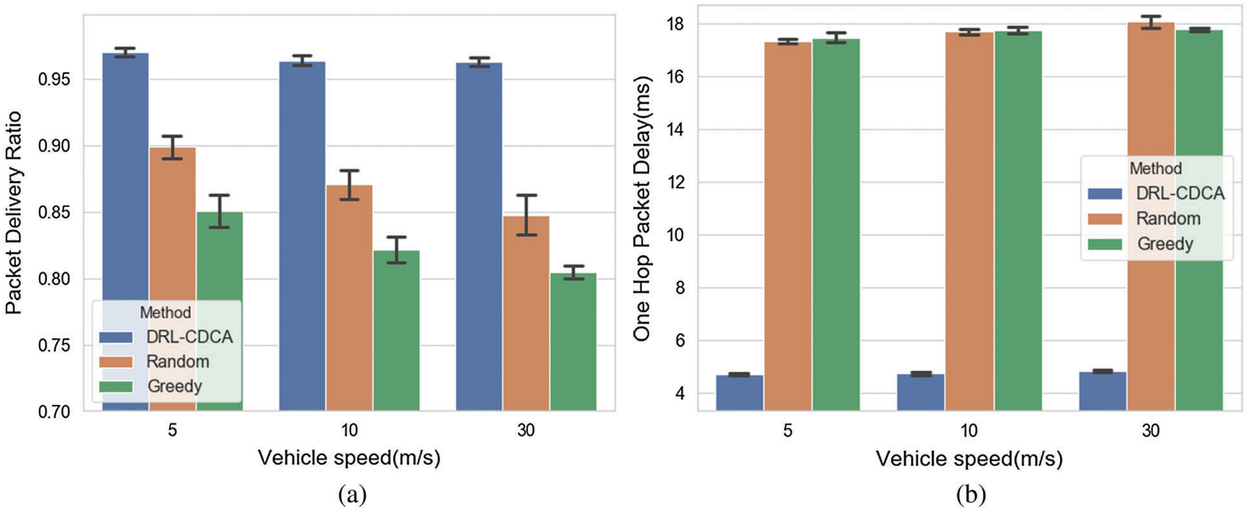


Figure 9: Simulation results: (a) and (b) compare the performance of different methods in terms of the mean and the standard deviation of packet delivery ratio and one hop packet delay under different vehicular speeds. (a) Packet delivery ratio (b) One hop packet delay

To sum up, the figures above confirm the advantage of our proposed method that can achieve higher efficiency performances. Random does not consider the network state, and only selects the SCH in a random manner. It may cause in very few nodes using a certain SCH, resulting in a huge waste of wireless communication resources. And it may also cause a large number of nodes using another certain SCH, resulting in severe data collisions that degrade network performance. Greedy may cause the SCH with the smallest load to become the SCH with the largest load, and make other SCHs underutilized.

Compared with other channel assignment mechanisms, the multi-agent DRL-CDCA is based on the dual Q-networks trained by the past experience including consensus reward to make a collaborative optimization. Obviously, the performance of multi-agent DRL-CDCA is significantly better than other channel assignment mechanisms.

7 Conclusion

In this paper, a dual reinforcement learning (DRL)-based cooperative DCA (DRL-CDCA) mechanism is proposed, which enable the nodes to learn the optimal channel selection and back-off adaptation strategy from past experiences. Then the performances of the proposed mechanisms are compared with two other existing mechanism under the same simulation scenario. The simulation results show that DRL-CDCA improves the overall performance compared with other two conventional baseline mechanism obviously.

Acknowledgement: The author(s) received no specific help from somebody other than the listed authors for this study.

Funding Statement: This research was supported in part by Beijing Municipal Natural Science Foundation Nos. L191001 and 4181002, the National Natural Science Foundation of China under Grant Nos. 61672082 and 61822101, the Newton Advanced Fellow-ship under Grant No. 62061130221.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] O. Mina, O. Zytoune and D. Aboutajdine, “Multi-channel coordination based MAC protocols in vehicular *ad hoc* networks (VANETs): A survey,” in *Advances in Ubiquitous Networking 2*. Singapore: Springer, 2017.

[2] J. H. Nie and S. Haykin, “A dynamic channel assignment policy through Q-learning,” *IEEE Transactions on Neural Networks*, vol. 6, no. 10, pp. 443–1455, 1999.

[3] S. J. Liu, X. Hu and W. D. Wang, “Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems,” *IEEE Access*, vol. 99, no. 6, pp. 15733–15742, 2018.

[4] Y. F. Wei, Z. Y. Wang, D. Guo and F. R. Yu, “Deep Q-learning based computation offloading strategy for mobile edge,” *Computers, Materials & Continua*, vol. 59, no. 1, pp. 89–104, 2019.

[5] T. Ahmed and L. M. Yannick, “A QoS optimization approach in cognitive body area networks for healthcare applications,” *Sensors*, vol. 17, no. 4, pp. 780, 2017.

[6] M. Louta, P. Sarigiannidis, S. Misra, P. Nicopolitidis and G. Papadimitriou, “RLAM: A dynamic and efficient reinforcement learning-based adaptive mapping scheme in mobile WiMAX networks,” *Mobile Information Systems*, vol. 10, no. 2, pp. 173–196, 2014.

[7] J. Su, Z. G. Sheng, L. B. Xie, G. Li and A. X. Liu, “Fast splitting-based tag identification algorithm for anti-collision in UHF RFID system,” *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2527–2538, 2019.

[8] J. Su, Z. G. Sheng, A. X. Liu and Y. R. Chen, “A group-based binary splitting algorithm for UHF RFID anti-collision systems,” *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 998–1012, 2019.

[9] J. Su, Z. G. Sheng, V. C. M. Leung and Y. R. Chen, “Energy efficient tag identification algorithms for RFID: Survey, motivation and new design,” *IEEE Wireless Communication*, vol. 26, no. 3, pp. 118–124, 2019.

[10] A. A. Almohammedi, N. K. Noordin, A. Sali, F. Hashim and M. Balfaqih, “An adaptive multi-channel assignment and coordination scheme for IEEE 802.11P/1609.4 in vehicular ad-hoc networks,” *IEEE Access*, vol. 1, no. 1, pp. 99, 2018.

[11] S. Chantaraskul, K. Chaitien, A. Nirapai and C. Tanwongvarl, “Safety communication based adaptive multi-channel assignment for VANETs,” *Wireless Personal Communications*, vol. 94, no. 1, pp. 83–98, 2017.

[12] X. H. Li, B. J. Hu, H. B. Chen, G. Andrieux and Z. H. Wei, "An RSU-coordinated synchronous multi-channel MAC scheme for vehicular *ad hoc* networks," *IEEE Access*, vol. 3, pp. 2794–2802, 2015.

[13] A. Ribal, C. Assi and M. Khabbaz, "Deep reinforcement learning-based scheduling for roadside communication networks," in *15th Int. Sym. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks IEEE*, Paris, France, 2017.

[14] Y. Liu, Z. Yang, X. Y. Yan, G. C. Liu and B. Hu, "A novel multi-hop algorithm for wireless network with unevenly distributed nodes," *Computers, Materials & Continua*, vol. 58, no. 1, pp. 79–100, 2019.

[15] X. L. Wang, J. M. Jiang, S. J. Zhao and L. Bai, "A fair blind signature scheme to revoke malicious vehicles in VANETs," *Computers, Materials & Continua*, vol. 58, no. 1, pp. 249–262, 2019.

[16] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054, 1998.