

Article

Deep Reinforcement Learning-Based Spectrum Allocation Algorithm in Internet of Vehicles Discriminating Services

Zheng Guan, Yuyang Wang  and Min He * 

School of Information Science and Engineering, Yunnan University, Kunming 650500, China; gz_627@sina.com (Z.G.); wyy925@mail.ynu.edu.cn (Y.W.)

* Correspondence: hemin@ynu.edu.cn; Tel.: +86-137-0844-1034

Abstract: With the rapid development of global automotive industry intelligence and networking, the Internet of Vehicles (IoV) service, as a key communication technology, has been faced with an increasing spectrum of resources shortage. In this paper, we consider a spectrum utilization problem, in which a number of co-existing cellular users (CUs) and prioritized device-to-device (D2D) users are equipped in a single antenna vehicle-mounted communication network. To ensure a business-aware spectrum access mechanism with delay granted in a complex dynamic environment, we consider optimizing a metric that maintains a trade off between maximizing the total capacity of vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) links and minimizing the interference of high priority links. A low complexity priority-based spectrum allocation scheme based on the deep reinforcement learning method is developed to solve the proposed formulation. We trained our algorithm using the deep Q-learning network (DQN) over a set of public bandwidths. Simulation results show that the proposed scheme can allocate spectrum resources quickly and effectively in a high dynamic vehicle network environment. Concerning improved channel transmission rate, the V2V link rate in this scheme is 2.54 times that of the traditional random spectrum allocation scheme, and the V2I link rate is 13.5% higher than that of the traditional random spectrum allocation scheme. The average total interference received by priority links decreased by 14.2 dB compared to common links, realized service priority distinction and has good robustness to communication noise.

Keywords: Internet of vehicles; spectrum allocation; reinforcement learning; neural network; priority



Citation: Guan, Z.; Wang, Y.; He, M. Deep Reinforcement Learning-Based Spectrum Allocation Algorithm in Internet of Vehicles Discriminating Services. *Appl. Sci.* **2022**, *12*, 1764. <https://doi.org/10.3390/app12031764>

Academic Editors: Gianni Pantaleo and Pierfrancesco Bellini

Received: 25 November 2021

Accepted: 23 January 2022

Published: 8 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, with the development of wireless communication and intelligent vehicle technology, the on-board network has attracted extensive attention from the industry and academia [1–4]. Vehicle to everything (V2X) includes vehicle to vehicle (V2V), vehicle to infrastructure (V2I), vehicle to pedestrian (V2P) and vehicle to network (V2N). The third generation partnership project (3GPP) already supports V2X services in long term evolution (LTE) and fifth generation mobile communication technology (5G) networks; through the wireless communication network formed by the vehicle and various nodes, real-time information such as driving assistance and accident avoidance can be transmitted, and data services such as on-board entertainment, real-time navigation and Internet access can be provided to provide people with a safer, more efficient, environmentally friendly and comfortable driving environment. Spectrum resources are limited natural resources. With the wide application of various radio technologies and services, the demand for spectrum resources in various industries and fields of the national economy keeps increasing. The development of new generation of information technologies such as mobile Internet and Internet of Things also puts forward new demands for spectrum resources. Therefore, spectrum resources are increasingly scarce and are wasted if not fully utilized or used improperly. The rational allocation of spectrum is the key to achieving high quality vehicle network communication. With the rapid and widespread development of advanced broadband wireless technologies and the increasing demand for high speed and quality services,

traditional static spectrum allocation policies are becoming obsolete. Dynamic spectrum sharing, as one of the key technologies to solve the problem of insufficient spectrum utilization, has received a lot of attention and research in recent years [5] to alleviate the current situation of insufficient spectrum resource utilization. The main goal of dynamic spectrum allocation is to design a flexible spectrum allocation strategy between existing users and new users without compromising the utilization of spectrum resources by existing users, so as to effectively allocate the idle spectrum to new users, so as to improve the efficiency of spectrum utilization [6]. However, with the continuous expansion of the application scope of the Internet of vehicles and the improved communication performance requirements in the network, it is necessary to design an effective spectrum resource allocation scheme to ensure the Internet of vehicles communication services with high reliability and low delay.

Vehicle nodes in vehicle-mounted networks have high mobility and complex time-varying characteristics. It is very challenging to provide high-quality services for vehicles, such as super-large capacity, ultra-high reliability and low delay. In order to solve these problems, it is necessary to provide an efficient spectrum resource allocation method for vehicles in V2X scenarios. The authors in [7] propose a spectrum resource allocation scheme that can adapt to the slow-changing large-scale channel fading, and maximizes the capacity of V2I by using the slow-fading statistical characteristics of channel status information (CSI). In [8], quality of service (QoS) requirements for different connections are different. Under the condition of ensuring V2V link reliability and waiting time constraints, the total traversal capacity of the V2I link is maximized to reduce network signaling overhead. In particular, this scheme allows spectrum resources to be shared not only between V2I and V2V links, but also between different V2V links. In [9], the authors propose a method to convert the actual delay and reliability requirements of V2V communication into optimization constraints, but the optimization constraints can only be calculated by slowly changing CSI. In [10], in areas where spectrum resources are in short supply, vehicle horizontal sharing is combined with available spectrum resources to improve the success rate of alarm information transmission, reduce transmission delay of alarm information and reduce noise to reduce the incidence of traffic accidents, making a great contribution to improving traffic safety. In [11], a spectrum allocation scheme oriented to service priority in the Internet of vehicles based on long-term evolution was proposed. The spectrum allocation problem was modeled as a mixed integer programming problem and solved by the immune cloning-based algorithm to maximize the system utility. The authors in [12] propose a spectrum allocation model based on pricing and auction. While the fairness of secondary users is effectively guaranteed, idle spectrum information needs to be obtained in advance, which requires high spectrum perception ability. The authors in [13] adopt the spectrum sharing model driven by the spectrum database to dynamically adjust the protection boundary of major users according to the geographical location database to improve spectrum utilization efficiency, but this has high requirements for real-time updating of spectrum information. In [14], the dynamic spectrum allocation problem is mapped to a graph coloring model by using topological graph relations, and the interference graph is used to reduce the interference caused by spectrum sharing. However, as long as the topological structure changes, re-mapping calculation is required, which is mostly suitable for a static network environment. However, in the future vehicular network with a high dynamic unknown environment, the traditional spectrum resource allocation scheme is often difficult to achieve.

Machine learning, as one of the powerful artificial intelligence tools, has been widely used in wireless communication networks in recent years, such as multiple input–multiple output (MIMO), D2D, heterogeneous network composed of femtocells and small cells, etc. [15]. In particular, reinforcement learning (RL), as a kind of machine learning, has the ability of adaptive adjustment and does not require real-time spectrum perception. Agents only need to observe the changes of environmental state and improve performance according to the reward feedback received after learning to take action [16], which greatly reduces the complexity of spectrum sharing. Moreover, they do not need the real-time spectrum perception—knowledge concerning only the status of environmental change—in learning

to take action after, accordingly as they receive feedback to improve performance, greatly reducing the complexity of spectrum sharing. This has achieved great success in many applications such as AlphaGo [17]. Inspired by its excellent performance, researchers began to use reinforcement learning methods to solve the spectrum resource allocation problem in the unknown high dynamic vehicle-mounted network environment. A spectrum allocation scheme based on distributed learning is proposed in [18], in which D2D users explore the environment and autonomously select spectrum resources to maximize throughput and spectrum efficiency and at the same time to meet the minimum interference caused to cellular users. In [19], a distributed spectrum resource allocation scheme based on multi-agent RL is proposed. Each V2V link is regarded as an agent, and each agent autonomously learns how to rationally select spectrum and power to improve the total capacity of the V2I link and the payload transmission rate of the V2V link. In [20], each vehicle is regarded as an agent, and multiple agents make decisions autonomously based on local observations in V2V broadcast communication to find the available spectrum. In [21], a deep RL method is developed to enable BS to centrally manage network, cache and computing resources. In [22], BS is used to summarize and compress vehicle observation data, and then the compressed information is fed back and the reinforcement learning process is carried out at the base station to improve the spectrum sharing decision performance in the network. In [23], the graph neural network (GNN) is used to build the V2X network; GNN extracts the features of each V2V pair. Based on the extracted features and local observations, the V2V pair can use the Q-network to make distributed decision-making. The authors in [24] propose a V2V communication wireless resource allocation system based on proximal policy optimization. In this radio resource allocation framework, continuous actions and multi-dimensional actions can be output to reduce the implementation complexity of large-scale communication scenarios. In [25], the DQN network is improved. Aiming at the non-stationarity problem caused by multi-agent parallel learning, lag Q-learning and parallel experience replay trajectory are introduced to stabilize the training process, and approximate regret reward (ARR) is added to stabilize the reward estimation. In order to improve the adaptability of the traditional deep reinforcement learning (DRL) algorithm in a dynamic environment, [26] further combines meta-learning with DRL and proposes a meta-based DRL algorithm. Compared with the DQN-based algorithm, our DRL-based algorithm can provide better performance on both V2I and V2V links. In addition, the DRL algorithm training strategy proposed in this paper has good generalization ability and can quickly adapt to the new environment with limited experience. The authors in [27] propose a centralized dynamic channel allocation method based on deep reinforcement learning for satellite Internet of Things. This method makes use of the strong representation ability of the deep neural network to make intelligent allocation decisions through continuous learning of allocation strategies so as to minimize the average transmission delay of all sensors. However, in the above methods, differentiated service design is carried out according to vehicle type or link service characteristics. However, in real life, special vehicles can be seen everywhere on the road, such as police cars, ambulances, fire trucks and so on; they need a better information transmission environment in the Internet of vehicles. Faced with these special vehicles with urgent business needs, compared with all V2V links in existing literature that compete fairly for spectrum resources, this paper proposes a V2V link and V2I link sharing strategy for the scenario that urgent services need to be handled first in the Internet of vehicles. Mode 4 in cellular V2X architecture is used for resource allocation, and vehicles share resource pools for communication between V2V and V2I. In this strategy, the link priority mechanism is introduced, and the higher priority link can get a better information interaction environment by reinforcing the reward design of learning.

The innovation points of this paper are summarized as follows:

- (1) Formulate the dynamic spectrum allocation problem in a CU and D2D co-existed vehicle network.
- (2) Develop a centralized low complexity algorithm based on the deep reinforcement learning method to achieve priority-based spectrum allocation.
- (3) Build a weighted sum reward function to realize the dynamically adaptive rates—interference between V2I and V2V links.

The simulation results show that the proposed control method can effectively improve the service quality of high-priority links while ensuring the overall performance of the system, and has good robustness to communication noise.

The rest of this paper is organized as follows. We depict the system model and formulate the optimization problem in Section 2. The proposed RL-based algorithm is presented in Section 3. Section 4 demonstrates simulation results and Section 5 concludes this paper.

2. System Model and Problem Formulations

On a V2X network, each V2V link can independently select a different channel to maximize the transmission rate. However, the global performance is poor due to interference between different V2V links. On the other side, considering the V2X scenarios, BS has enough computing and storage resources and can achieve the efficient allocation of resources. With the help of reinforcement learning, this paper uses BS as an agent to interact with the unknown vehicle-mounted network environment.

Suppose there are CU and D2D users co-existing in a vehicle-mounted communication network, where each device is equipped with a single antenna. In this paper, the sets of CU and D2D users are, respectively, expressed as $I = \{1, 2, \dots, i, \dots, I\}$ and $J = \{1, 2, \dots, j, \dots, J\}$. Each CU establishes V2I links with BS to support high-quality services, and each D2D user pair transmits information by establishing V2V links. In order to ensure high quality V2I link communication, it is assumed that each V2I link has been pre-assigned different orthogonal spectral subcarriers to eliminate the interference between V2I links in the network. Without sacrificing performance, V2V links and V2I links share the same spectrum resources. To improve the communication quality of V2V links, each V2V link needs to select its occupied spectrum subcarriers and transmitted power.

The channel power gain of the V2I link established between the i -th CU user and BS on authorized channel i is defined as $G_i[i]$. $H_{j,B}[i]$ Represents the interference channel gain from the vehicle transmitter of the i -th V2I link to the vehicle receiver of the j -th V2V link occupying the i -th subcarrier. P_i^c and P_j^d denote the transmitting power of the i -th V2I link vehicle transmitter and the j -th V2V link vehicle transmitter, respectively; σ^2 represents the noise power. $\rho_j^i = \{0, 1\}$ represents the spectrum allocation scheme, if the j -th V2V link chooses the i -th channel, $\rho_j^i = 1$; otherwise $\rho_j^i = 0$.

In this case, the reception signal-to-noise ratio (SINR) of V2I link i using the i -th subcarrier can be expressed as follows:

$$\gamma_i^c[i] = \frac{P_i^c G_i[i]}{\sum_{j=1}^J \rho_j^i P_j^d H_{j,B}[i] + \sigma^2}. \tag{1}$$

Assume that each V2V link occupies only one channel. According to the Shannon formula, the transmission rate of the i -th V2I link using the i -th channel can be expressed as:

$$C_i^c[i] = W \log_2(1 + \gamma_i^c[i]), \tag{2}$$

where W is the bandwidth of each channel.

Similarly, $H_j[i]$ represents the interference channel gain of the j -th V2V link occupying the i -th subcarrier. $H_{k,j}[i]$ denotes the interference channel gain from the j -th V2V link vehicle transmitter to the j -th V2V link vehicle receiver on the i -th channel. $G_{i,j}[i]$ indicates the interference channel gain of the i -th V2I link vehicle transmitter to the j -th V2V link receiver on the i -th subcarrier. In summary, the SINR of the V2V link j occupying the i -th subchannel can be represented as:

$$\gamma_j^d[i] = \frac{\rho_j[i]P_j^d H_j[i]}{I_j[i] + \sigma^2}, \tag{3}$$

where $I_j[i]$ denotes the interference power received by the j -th V2V link from other V2V links and from all V2I links:

$$I_j[i] = \sum_{k \neq j}^J \rho_k[i]P_k^d H_{k,j}[i] + P_i^c G_{i,j}[i], \tag{4}$$

$\sum_{k \neq j}^J \rho_k[i]P_k^d H_{k,j}[i]$ indicates the interference power between V2V links; $P_i^c G_{i,j}[i]$ represents V2I link interference power. Therefore, the transmission rate of the j -th V2V link on the i -th channel can be expressed as:

$$C_j^d[i] = W \log_2(1 + \gamma_j^d[i]). \tag{5}$$

To account for overall link performance and high-priority link interference requirements, we maximize an objective function that is a weighted sum of two terms and subtract one term. The first term is the sum rate of V2I links, the second term is the sum rate of V2V links and the third term is the total interference of priority links. Meanwhile, to reflect the advantages of high-priority links without affecting the performance of other links, we introduce some constraints.

Therefore, the overall optimize problem is:

$$\begin{aligned} \text{Maximize : } & \lambda_1 \sum_i C_i^c[i] + \lambda_2 \sum_j C_j^d[j] - \lambda_3 I_f, \\ \text{Subject to : } & I_f \leq I_{\max} \\ & C_i^c > C_{RA}^c \\ & C_j^d > C_{RA}^d \end{aligned} \tag{6}$$

where λ_1, λ_2 and λ_3 are the weight constant used to define the priority between the three targets, I_f refers to the total interference received by a priority link, I_{\max} represents the maximum total interference that we set. C_{RA}^c and C_{RA}^d , respectively represent the rates of V2I link and V2V link when the random spectrum allocation scheme is used.

3. RL-Based Resource Allocation Algorithm

As shown in Figure 1, based on the network model in [22], the first V2V link is set as a priority link, and the other V2V links are set as common links. Deep neural network (DNN) is designed to compress the local information observed by each V2V; this information includes its own channel power gain, interference from other links and the transmitted power of its own vehicle transmitter. The compressed information is then fed back to BS. Based on the feedback from all V2V links, BS will use RL to make an optimal decision on spectrum allocation for all V2V links. Finally, the BS sends the decision information to each V2V link.

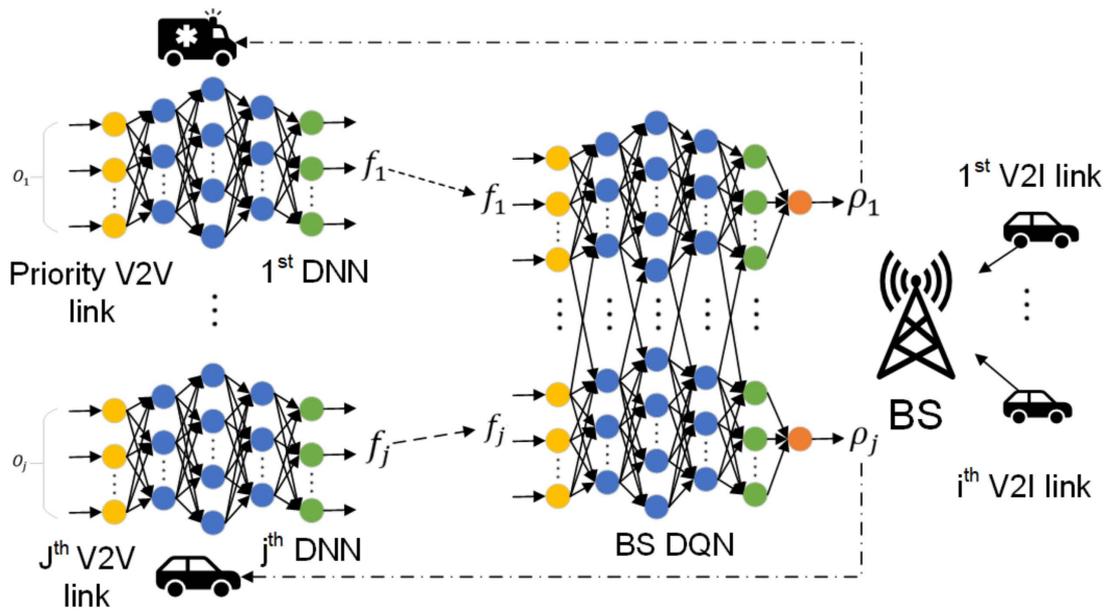


Figure 1. Schematic diagram of neural network structure.

3.1. Link Data Compression

DNN compresses the local observation data of each V2V link. Each V2V link first observes the surrounding environment to acquire its transmitting power P_j^d and the combined interference power from other links $I_j = (I_j[1], \dots, I_j[i], \dots, I_j[I])$. Considering the impact of V2V links on V2I links, the observation data of the j -th V2V link must also include the interference channel gain from the V2V link to all V2I links which are denoted as $H_{j,B} = (H_{j,B}[1], \dots, H_{j,B}[i], \dots, H_{j,B}[I])$; it can be estimated at BS and broadcast to all V2V links within the coverage area of BS, resulting in low signaling overhead [10]. In [22], it is assumed that the power gain of the current channel can be obtained by delayless feedback at the vehicle transmitter on the j -th V2V link, which is represented as $H_j = (H_j[1], \dots, H_j[i], \dots, H_j[I])$.

In this case, the observed data of the j -th V2V link can be written as:

$$O_j = \{P_j^d, I_j, H_{j,B}, H_j\}. \tag{7}$$

The observed data O_j is compressed using DNN on each V2V link; the compressed data f_j output by DNN is fed back to DQN at BS. Here, $f_j = \{f_{j,k}\}$ is also known as the feedback vector of the j -th V2V link, where $f_{j,k}, \forall k \in \{1, 2, \dots, N_j\}$ refers to the k -th feedback element of the j -th V2V link, N_j represents the feedback number learned by the j -th V2V link. f_j will also serve as input to DQN for the reinforcement learning process.

3.2. Link Configuration Decision

Reinforcement learning can maximize long-term returns in sequential decision-making problems, and enable agents to interact with complex environments and seek optimal spectrum allocation strategies through continuous trial and error. In [28], a kind of deep Q network is proposed. With the help of the end to end reinforcement learning method, an excellent policy can be directly learned from the high-dimensional output, enabling agents to solve a series of challenging tasks.

This paper treats BS as an Agent in RL. The agent seeks a spectrum selection strategy with maximum cumulative reward by trial and error through continuous interaction with a complex unknown environment. RL can be modeled as a Markov decision process (MDP).

As shown in Figure 2, the specific process can be divided into three steps:

- (1) At the time step t , select the action A_t to be performed according to the current state S_t .
- (2) Select the state S_{t+1} after the transfer based on the current state S_t and action A_t .
- (3) According to the action A_t taken in the current state S_t gives the corresponding reward R_{t+1} .

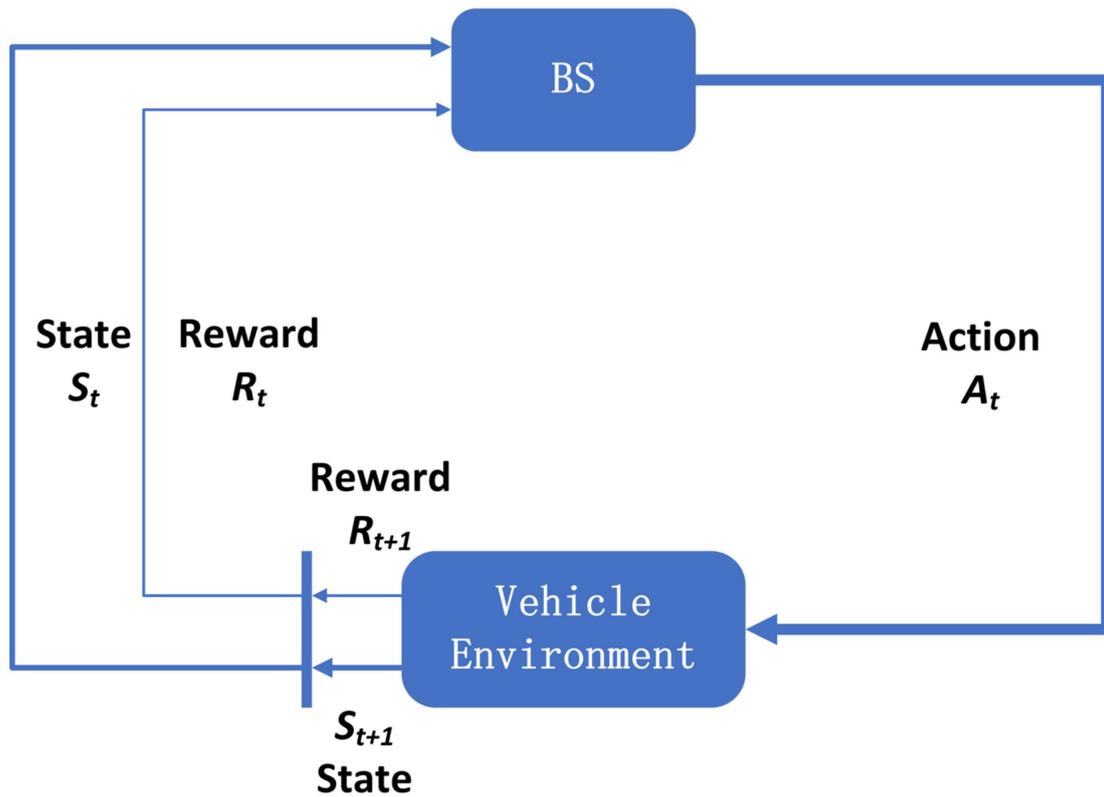


Figure 2. The interaction between an agent and the environment in vehicle network.

3.2.1. State Space

BS takes all compressed data fed back by DNN as the current state, which can be expressed as:

$$S = \{f_1, \dots, f_j, \dots, f_J\}. \tag{8}$$

3.2.2. Action Space

The spectrum allocation scheme of all V2V links will be decided at BS using DQN. So the action is defined as:

$$A = \{\rho_1, \dots, \rho_j, \dots, \rho_J\}, \tag{9}$$

where $\rho_j = \{\rho_j[i]\}, \forall i \in I$ represents the spectrum allocation scheme.

3.2.3. Reward Design

The ultimate goal of this paper is to maximize the long-term sum rate of V2V links, ensure the QoS of V2I links in V2X scenarios and ensure the transmission performance of priority links. V2V links are generally used to transfer key information during vehicle running, such as vehicle speed, location, driving direction and braking. V2I communication

is mainly used for real-time information services, vehicle monitoring and management, and charging without parking. Therefore, the primary goal should be to ensure V2V link transmission, while ensuring that V2V transmission should not cause too much influence on V2I links, and priority links in the network of vehicles should be interfered with less. To achieve this goal, this paper assumes that the first V2V link is a priority link, and the reward of RL can be designed as follows:

$$R = \lambda_c \sum_{i=1}^I C_i^c [i] + \lambda_d \sum_{j=1}^J C_j^d - \lambda_f I_1, \tag{10}$$

where λ_c , λ_d and λ_f , respectively, correspond to the non-negative weight of the total rate of V2I links, total rate of V2V links and total interference of priority links.

The optimization goal of the reinforcement learning algorithm is to find an optimal strategy $\pi(a, s)$ to maximize the cumulative reward value returned by the training set. Strategy $\pi(a, s)$ refers to the probability distribution of action a given state s . In reinforcement learning, there is always an optimal strategy $\pi^*(a, s)$ that maximizes the expected reward. The expected reward can be expressed as $R_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$, where γ is the attenuation of future rewards.

In this paper, Q learning is used to solve RL problems because it has model-independent properties where for $P(s', r|s, a)$ no prior is required. Concerning Q learning based on a given strategy π

$$Q^\pi(s, a) = E_\pi[R_t | S_t = s, A_t = a]$$

So the agent takes action a in state S and gets a reward based on probability π . The optimal action value function $Q^*(s, a)$ under the optimal strategy π^* satisfies the Bellman equation, which can be approximated by the iterative updating method:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)], \tag{11}$$

where $0 < \alpha < 1$ represents the learning rate, which determines how much of the error is to be learned.

In [28], the Q -table was turned into a network model to obtain a Q value. In the past, a Q value needed to be queried in a Q -table, but now we only need to input the state and action to obtain the corresponding Q value.

DQN uses the ϵ -greedy strategy to store the transfer samples (S_t, A_t, R_t, S_{t+1}) obtained by the interaction between each time-stepping agent and the environment in the empirical memory unit. At each slot, according to the behavioral strategy and feedback from the environment, a set of sample data D of the agent is stored in the memory bank every training session, and network parameters θ are updated with random gradient descent variables to minimize the squared error:

$$L = \sum_{t \in D} [R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \theta^-) - Q(S_t, A_t; \theta)]^2, \tag{12}$$

where θ^- represents the parameter of the target Q -network, which is updated synchronously with Q -network parameter θ every time step. Meanwhile, in order to further improve the stability of DQN, DQN is updated several times. During training, a certain number of data are randomly extracted from the empirical memory unit for training. So you can constantly optimize the network model.

3.3. Algorithm Flow

In this paper, the observed value of the j -th V2V link in the time step $t \in \{1, 2, \dots, T\}$ is defined as O_j^t . At this time, the observed values of all V2V links at the time step t can be expressed as $O_t = \{O_j^t\}, \forall j \in J$. According to literature [16], the approximate target value of return can be expressed as:

$$y_t = R_{t+1} + \gamma \max_a Q(O_{t+1}, a; \theta^-), \quad (13)$$

Then the update process of DQN located at BS can be expressed as:

$$\theta \leftarrow \theta + \beta \sum_{t \in D} \frac{\partial Q(O_t, a_t; \theta)}{\partial \theta} [y_t - Q(O_t, a_t; \theta)], \quad (14)$$

where β refers to the time step length in the strategy gradient iteration.

At each episode t , each V2V link observes the local data O_j^t first, then uses it as input to DNN to get feedback f_j^t , which is then transmitted to BS. The BS will then serve $\{f_j^t\}$ as the input to the DQN to generate the decision a_t , which will be broadcast to all V2Vs. Finally, each V2V link selects its own spectrum according to the decision result.

4. Simulations

This paper designs a simulator according to the evaluation method defined for urban cases in Annex A of 3GPP TR 36.885 [29], which describes in detail the vehicle fading model, density, speed, direction of movement, vehicle passage, V2V data flow, etc. The simulation considers the topology scene of the Internet of vehicles in the two-way and one-way lane area of 375 m wide and 649 m long at the intersection. There is a BS in the center of the scene, and the starting position and driving direction of vehicles are randomly initialized within the region. Other simulation parameters of the system are shown in Table 1. The hardware environment of the simulator is Intel Core I9-10900F processor + 32G memory + Nvidia GeForce RTX3090 graphics card. Tensorflow 1.12.0 and Keras 2.24 are used to build and train the neural network.

Table 1. Simulation parameters.

Parameters	Values
Number of V2I links	4
Number of V2V links	4
Carrier frequency	2 GHz
Normalized channel bandwidth	1
BS antenna height	25 m
BS antenna gain	8 dBi
BS receiver noise factor	5 dB
Vehicle antenna height	1.5 m
Vehicle antenna gain	3 dBi
Vehicle receiver noise factor	9 dB
Vehicle speed	10~15 km/h
Vehicle fading and movement models	Urban case of A.1.2 in [29]
V2I transmit power	23 dBm
V2V transmit power	10 dBm
V2I and V2V link models	Table 2 in [22]

The above algorithm adopts a five-layer fully connected neural network. According to literature [22], the hidden layers of DNN and DQN are set as 3 in this paper. The number of neurons in the three hidden layers of DNN is set to 16, 32 and 16, respectively. The number of neurons in the three hidden layers of DQN was set to 1200, 800 and 600, respectively.

Here, the RELU activation function $f(x) = \max(0, x)$ is used for both DNN and DQN, and the linear function is set to the activation function for the output layer in DNN and DQN. In addition, the RMSProp optimizer [30] was used for renewal network parameters, and the study rate was 0.001. The loss function is set as Huber loss [31]. In addition, the exploration rate of the whole neural network was set as linear decay from 1 to 0.01 during training. The number of steps T of each training set is set to 1000, and the update frequency of the target network Q is set to 500 steps. The discount rate γ for training is set to 0.05.

The size of the empirical memory unit is set to be 1×10^6 , and the size of the small sample data is set to 512.

In order to facilitate performance comparison, the corresponding weights of total V2I link rates and total V2V link rates are set to $\lambda_c = 0.1$, $\lambda_d = 1$ in accordance with reference [22] in the experiment. This algorithm introduces interference weight of priority link λ_f to distinguish link priorities. If λ_f is too small, it cannot reflect the superiority of the high-priority link. In contrast, high-priority links overoccupy common link resources, affecting fairness and significantly degrading overall performance. Through the experimental test from Figure 3, the selection $\lambda_f = 8 \times 10^5$ can distinguish the priority well and take into account the overall performance of the system.

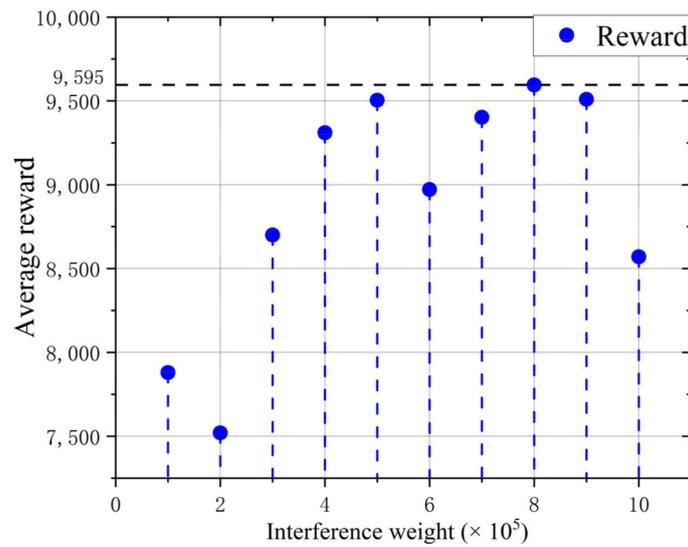


Figure 3. Influence of interference weight on reward.

In Figure 4, the reward for each episode increases as the training set increases, and eventually converges. However, the reason why rewards occasionally get smaller on the way is that greedy strategies are adopted in the training, which may lead to poor results in the exploration of unknown environments. However, as the exploration rate decreases in the later training period, the rewards per episode are not particularly bad and tend to converge, indicating the stability of the training process.

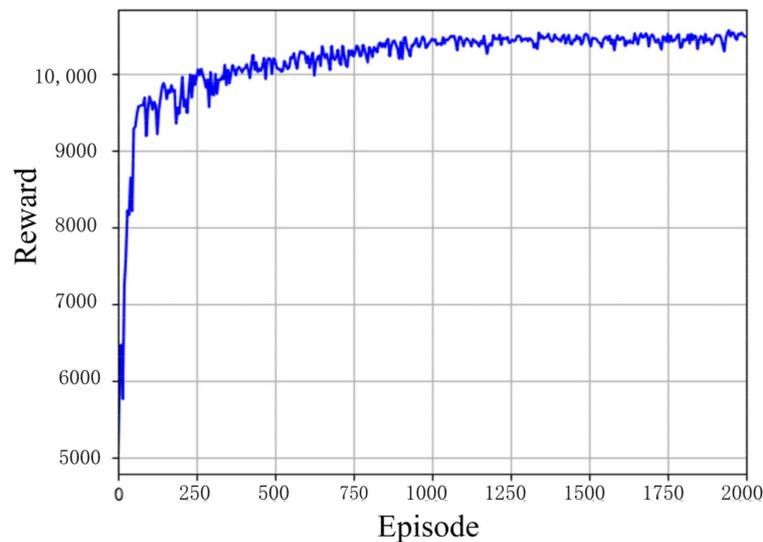


Figure 4. Reward per episode.

5. Discussion

The following four comparison schemes are considered in the simulation:

- Scheme of this paper;
- C-Decision [22];
- SOLEN [9];
- GNN-RL [23];
- Random action.

Firstly, the influence of feedback number on V2V link is considered.

As shown in Figure 5, when the number of feedback is 3, the average total interference received by the high-priority link in this scheme is the smallest, obviously reflecting the advantage of priority, and the average total interference received by the ordinary link is also slightly lower than that received by C-Decision. Priority discrimination is implemented on the premise that the average performance (that is, the overall performance) does not degrade, or priority discrimination ensures fairness. Then, with the increase of the feedback number, the average total interference of each V2V link basically remains unchanged and is larger than the total interference of the feedback number 3. In summary, the number of feedback is set as 3 in the simulation.

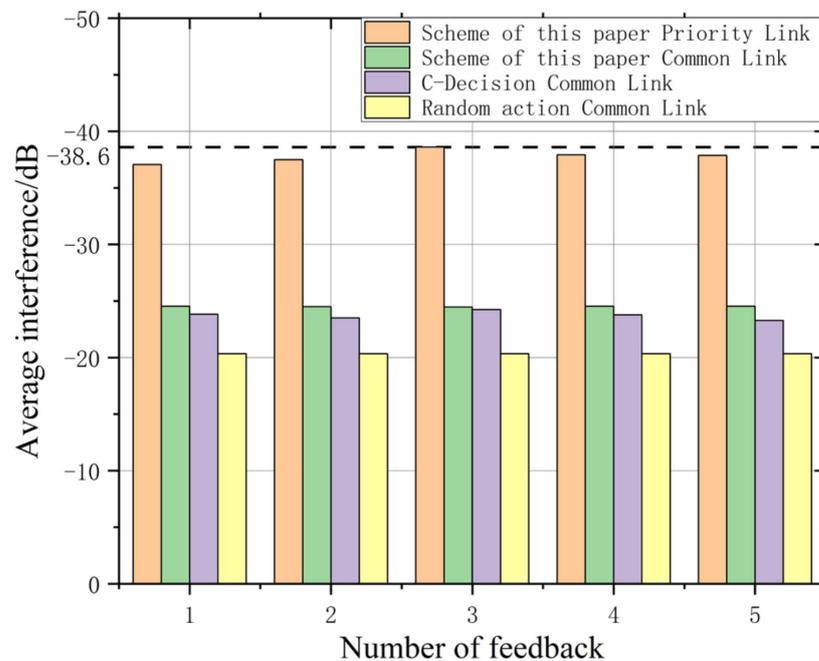


Figure 5. Impact of feedback number on V2V links.

Table 2 lists the performance comparison of the four schemes when the number of feedback is 3 and there is no input noise and feedback noise. Input noise refers to the Gaussian white noise received by each V2V during local observation, and feedback noise refers to the noise generated when DNN's output is fed back to DQN's input. Because the last four schemes consider the links equally, they do not distinguish the priority links, so we treat the average rate as the priority link rate. For the scheme in this paper, the average total interference received by the high-priority link is significantly smaller than that received by the common link, that is, the priority link has a better information transmission environment, reflecting the advantage of priority. The average total interference received by the ordinary link is slightly lower than that of C-Decision, SOLEN scheme and GNN-RL scheme and better than that of random scheme. In addition, the average rate of the V2V link in this scheme is improved by about 2.12% compared with C-Decision, which is higher than the SOLEN scheme and obviously better than the GNN-RL scheme and random action scheme. Similarly, the average rate of the V2I link in this scheme increases by about

7.51% compared with C-Decision, slightly higher than the SOLEN scheme and higher than the GNN-RL scheme and random action scheme. In summary, the scheme presented in this paper obviously shows the advantage of priority and is superior to C-Decision, GNN-RL and random spectrum allocation schemes in performance. While the proposed scheme is only slightly better than the SOLEN scheme, the proposed scheme does not need to obtain the CSI of all links in the base station. The scheme in this paper has no such limitation. It indicates that the scheme in this paper is more suitable for the Internet of vehicles environment with urgent business needs in real life.

Table 2. Performance of four schemes.

	Scheme of This Paper	C-Decision	SOLEN	GNN-RL	Random Action
Average total interference received by a high-priority link	−38.6 dB	−23.5 dB	−24.2 dB	−24.3 dB	−20.3 dB
Average total interference received by a common link	−24.4 dB	−23.5 dB	−24.2 dB	−24.3 dB	−20.3 dB
Average rate of V2V links	2023 bps	1981 bps	2012 bps	1726 bps	794 bps
Average rate of V2I links	7205 bps	6702 bps	7108 bps	6577 bps	6348 bps

Finally, we study the influence of input noise and feedback noise on priority-enabled schemes.

As can be seen from Figure 6, with the increase of input noise, the average interference received by high-priority V2V links in this scheme rises gradually and tends to be stable when the input noise is greater than 10 dB, while the average total interference received by non-priority links basically remains unchanged. This indicates that the scheme with priority enabled when the input noise is low can obviously show the advantage of priority, and the advantage of priority link gradually decreases with the increase of input noise.

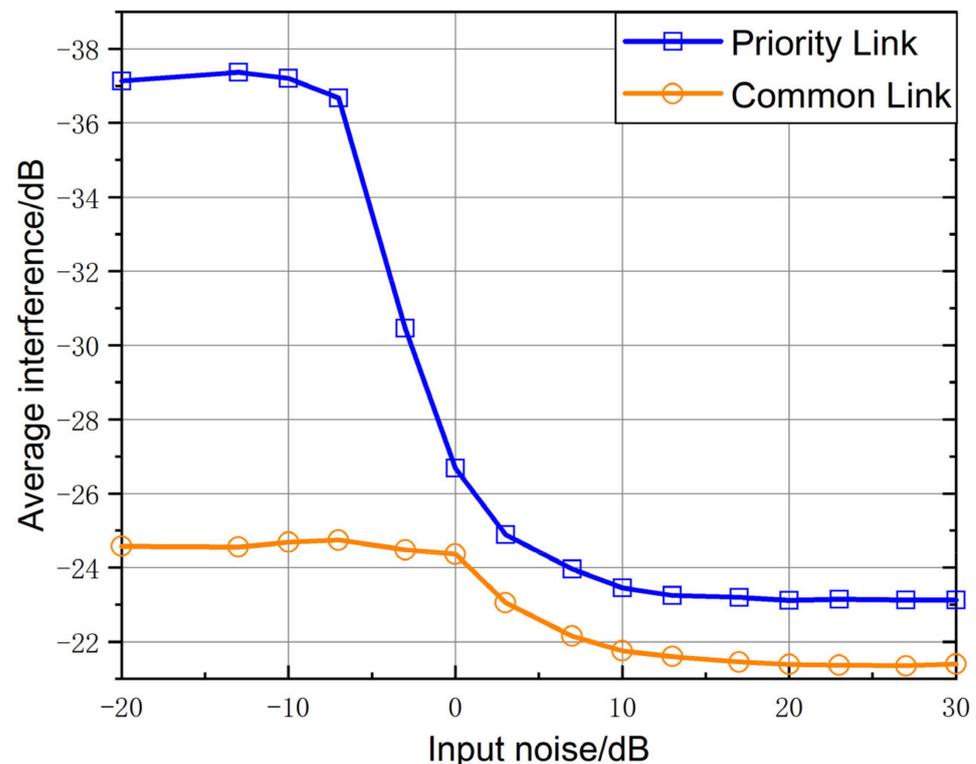


Figure 6. Impact of input noise on V2V links.

As shown in Figure 7, in this scheme, the average rate of the V2I link decreases with the increase of input noise, and tends to be stable when the input noise is greater than 3 dB. When the input noise is large, the rate of the V2I link can still maintain 56.78%. Similarly, the average rate of V2V links decreases with the increase of input noise, and remains stable when the input noise is greater than 10 dB. When the input noise is large, the average rate of V2V links remains 82.04%. V2I links maintain better performance than V2V links.

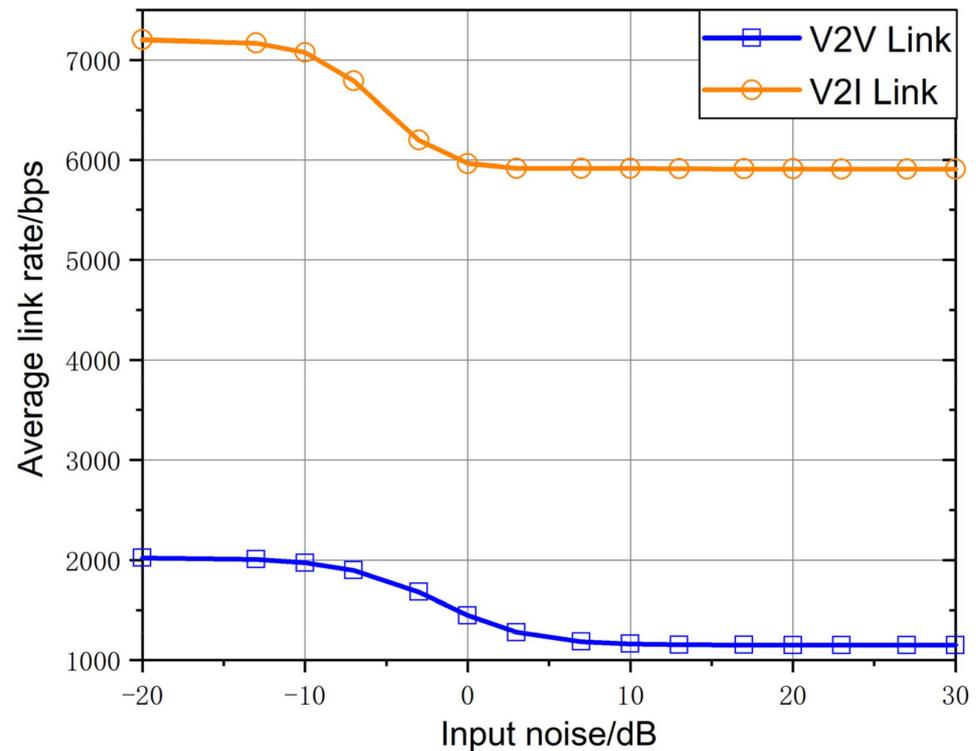


Figure 7. Impact of input noise on V2V links and V2I links.

It can be seen from Figure 8 that the average interference received by the high-priority link increases with the increase of feedback noise, and there is no significant difference between the high-priority link and the ordinary link when the feedback noise is greater than 3 dB. This indicates that the effective information of link differentiation cannot be resolved when the feedback noise is high, but the priority can still be reflected when the feedback noise is low. When the feedback noise is greater than 10 dB, the average interference received by both of them is basically stable.

As shown in Figure 9, with the increase of feedback noise, the average rate of the V2I link in this scheme decreases gradually, and is basically unchanged when the feedback noise is greater than 10 dB. When the input noise is large, the rate can be maintained at 85.55%. The average rate of the V2V link decreases with the increase of feedback noise, and remains stable when the feedback noise is greater than 10 dB. When the input noise is large, the average rate of the V2V link remains 57.59%. The V2I link is less affected by feedback noise.

Compared with the random spectrum allocation scheme, C-Decision and SOLEN algorithm have improved transmission efficiency and anti-interference. However, C-Decision does not differentiate services between V2V links. The SOLEN scheme relies on CSI of all V2V and V2I links. The proposed scheme improves the transmission performance of high-priority links while ensuring the performance of common links, and is superior to C-Decision, SOLEN and traditional random spectrum allocation schemes in overall performance. At the same time, it has good robustness to input noise and feedback noise, and is more suitable for practical vehicle-mounted network scenarios.

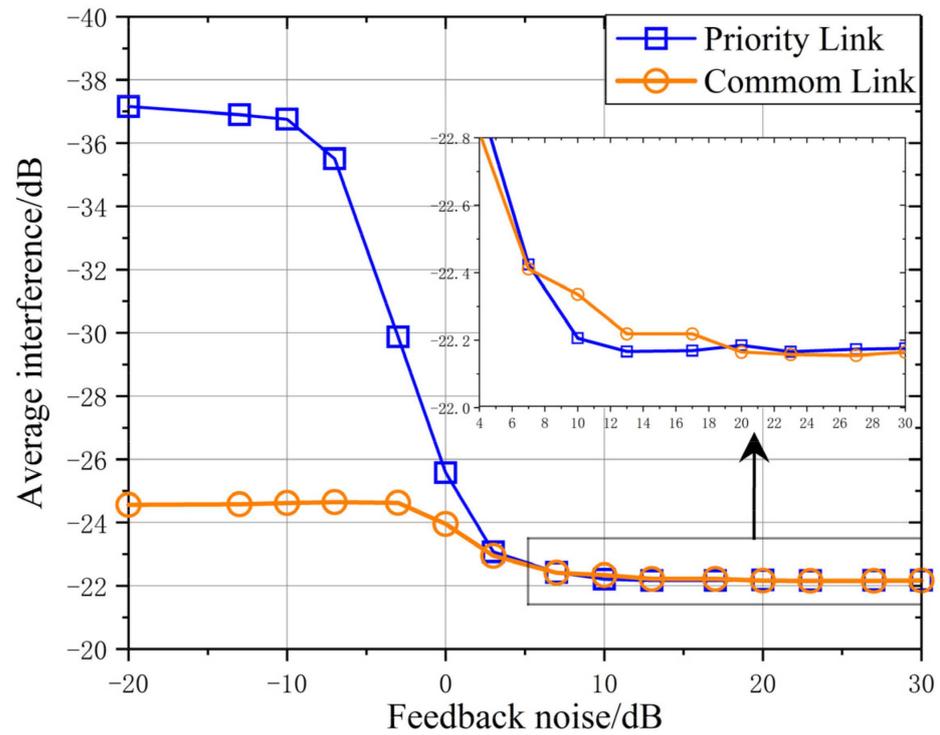


Figure 8. Impact of feedback noise on V2V links.

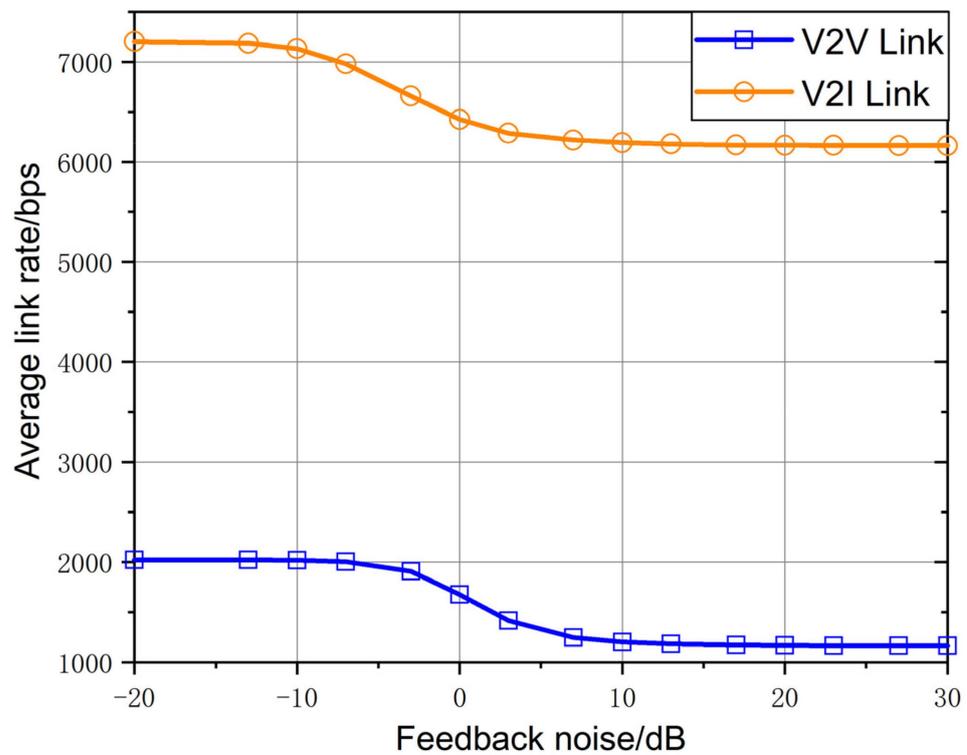


Figure 9. Impact of feedback noise on V2V links and V2I links.

At present, the proposed algorithm makes centralized spectrum allocation decisions at the base station. In order to reduce computational complexity, we can consider distributed spectrum resource allocation in the next step, and each V2V link makes local spectrum allocation decisions. In the future, we will also consider improving our algorithm to be

more suitable for joint V2I link and V2V link resource allocation problems in the future Internet of vehicles.

6. Conclusions

In order to realize dynamic spectrum resource allocation based on service types, this paper proposes a link priority-based spectrum resource allocation scheme based on reinforcement learning to maximize the total capacity of V2V and V2I links and minimize the received interference of priority links to achieve optimal spectrum allocation. Simulation results show that this method effectively solves the complex spectrum of on-board network resource allocation problems, maximizes V2V and V2I link capacity—the V2V link rate is 2.54 times that of the traditional random spectrum allocation scheme, and the V2I link rate is 13.5% higher than that of the traditional random spectrum allocation scheme—reduces the priority link interference by 14.2 dB relative to the common link and not only provides the high priority link the transmission quality of service guarantee, but there is no harm to ordinary link performance; in addition, the scheme of communication has good noise robustness and better universality in the actual scene. In the future, the proposed scheme will also be improved, and further in-depth research will be carried out in terms of diversified business types and distributed decision control so as to explore resource scheduling and access control technologies that are more suitable for future high-dynamic complex vehicle-mounted networks.

Author Contributions: Conceptualization, Z.G. and M.H.; Formal analysis, M.H.; Funding acquisition, Z.G.; Investigation, Y.W.; Methodology, Z.G. and M.H.; Project administration, Z.G.; Resources, Y.W. and M.H.; Visualization, Y.W.; Writing—original draft, Y.W.; Writing—review & editing, Z.G. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation of China (NSFC), grant number 61761045.

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Nurcahyani, I.; Lee, J.W. Role of Machine Learning in Resource Allocation Strategy over Vehicular Networks: A Survey. *Sensors* **2021**, *21*, 6542. [[CrossRef](#)]
2. Xiao, S.; Wang, S.; Zhuang, J.; Wang, T.; Liu, J. Research on a Task Offloading Strategy for the Internet of Vehicles Based on Reinforcement Learning. *Sensors* **2021**, *21*, 6058. [[CrossRef](#)]
3. Li, D.; Xu, S.; Li, P. Deep Reinforcement Learning-Empowered Resource Allocation for Mobile Edge Computing in Cellular V2X Networks. *Sensors* **2021**, *21*, 372. [[CrossRef](#)]
4. Park, H.; Lim, Y. Reinforcement Learning for Energy Optimization with 5G Communications in Vehicular Social Networks. *Sensors* **2020**, *20*, 2361. [[CrossRef](#)] [[PubMed](#)]
5. Tarek, D.; Benslimane, A.; Darwish, M.; Kotb, A.M. Survey on spectrum sharing/allocation for cognitive radio networks Internet of Things. *Egypt. Inform. J.* **2020**, *21*, 231–239. [[CrossRef](#)]
6. Zhang, Z.; Xiao, Y.; Ma, Z.; Xiao, M.; Ding, Z.; Lei, X.; Karagiannidis, G.K.; Fan, P. 6G wireless networks: Vision, requirements, architecture, and key technologies. *IEEE Veh. Technol. Mag.* **2019**, *14*, 28–41. [[CrossRef](#)]
7. Liang, L.; Li, G.Y.; Xu, W. Resource allocation for D2D-enabled vehicular communications. *IEEE Trans. Commun.* **2017**, *65*, 3186–3197. [[CrossRef](#)]
8. Guo, C.; Liang, L.; Li, G.Y. Resource allocation for vehicular communications with low latency and high reliability. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 3887–3902. [[CrossRef](#)]
9. Sun, W.; Strom, E.G.; Brannstrom, F.; Sou, K.C.; Sui, Y. Radio resource management for D2D-based V2V communication. *IEEE Trans. Veh. Technol.* **2015**, *65*, 6636–6650. [[CrossRef](#)]
10. Li, B.; He, D.; Feng, Y.; Xu, Y.; Zheng, H. Spectrum resource allocation scheme for alarm information delivery in V2V communication. In Proceedings of the 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 27–30 August 2018; pp. 1–5.
11. Luo, Q.; Li, C.; Luan, T.H.; Wen, Y. Optimal utility of vehicles in LTE-V scenario: An immune clone-based spectrum allocation approach. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 1942–1953. [[CrossRef](#)]

12. Farshbafan, M.K.; Bahonar, M.H.; Khaiehraveni, F. Spectrum trading for Device-to-Device communication in cellular networks using incomplete information bandwidth-auction game. In Proceedings of the 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, 30 April–2 May 2019; pp. 1441–1447.
13. Bhattarai, S.; Park, J.M.; Lehr, W. Dynamic exclusion zones for protecting primary users in database-driven spectrum sharing. *IEEE/ACM Trans. Netw.* **2020**, *28*, 1506–1519. [[CrossRef](#)]
14. Du, B.; Xue, R.; Zhao, L.; Leung, V.C.M. Coalitional graph game for air-to-air and air-to-ground cognitive spectrum sharing. *IEEE Trans. Aerosp. Electron. Syst.* **2019**, *56*, 2959–2977. [[CrossRef](#)]
15. Jiang, C.; Zhang, H.; Ren, Y.; Han, Z.; Chen, K.; Hanzo, L. Machine learning paradigms for next-generation wireless networks. *IEEE Wirel. Commun.* **2016**, *24*, 98–105. [[CrossRef](#)]
16. Karmakar, R.; Chattopadhyay, S.; Chakraborty, S. Dynamic link adaptation in IEEE 802.11 ac: A distributed learning based approach. In Proceedings of the 2016 IEEE 41st Conference on Local Computer Networks (LCN), Dubai, United Arab Emirates, 7–10 November 2016; pp. 87–94.
17. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [[CrossRef](#)] [[PubMed](#)]
18. Zia, K.; Javed, N.; Sial, M.N.; Ahmed, S.; Pervez, F. Multi-agent RL based user-centric spectrum allocation scheme in D2D enabled hetnets. In Proceedings of the 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Barcelona, Spain, 17–19 September 2018; pp. 1–6.
19. Liang, L.; Ye, H.; Li, G.Y. Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2282–2292. [[CrossRef](#)]
20. Ye, H.; Li, G.Y.; Juang, B.H.F. Deep reinforcement learning based resource allocation for V2V communications. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3163–3173. [[CrossRef](#)]
21. He, Y.; Zhao, N.; Yin, H. Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach. *IEEE Trans. Veh. Technol.* **2017**, *67*, 44–55. [[CrossRef](#)]
22. Wang, L.; Ye, H.; Liang, L.; Li, G.Y. Learn to compress CSI and allocate resources in vehicular networks. *IEEE Trans. Commun.* **2020**, *68*, 3640–3653. [[CrossRef](#)]
23. He, Z.; Wang, L.; Ye, H.; Li, G.Y.; Juang, B.-H.F. Resource Allocation based on Graph Neural Networks in Vehicular Communications. In Proceedings of the GLOBECOM 2020—2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020; pp. 1–5.
24. Hu, X.; Xu, S.; Wang, L.; Wang, Y.; Liu, Z.; Xu, L.; Li, Y.; Wang, W. A joint power and bandwidth allocation method based on deep reinforcement learning for V2V communications in 5G. *China Commun.* **2021**, *18*, 25–35. [[CrossRef](#)]
25. Xiang, P.; Shan, H.; Wang, M.; Xiang, Z.; Zhu, Z. Multi-Agent RL Enables Decentralized Spectrum Access in Vehicular Networks. *IEEE Trans. Veh. Technol.* **2021**, *70*, 10750–10762. [[CrossRef](#)]
26. Yuan, Y.; Zheng, G.; Wong, K.K.; Letaief, K.B. Meta-reinforcement learning based resource allocation for dynamic V2X communications. *IEEE Trans. Veh. Technol.* **2021**, *70*, 8964–8977. [[CrossRef](#)]
27. Liu, J.; Zhao, B.; Xin, Q.; Liu, H. Dynamic channel allocation for satellite internet of things via deep reinforcement learning. In Proceedings of the 2020 International Conference on Information Networking (ICOIN), Barcelona, Spain, 7–10 January 2020; pp. 465–470.
28. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)] [[PubMed](#)]
29. 3rd Generation Partnership Project (3GPP). In *Study on LTE-based V2X Services (Release 14)*; Technical Specification Group Radio Access Network; 3GPP: Sophia Antipolis, France, 2016.
30. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
31. Franklin, J. The elements of statistical learning: Data mining, inference and prediction. *Math. Intell.* **2005**, *27*, 83–85. [[CrossRef](#)]