Article

Deep Reinforcement Learning-Based Resource Allocation for Cellular Vehicular Network Mode 3 with Underlay Approach

Jinjuan Fu, Xizhong Qin, Yan Huang, Li Tang and Yan Liu

1 College of Information Science and Engineering, Xinjiang University, Urumqi 830000, China; fujinj@stu.xju.edu.cn (J.F.); dandelion@stu.xju.edu.cn (Y.L.)
2 Network Department, China Mobile Communications Group Xinjiang Co., Ltd., Urumqi 830000, China; huangyan@sj.chinamobile.xh.com (Y.H.); tangli@sj.chinamobile.com (L.T.)
* Correspondence: qinxz@xju.edu.cn

Abstract: Vehicle-to-vehicle (V2V) communication has attracted increasing attention since it can improve road safety and traffic efficiency. In the underlay approach of mode 3, the V2V links need to reuse the spectrum resources preoccupied with vehicle-to-infrastructure (V2I) links, which will interfere with the V2I links. Therefore, how to allocate wireless resources flexibly and improve the throughput of the V2I links while meeting the low latency requirements of the V2V links needs to be determined. This paper proposes a V2V resource allocation framework based on deep reinforcement learning. The base station (BS) uses a double deep Q network to allocate resources intelligently. In particular, to reduce the signaling overhead for the BS to acquire channel state information (CSI) in mode 3, the BS optimizes the resource allocation strategy based on partial CSI in the framework of this article. The simulation results indicate that the proposed scheme can meet the low latency requirements of V2V links while increasing the capacity of the V2I links compared with the other methods. In addition, the proposed partial CSI design has comparable performance to complete CSI.

Keywords: vehicular network; deep reinforcement learning; resource management; low latency

1. Introduction

Vehicular communication is a vital realization technology of automatic driving and intelligent transportation systems [1]. Various candidate technical solutions have been proposed, such as cellular vehicular communication and dedicated short-range communication (DSRC) based on IEEE 802.11p [2,3]. DSRC has the disadvantages of short coverage distance and high infrastructure cost [4,5]. Despite the deployment of DSRC-based vehicular communication prototypes in the United States, inherent issues with DSRC and recent developments in cellular technology have encouraged research and industry research on cellular-based vehicular communications [6]. 5G Automotive Association (5GAA), as a global and cross-industry organization, was created in September 2016 and has presented the clear position of supporting cellular vehicular communication as the feasible solution for future mobility and transportation services [7]. Different regions such as the United States, Europe, and Japan have carried out pilot construction of cellular vehicular communication in actual operation [8]. Through cellular vehicular communication, road safety and traffic efficiency can be improved [9]. Therefore, cellular vehicular communication has attracted more attention from industry and academia. 3GPP Release 14 introduced modes 3 and 4 for vehicular communication [10]. These modes enable vehicles to directly communicate via licensed cellular frequency bands, bypassing the BS. Modes 3 and 4 both support direct vehicular communications, but their radio resource allocation methods are different. In mode 3, the BS allocates resources to vehicle users, while in mode 4, vehicles autonomously select radio resources [10,11]. Mode 3 can be divided into two types: overlay method and underlay method. In the overlay approach, V2I users and V2V users transmit...
in different frequency bands, while in the underlay method, V2V users reuse spectrum resources allocated to V2I users, which improves the spectrum utilization but interferes with V2I users. Therefore, in the underlay approach of mode 3 how to flexibly allocate resources to meet the low latency requirements of V2V users while increasing the throughput of V2I users is challenging.

2. Related Work

There have been more studies on the allocation of wireless resources for vehicular communication. The existing spectrum resource sharing schemes are mainly divided into two categories: traditional methods [12–14] and machine learning methods [15–17]. In conventional methods, under different constraints such as transmit power, quality of service, mathematical programming is used to optimize design goals of interest, such as maximizing spectrum efficiency or minimizing interference. Under normal circumstances, resource allocation problems are usually non-convex or mixed-integer nonlinear programming. Due to its NP-hardness, it is difficult to solve using effective methods. In order to deal with such problems, heuristic methods are usually used or sub-optimal solutions are pursued. In addition, the solution is highly dependent on the model. If you consider different objective functions or add new constraints, you need to develop a new solution [18]. Therefore, these many reasons have promoted the exploration of new resource allocation schemes.

In recent years, machine learning has become a potential implementation technology for future wireless communications. Reinforcement learning (RL) in machine learning is widely used in wireless communication resource allocation to solve the challenges in traditional optimization methods. In a high-speed moving vehicular network, RL can perceive changes through interaction with the environment and make corresponding decisions. In addition, the goal problems and constraints that are difficult to optimize can be solved by designing corresponding training rewards [19]. Q-learning is a typical RL algorithm, which avoids the difficulty of acquiring dynamics by adopting the iterative update method. However, since Q-learning stores the state in the form of a table, it is not suitable for application in scenarios with too large a state space. By combining deep neural network (DNN) with Q-learning, deep Q-learning (DQN) is designed to handle larger state spaces. In [20], the problem of vehicular network spectrum sharing based on multi-agent RL was studied, resource sharing was modeled as a multi-agent RL problem, and then the fingerprint-based DQN method was used to solve it. In [21], the spectrum allocation problem in the vehicular network was studied, and the graph neural network (GNN) and DQN learning were used to optimize the sum capacity of the vehicular network. In [22], the problem of shared resource allocation between V2I and V2V links is studied.

Under the deep reinforcement learning (DRL) framework, the existing resource allocation methods can be divided into modes 3 and 4. In the mode 3 scheme [23–25], the BS acts as a central controller to allocate resources for each user. In [23], a joint DRL algorithm was proposed, which can effectively deal with the discrete–continuous mixed action space in wireless communication systems and solve the power and sub-channel allocation problems of non-orthogonal multiple access systems. In [24], a resource allocation method based on graphs and DRL is proposed, in which channel resources are centrally allocated by BS, and the vehicle user uses DRL for distributed power control. In [25], the joint channel allocation and power allocation problem was transformed into an optimization problem, and a DRL framework based on an attention neural network was proposed to solve the problem. In the above-mentioned schemes, the BS needs to know the CSI of all users, that is, V2V user’s communication channel gain, the interference link information, and the user’s transmit power to determine the spectrum sharing decision. Feeding this information back to the BS will result in more significant overhead, especially when there are many users. This limits the application of the centralized solution in practice.

In the mode 4 resource scheme [26–28], the users independently select the transmission spectrum without the participation of the BS, which reduces the signaling overhead of the BS. In [26], a DRL-based V2V communication distributed resource allocation mechanism
that can be used in both unicast and broadcast scenarios was developed to learn how to meet V2V latency constraints while minimizing the interference to V2I communication. In [27], a DRL-based mode selection and resource allocation method are proposed to maximize the total capacity of V2I users while ensuring the delay and reliability requirements of V2V users. In these schemes, each V2V pair independently updates its strategy at different times to maximize its reward. However, when two or more V2V pairs update their policies at the same time, it becomes a seemingly unstable environment. In [28], an effective transfer actor–critic learning approach was proposed to ensure the ultra-reliable and low-latency communication requirements of the V2V links, but this work can only be applied to low-dimensional state–action mapping. Using a distributed scheme can significantly reduce signaling overhead, but distributed DRL may lead to local optimization. Based on the research work analysis of the above-mentioned mode 3 and mode 4 resource allocation schemes, how to design an effective V2V resource allocation framework is the key to improving the data rate of V2I communication and meeting the requirements of V2V communication delay.

The analysis of the research work in [23–28] shows that DRL-based resource allocation with delay as the constraint is more common in the mode 4 scheme. However, due to the high signaling overhead, there are few studies on the mode 3 solution and primarily focus on solving the rate problem between V2I users and V2V users, without considering the delay requirements of V2V users. Compared with mode 4, the advantage of mode 3 is that the BS collects the CSI of all vehicles in the coverage area and can allocate spectrum resources more effectively. However, in a high-speed moving vehicle environment, it is a challenge for the BS to obtain complete CSI. Based on the above considerations, this paper proposes a DRL-based V2V resource allocation framework in the underlay of mode 3. In this framework, the BS uses double DQN for resource allocation. In particular, to reduce the overhead of CSI feedback, in the framework of this article, the V2V links only feedback partial CSI, and the BS only optimizes the resource allocation strategy based on the feedback information.

The main contributions of this work are summarized in the following aspects:

1. In view of the low latency requirements of V2V communication, a DRL-based resource allocation framework is proposed in the underlay method of mode 3, which aims to meet the V2V links latency requirements with low signaling overhead while increasing the V2I link throughput. It is novel to consider the delay requirement of V2V links in the DRL-based mode 3 spectrum allocation scheme.

2. To reduce the signaling overhead of BS acquiring CSI in the resource allocation scheme of mode 3, in this design, each V2V link only needs to feed back interference links information and delay information to the BS, instead of feeding back complete CSI signaling. The BS only optimizes the resource allocation strategy based on the feedback information, which significantly reduces the CSI feedback overhead.

3. The simulation results show that, compared with other solutions, the proposed solution can not only meet the low latency requirements of the V2V links but also increase the throughput of the V2I links, and the proposed partial CSI has comparable performance compared with the complete CSI.

The rest of the paper is organized as follows. The system model is introduced in Section 3. Section 4 presents the DRL framework to solve the problem formulation of spectrum resource allocation. The simulation results and analysis are shown in Section 5. Section 6 is the conclusion.

3. System Model

Since the uplink resources are under-utilized compared with the downlink ones [29], we only consider uplink cellular resource allocation.

As illustrated in Figure 1, an uplink scenario in a single cell system is considered. BS was lying in the center of the cell. The road configuration for vehicular communication is defined as an urban setting [30]. The cell consists of multiple V2I links, denoted as
$M = \{1, \ldots, M\}$ and multiple V2V pairs denoted as $N = \{1, \ldots, N\}$. In order to improve the spectrum efficiency and ensure the data rate requirements of the V2I links. We assume that each V2I user has assigned one orthogonal subcarrier resource and each V2I user allows more than one V2V pair to share its resource.

![System model for vehicular communication underlaying cellular network in uplink.](image)

**Figure 1.** System model for vehicular communication underlaying cellular network in uplink.

We define the channel power gain of the desired transmission for $m$th V2I link and $n$th V2V pair on the $n$th channel as $g_{mn}^m$. Similarly, let $g_{mn}$ represent the interfering channel power gain from the $m$th V2I transmitter to the V2V pair $n$’s receiver. The interference channel power gain between $n$th V2V link and the BS is denoted as $g_{n}$, and the interference channel gain from the V2V pair $l$’s transmitter to the V2V pair $n$’s receiver is denoted as $g_{ln}^n$. In the simulation in this article, all the channel gains mentioned above include path loss, shadow fading, and small-scale fading.

Then, we can respectively express the signal-to-interference-plus-noise-ratio (SINR) for the $m$th V2I link and $n$th V2V pair on subcarrier $m$ as

$$\gamma_m = \frac{p_m s_m^m}{\sum_{n=1}^{N} p_n s_n^m + \sigma^2}$$  \hspace{1cm} (1)

$$\gamma_n = \frac{p_n s_n^m}{\sum_{n \neq n} p_l^{m} s_{l,n}^{m} + p_m s_m^m + \sigma^2}$$  \hspace{1cm} (2)

where $p_m, p_n, p_l$ denotes the transmit powers of the $m$th V2I, $n$th V2V pair, and $l$th V2V pair. $\rho_n^m$ is the subcarrier access indicator, $\rho_n^m \in \{0, 1\}$, $\rho_n^m = 1$ if $n$th V2V pair is allowed to use subcarrier $m$, otherwise, $\rho_n^m = 0$. $\sigma^2$ indicates the noise power.

Furthermore, to reduce interference between V2V links, we assume each V2V link can only access one V2I link resource. Accordingly, the sum data rate of $m$th V2I link and $n$th V2V pair on subcarrier $m$ can be respectively given by

$$R_m = \log_2 (1 + \gamma_m)$$  \hspace{1cm} (3)

$$R_n = \log_2 (1 + \gamma_n)$$  \hspace{1cm} (4)
We model the latency requirements of V2V links as the successful delivery of packets of size L within a limited time budget $T_{\text{max}}$. Therefore, if the following constraint holds, the latency requirements of a V2V link will be satisfied:

$$\Pr \left\{ \sum_{t=1}^{T_{\text{max}}} \sum_{m=1}^{M} \rho_n[m] R_{n}^{m}[t] \geq L / \Delta T \right\}$$

(5)

where $\Delta T$ is channel coherence time, and the index $t$ is added in $R_{n}^{m}[t]$ to indicate the capacity of the $n$th V2V link at different coherence time slots.

4. DRL for Resource Management

In this section, we describe how to use the DRL method within the proposed framework to solve the spectrum allocation, including the proposed partial CSI observations and double DQN learning algorithm.

For mode 3 resource allocation, our objective is to find the optimal resource allocation policy that can meet the latency demand of V2V links without excessively affecting the V2I links data rate under low signaling overhead. In addition, the vehicular communication network needs an intelligent resource management framework to make decisions intelligently, therefore, we adopt DRL architecture to make resource-sharing decisions. In the RL, an agent connects its environment, at each moment, the agent takes action based on the current environment state. The action will impact the state of the environment, causing certain changes in the environment. Then the agent receives numerical rewards and observes a new environment state. Through continuously exploiting the environment, agents learn how the environment generates rewards and then update their policy accordingly. In RL, the agent’s objective is to maximize the expected sum of rewards, which is defined as

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

(6)

where $\gamma \in [0,1]$ is called the discounter factor.

We adopt double DQN with experience replay architecture for subcarrier assignment, as shown in Figure 2. BS is regarded as a learning agent and the vehicular network acts as the environment state. Three key elements of the proposed model, i.e., state, action and reward, are defined, respectively, as follows.

**Figure 2. Proposed framework for resource allocation.**

As shown in Figure 2, each V2V pair observed local information, which includes communication interference channel power gain and latency information. We consider that it is difficult to obtain a complete CSI in the vehicle network, so only part of the CSI is considered in our observation space design. This helps to reduce the signaling overhead caused by channel estimation and CSI feedback. To consider the low latency communication
requirements, key elements related to latency should be involved. In Figure 2, \( I_{n,t-1}, L_{n,t}, T_{n,t} \) represents the interference power gain received by each sub-channel of the \( n \)th V2V link in the previous time slot, remaining payload size for transmission and the time left before violating the latency constraint; among them, \( I_{n,t}^m \) can be written as follows

\[
I_{n,t}^m = \sum_{n \neq m}^{N} p_{l,n}^m p_{l,n,t} + p_m g_{m,n,t}^m
\]

In summary, the partial CSI observation space of the \( n \)th V2V pair can be expressed as

\[
o_{n,t} = \{ I_{n,t-1}, L_{n,t}, T_{n,t} \}
\]

where \( I_{n,t} = (I_{1,n,t}, I_{2,n,t}, I_{3,n,t}, \ldots , I_{M,n,t}) \).

Specifically, the complete CSI observation space is defined as

\[
 o_{n,t} = \{ G_{n,t}, I_{n,t-1}, L_{n,t}, T_{n,t} \}
\]

where \( G_{n,t} = (g_{n,t,1}, g_{m,n,t,1}, g_{m,n,t,2}) \). Feeding back complete CSI to the BS will generate a large signaling overhead, and it is unrealistic in the dynamic environment of the vehicle, especially the links not connected to the BS, such as the V2V communication links and the interference of other vehicles on the V2V links. To obtain the \( g_{n,t,1}, g_{m,n,t,1}, g_{m,n,t,2} \) channel estimation at the receiver side is necessary, after which the estimation results can be fed back to the transmitter. Then the transmitter feeds back to BS. Therefore, we consider not using complete CSI. On the other hand, in mode 4, the user selects the spectrum resource with sensing channel measurements, the level of interference that will be experienced if the sensing user transmits in the corresponding spectrum resource [31]. Therefore, we are motivated to consider designing a partial CSI resource allocation mechanism using interference power measurement. Compared with the complete CSI version designed as (9), the partial CSI observation space designed as (8) can help reduce the signaling overhead.

(1) State Space: As stated above, to reduce signaling overhead in a centralized solution, in our solution, each V2V link feeds interference and delay information back to the BS. Hence, BS considers the partial CSI feedback from all V2V pairs as the current state. Thus, the state space of the BS is described as follows

\[
s_t = \{ o_{1,t}, o_{2,t}, o_{3,t}, \ldots , o_{N,t} \}
\]

(2) Action Space: In this paper, we focus on the subcarrier assignment issues in the V2V communication network. Hence, the action is defined as

\[
a_t = \{ b_{1,t}, b_{2,t}, b_{3,t}, \ldots , b_{N,t} \}
\]

where \( b_{n,t} = m, \forall m \in M \) represents that the \( n \)th subcarrier has been selected by the \( n \)th V2V pair.

(3) Reward: In the RL framework, the training process is driven by the reward function. An agent searches decision-making policy by maximizing reward under the interaction with the environment. Hence, to meet the different requirements of communication devices, corresponding reward functions need to be formulated. In vehicular communications, V2V links exchange critical safety information and have strict latency requirements, whereas V2I links support high-rate data transmission [32]. Our objective is to improve the sum throughput of V2I links while meeting the latency requirements of V2V links. Therefore, we propose the following reward function

\[
r_t = \lambda_s \sum_{m=1}^{M} R_{m,t} + \lambda_l \sum_{n=1}^{N} G_n
\]

where \( \lambda_s \) and \( \lambda_l \) are the weight factors for the contributions of these two parts to the reward function composition. In response to the first objective, we simply include the sum capacity of all V2I links.
To achieve V2V link low latency requirements, for each V2V link, we set the reward $G_n$ equal to the effective V2V transmission rate until the V2V links deliver the payload successfully within the delay constraint, after which the reward is set to a constant number, $c$, that is greater than the largest possible V2V transmission rate. Therefore, the sooner the V2V link completes the transmission, the more rewards can be obtained. As such, the V2V-related reward at each time step $t$ is

$$G_n(t) = \begin{cases} R^m_n(t) & L_n,t > 0 \\ c & L_n,t \leq 0 \end{cases}$$

(13)

Q-learning [33] is a typical RL algorithm, which can be used to learn the optimal strategy when the state–action space is small. In the Q-learning algorithm, the action-value function is used to calculate the expected accumulative rewards for starting from a state $s$ by taking action $a$ under policy $\pi$, which can be expressed by

$$Q^\pi(s,a) = E[R_t | s_t = s, a_t = a]$$

(14)

Similarly, the optimal action-value function is obtained by

$$Q^*(s,a) = E[r_{t+1} + \gamma \max_{a_{t+1}} Q^*(s_{t+1},a_{t+1}) | s_t = s, a_t = a]$$

(15)

where $s_{t+1}$ is the new state after taking action $a$. The action-value function is updated by

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha \left[ r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t) \right]$$

(16)

where $\alpha$ is the step-size parameter. Moreover, the choice of action $a_t$ in state $s_t$ follows exploitation and exploration policies, a widely used algorithm is the $\epsilon$-greedy algorithm [34], which is defined as

$$a \leftarrow \begin{cases} \text{argmax}_a Q(s,a) & \text{with probability } 1-\epsilon \\ \text{random action} & \text{with probability } \epsilon \end{cases}$$

(17)

Here, $\epsilon$ is the exploration rate. However, in larger networks, complex states and optional actions make Q-learning maintain a large Q-table and slow convergence speed, which limits the application scenario. Thus, we apply Q-learning techniques with a DNN parameterized by $\theta$ as the action-value function approximator to learn the optimal policies, thus called DQN [35]. To accelerate the learning convergence rate, we adopt double DQN [36] with three key techniques.

(1) Replay Memory: At each time step, BS observes the state $s_t$, determines resource allocation $a_t$, and broadcasts it to all V2V users. After the execution of the V2V links, the BS gets the corresponding reward $r_{t+1}$, and the environment reaches the next state $s_{t+1}$. In this way, experience samples are formed, that is, $(s_t, a_t, r_{t+1}, s_{t+1})$. Experience samples are stored in the replay memory. The replay memory accumulates the agent’s experiences over many episodes. After that, a mini-batch is uniformly sampled from memory for neural network training.

(2) Fixed Q-Target: As shown in Figure 2, our proposed scheme consists of double DQN, a target DQN and a training DQN, which have the same structure. The weight value $\theta_{\text{target}}$ in the target network is updated by the training network weight $\theta_{\text{train}}$ at regular intervals. This improves network stability and convergence.

(3) Action Selection–Evaluation Decoupling: In double DQN, the training DQN is used to select the optimal action for the next state, then the action selected by the training DQN is input and the next state into the target DQN to generate the target value to calculate the training loss. By changing the action of target DQN selection to training DQN selection, the risk of over-estimation of the Q value can be reduced.
After gathering sufficient experience samples, a mini-batch that consists of $D$ experience samples is retrieved from the buffer to minimize the sum-squared error

$$\sum_{t\in D} \left[ y_t - Q(s_t, a_t \mid \theta^{\text{train}}) \right]^2$$

(18)

where $y_t$ is the target $Q$-value, accordingly, the $y_t$ can be respectively given by

$$y_t = r_{t+1} + \gamma Q\left(s_{t+1}, \underset{a_{t+1}}{\text{argmax}} Q\left(s_{t+1}, a_{t+1} \mid \theta^{\text{train}} \right) \mid \theta^{\text{target}} \right)$$

(19)

Then, the updating process for the BS double DQN can be written as [36]

$$\theta^{\text{train}} = \theta^{\text{train}} + \beta \sum_{t\in D} \frac{\partial Q(s_t, a_t \mid \theta^{\text{train}})}{\partial \theta^{\text{train}}} \left[ y_t - Q(s_t, a_t \mid \theta^{\text{train}}) \right]$$

(20)

where $\beta$ is a nonnegative step size for each adjustment. Algorithm 1 summarizes the training process of the proposed scheme.

---

**Algorithm 1 Training Process for the Proposed Scheme**

1: Input: double DQN structure, vehicular environment simulator and V2V pair delay requirements
2: Output: double DQN networks’ weights
3: Initialize: experience replay buffer, the weights of train DQN $\theta^{\text{train}}$ and target DQN $\theta^{\text{target}}$
4: for each episode $j = 1, 2, \ldots$ do
5: Start the V2X environment simulator for each episode
6: Reset $L_{n,t} = L$ and $T_{n,t} = T_{\text{max}}$, for all $n \in N$
7: for each iteration step $t = 1, 2, \ldots$ do
8: Each V2V observes the observation $o_{n,t}$, sends it to BS
9: BS based on the current state $s_t = \{ o_{1,t}, \ldots, o_{n,t}, \ldots \}$, select the action according to the $\epsilon$-greedy, then gets a reward $r_{t+1}$, transforms to new state $s_{t+1}$
10: Store transition $(s_t, a_t, r_{t+1}, s_{t+1})$ into experience replay buffer
11: Sample a mini-batch of $D$ transition samples from experience replay buffer
12: Calculate the target value according to Equation (19)
13: Update $\theta^{\text{train}}$ according to Equation (20)
14: Update $\theta^{\text{target}}$ by setting $\theta^{\text{target}} = \theta^{\text{train}}$ every $K$ steps
15: end for
16: end for

---

5. Simulation Results and Analysis

In this section, we provide the simulation results to demonstrate the performance of the proposed resource allocation method.

5.1. Simulation Settings

We considered a single-cell scenario. BS is located in the center of the region. The simulation setup we used was based on the urban case in 3GPP TR 36.885 [30]. We followed the main simulation setup in [22]. The main simulation parameters are summarized in Table 1, and the channel models of V2V and V2I links are shown in Table 2.
Table 1. Simulation parameters.

| Parameters                        | Values       |
|-----------------------------------|--------------|
| Carrier frequency                 | 2 GHz        |
| Subcarrier bandwidth              | 1 MHz        |
| BS antenna height                 | 25 m         |
| BS antenna gain                   | 8 dBi        |
| BS receive noise figure           | 5 dB         |
| Vehicle antenna height            | 1.5 m        |
| Vehicle antenna gain              | 3 dBi        |
| Vehicle receive noise figure      | 9 dB         |
| Transmit power of V2I             | 35 dBm       |
| Transmit power of V2V             | 23 dBm       |
| Number of V2I links               | 4            |
| Number of V2V pairs               | 4            |
| $[\lambda_c, \lambda_v]$         | [0.1, 0.9]   |
| Noise power                       | $-114 \text{dBm}$ |
| Vehicle speed                     | 50 km/h      |
| Latency constraint of V2V links   | 100 ms       |

Table 2. Channel models.

| Parameters                        | V2I Link | V2V Link |
|-----------------------------------|----------|----------|
| Path loss model                   | $128.1 + 37.6 \log_{10}(d), d \text{ in km}$ | WINNER + B1 |
| Shadowing distribution            | Log-normal | Log-normal |
| Shadowing standard deviation      | 8 dB      | 3 dB      |
| Decorrelation distance            | 50 m      | 10 m      |
| Fast fading                       | Rayleigh fading | Rayleigh fading |

The double DQN for BS consists of three fully connected hidden layers, containing 1200, 800, and 600 neurons, respectively. The activation function used by the hidden layer of double DQN networks is ReLu $f(x) = \max(0, x)$. The RMSProp optimizer [37] is used to update network parameters. The discount factor of the double DQN algorithm is set to 0.05, and the learning rate is set to 0.001. Moreover, $\theta_{\text{target}}$ is updated with $\theta_{\text{train}}$ every 500 steps.

We trained the whole neural network for 4000 episodes and fixed the payload size for V2V links during the training phase to be 1060 bytes, which is varied during the test phase.

5.2. Performance Comparisons under Different Parameters

To verify the effectiveness of the algorithm in this paper, we compared the proposed algorithm with the following three algorithms.

1. meta-DRL [22]: in this scheme, using DQN to solve the problem of spectrum resource allocation, and applying deep deterministic policy gradient (DDPG) to solve the problem of continuous power allocation.

2. Brute-Force method: the action (including channel and power) is searched exhaustively to maximize the rate of V2V links.

3. Random method: the channel and power are randomly selected.

Since we used the interference power measurement to design a partial CSI resource allocation mechanism, to prove the performance advantage of partial CSI, we compared it with complete CSI.

Figure 3 shows the changes in the V2V link’s successful transmission probability, V2I links throughput as the payload changes. Figure 3a reflects that as the payload size increases, the probability of successful transmission of the V2V links of all schemes decreases, including brute-force. This is because more information needs to be transmitted within the same delay constraint, so performance will decrease. However, it is found that the method in this paper is very close to brute-force and far superior to the random scheme. Since [22] considers the power control of the V2V links, the successful transmission probability of the
V2V links is slightly higher than the scheme proposed in this paper. Furthermore, as can be seen from Figure 3a,b, the partial CSI proposed in this paper has comparable performance to complete CSI in meeting the V2V links delay.

![Figure 3a](image1.png)  
**Figure 3a.** Performance comparisons with different payload sizes. (a) V2V payload transmission success probability with varying payload sizes. (b) Sum capacity performance of V2I links with varying payload sizes.

Figure 3b reflects that as the payload increases, the total throughput of the V2I link gradually decreases. This is because when the V2V payload increases, in order to obtain higher rewards, the BS will select actions to increase the V2V rate to meet the delay constraint, which will increase the interference of the V2I links and cause its rate to decrease. However, it can be seen that the throughput of the V2I links obtained by the proposed scheme is still higher than that of the random scheme and has comparable performance to the complete CSI scheme. Furthermore, the proposed scheme outperforms the literature [22] in terms of V2I throughput. In summary, the scheme proposed in this paper is close to brute-force in terms of the V2V links’ successful transmission probability and the throughput of the V2I links, which is better than the random scheme and has comparable performance to complete CSI.

Figure 4 shows the change in the probability of successful transmission of the V2V links as the vehicle speed increases. It can be seen from the figure that the probability
of successful transmission of the V2V links in the proposed scheme decreases. This is because compared to the low-speed state, when the vehicle is in the high-speed state, the environmental state changes more significantly, leading to higher observation uncertainty and reducing the learning efficiency. Therefore, when the vehicle speed gradually increases, the V2V links’ successful transmission probability declines. However, the proposed scheme can still maintain a higher probability of successful transmission, which shows that the proposed algorithm has better stability in a highly dynamic environment.

![Figure 4](image-url). V2V payload transmission success probability with varying vehicle speeds.

In order to more clearly explain the reason why the solution in this article is better than the random solution, we randomly selected an episode in the test and plotted the changes in the remaining load of the V2V links over time. Among them, the delay constraint T = 100 ms, and the payload size is 3 × 1060 bytes. Since in the randomly selected episode, all V2V links in the proposed scheme and the random scheme have completed the transmission task within 50ms, we only show the data within 0–50 ms. Comparing Figure 5a, b, it can be seen that although all V2V links of the proposed scheme and the random scheme have completed data transmission within the delay constraint, the transmission time required by the scheme of this paper is much shorter than that of the random scheme. As shown in Figure 5a, all vehicles in the scheme in this paper are transmitted within 15ms, while the random scheme in Figure 5b is completed within 42 ms. This shows that the solution in this paper is more suitable for transmitting delay sensitive services and meeting the delay requirements of the V2V links.

![Figure 5](image-url). Cont.
Figure 5. Remaining V2V payload comparisons between proposed and random schemes. (a) The remaining payload of the proposed. (b) The remaining payload of the random.

Figure 6 reflects the impact of V2I power changes on V2I throughput and the probability of successful V2V link transmission. It can be seen that with the increase in V2I power, the throughput of the V2 link increases, and at the same time, the interference to the V2V links increases, so the probability of successful transmission of the V2V links decreases. Therefore, it is necessary to reasonably set the power of the V2I links to meet the throughput requirements of the V2I links and the delay requirements of the V2V links.

Figure 6. Performance comparisons with varying power. (a) Sum capacity performance of V2I links with varying power. (b) V2V payload transmission success probability with varying power.
6. Conclusions

In this article, we developed a DRL-based resource sharing scheme in the underlay approach of mode 3, in which the V2V links reuse the V2I links spectrum. Our goal is to improve the throughput of V2I links while ensuring the V2V links delay constraint. In particular, to reduce the signaling overhead for the BS to acquire complete CSI in mode 3, an intelligent resource allocation strategy based on partial CSI is proposed. The BS only allocates a spectrum based on the feedback CSI, which significantly reduces the signaling overhead. The simulation results show that compared with other methods, the proposed scheme can meet the V2V links delay constraint and has a higher V2I links throughput, and the proposed partial CSI scheme has comparable performance as the complete CSI scheme.

Author Contributions: Conceptualization, J.F. and X.Q.; methodology, J.F.; software, J.F. and X.Q.; validation, J.F., X.Q. and Y.L.; formal analysis and investigation, J.F.; resources, Y.H. and L.T.; writing—original draft preparation, J.F.; writing—review and editing, J.F. and X.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Xinjiang Uygur Autonomous Region Major Science and Technology Special Fund Project. The funded project number is: 2020A03001-2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Balkus, S.V.; Wang, H.; Cornet, B.D.; Mahabal, C.; Ngo, H.; Fang, H. A Survey of Collaborative Machine Learning Using 5G Vehicular Communications. *IEEE Commun. Surv. Tutor.* 2022. [CrossRef]

2. Kimura, T. Performance Analysis of Cellular-Relay Vehicle-to-Vehicle Communications. *IEEE Trans. Veh. Technol.* 2021, 70, 3396–3411. [CrossRef]

3. Rahim, N.-A.-R.; Liu, Z.; Lee, H.; Ali, G.G.M.N.; Pesch, D.; Xiao, P. A Survey on Resource Allocation in Vehicular Networks. *IEEE Trans. Intell. Transp. Syst.* 2020, 23, 701–721. [CrossRef]

4. Yang, Y.; Hua, K. Emerging Technologies for 5G-Enabled Vehicular Networks. *IEEE Access* 2019, 7, 181117–181141. [CrossRef]

5. Le, T.T.T.; Moh, S. Comprehensive Survey of Radio Resource Allocation Schemes for 5G V2X Communications. *IEEE Access* 2021, 9, 123117–123133.

6. Gyawali, S.; Xu, S.; Qian, Y.; Hu, R.Q. Challenges and Solutions for Cellular Based V2X Communications. *IEEE Commun. Surv. Tutor.* 2021, 23, 222–255. [CrossRef]

7. Kumar, A.S.; Zhao, L.; Fernando, X. Multi-Agent Deep Reinforcement Learning-Empowered Channel Allocation in Vehicular Networks. *IEEE Trans. Veh. Technol.* 2022, 71, 1726–1736. [CrossRef]

8. Chen, S.; Hu, J.; Shi, Y.; Zhao, L.; Li, W. A Vision of C-V2X: Technologies, Field Testing, and Challenges with Chinese Development. *IEEE Internet Things J.* 2020, 7, 3872–3881. [CrossRef]

9. Li, X.; Ma, L.; Shankaran, R.; Xu, Y.; Orgun, M.A. Joint Power Control and Resource Allocation Mode Selection for Safety-Related V2X Communication. *IEEE Trans. Veh. Technol.* 2019, 68, 7970–7986. [CrossRef]

10. Molina-Masegosa, R.; Gozalvez, J. LTE-V for Sidelink 5G V2X Vehicular Communications: A New 5G Technology for Short-Range Vehicle-to-Everything Communications. *IEEE Veh. Technol. Mag.* 2017, 12, 30–39. [CrossRef]

11. Abbas, F.; Fan, P.; Khan, Z. A Novel Low-Latency V2V Resource Allocation Scheme Based on Cellular V2X Communications. *IEEE Trans. Intell. Transp. Syst.* 2019, 20, 2185–2197. [CrossRef]

12. Li, X.; Ma, L.; Xu, Y.; Shankaran, R. Resource Allocation for D2D-Based V2X Communication with Imperfect CSI. *IEEE Internet Things J.* 2020, 7, 3545–3558. [CrossRef]

13. Aslani, R.; Saberinia, E.; Rasti, M. Resource Allocation for Cellular V2X Networks Mode-3 with Underlay Approach in LTE-V Standard. *IEEE Trans. Veh. Technol.* 2020, 69, 8601–8612. [CrossRef]

14. Jameel, F.; Khan, W.U.; Kumar, N.; Jäntti, R. Efficient Power-Splitting and Resource Allocation for Cellular V2X Communications. *IEEE Trans. Intell. Transp. Syst.* 2021, 22, 3547–3556. [CrossRef]

15. Tang, F.; Kawamoto, Y.; Kato, N.; Liu, J. Future Intelligent and Secure Vehicular Network toward 6G: Machine-Learning Approaches. *Proc. IEEE* 2019, 108, 292–307. [CrossRef]

16. Hussain, F.; Hassan, S.A.; Hussain, R.; Hossain, E. Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges. *IEEE Commun. Surv. Tutor.* 2020, 22, 1251–1275. [CrossRef]
17. Tang, F.; Mao, B.; Kato, N.; Gui, G. Comprehensive Survey on Machine Learning in Vehicular Network: Technology, Applications and Challenges. *IEEE Commun. Surv. Tutor.* 2021, 23, 2027–2057. [CrossRef]
18. Liang, L.; Ye, H.; Yu, G.; Li, G.Y. Deep-Learning-Based Wireless Resource Allocation with Application to Vehicular Networks. *Proc. IEEE* 2020, 108, 341–356. [CrossRef]
19. Li, Z.; Guo, C. Multi-Agent Deep Reinforcement Learning Based Spectrum Allocation for D2D Underlay Communications. *IEEE Trans. Veh. Technol.* 2020, 69, 1828–1840. [CrossRef]
20. 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]
21. He, Z.; Wang, L.; Ye, H.; Li, G.Y.; Juang, B.-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.
22. 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]
23. Wang, X.; Zhang, Y.; Shen, R.; Xu, Y.; Zheng, F.-C. DRL-Based Energy-Efficient Resource Allocation Frameworks for Uplink NOMA Systems. *IEEE Internet Things J.* 2020, 7, 7279–7294. [CrossRef]
24. Gyawali, S.; Qian, Y.; Hu, R.Q. Resource Allocation in Vehicular Communications Using Graph and Deep Reinforcement Learning. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
25. He, C.; Hu, Y.; Chen, Y.; Zeng, B. Joint Power Allocation and Channel Assignment for NOMA with Deep Reinforcement Learning. *IEEE J. Sel. Areas Commun.* 2019, 37, 2200–2210. [CrossRef]
26. Ye, H.; Li, G.Y.; Juang, B.-F. Deep Reinforcement Learning Based Resource Allocation for V2V Communications. *IEEE Trans. Veh. Technol.* 2019, 68, 3163–3173. [CrossRef]
27. Zhang, X.; Peng, M.; Yan, S.; Sun, Y. Deep-Reinforcement-Learning-Based Mode Selection and Resource Allocation for Cellular V2X Communications. *IEEE Internet Things J.* 2020, 7, 6380–6391. [CrossRef]
28. Yang, H.; Xie, X.; Kadoch, M.; Rong, B. Intelligent Resource Management Based on Reinforcement Learning for Ultra-Reliable and Low-Latency IoV Communication Networks. *IEEE Trans. Veh. Technol.* 2019, 68, 4157–4169. [CrossRef]
29. Zhao, D.; Qin, H.; Song, B.; Zhang, Y.; Du, X.; Guizani, M. A Reinforcement Learning Method for Joint Mode Selection and Power Adaptation in the V2V Communication Network in 5G. *IEEE Trans. Cogn. Commun. Netw.* 2020, 6, 452–463. [CrossRef]
30. Technical Specification Group Radio Access Network; Study LTE-Based V2X Services; (Release 14), Document 3GPP TR 36.885 V14.0.0, 3rd Generation Partnership Project, June 2016. Available online: http://www.doc88.com/p-67387023571695.html (accessed on 23 February 2022).
31. Garcia, M.H.C.; Molina-Galan, A.; Boban, M.; Gozalvez, J.; Coll-Perales, B.; Sahin, T.; Kousaridas, A. A Tutorial on 5G NR V2X Communications. *IEEE Commun. Surv. Tutor.* 2021, 23, 1972–2026. [CrossRef]
32. Wu, W.; Liu, R.; Yang, Q.; Shan, H.; Quek, T.Q.S. Learning-Based Robust Resource Allocation for Ultra-Reliable V2X Communications. *IEEE Trans. Wirel. Commun.* 2021, 20, 5199–5211. [CrossRef]
33. Watkins, C.J.C.H.; Dayan, P. Q-learning. *Mach. Learn.* 1992, 8, 279–292. [CrossRef]
34. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
35. 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]
36. van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double Q-learning. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2094–2100.
37. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* 2016, arXiv:1609.04747. Available online: https://arxiv.org/abs/1609.04747 (accessed on 23 February 2022).