mathematics

MDPI

Article

Resource Allocation in V2X Communications Based on Multi-Agent Reinforcement Learning with Attention Mechanism

Yuanfeng Ding [1], Yan Huang [2], Li Tang [2], Xizhong Qin [1,*] and Zhenhong Jia [1]

[1] College of Information Science and Engineering, Xinjiang University, Urumqi 830000, China
[2] Network Department, China Mobile Communications Group Xinjiang Co., Ltd., Urumqi 830000, China
* Correspondence: qinxz@xju.edu.cn

Abstract: In this paper, we study the joint optimization problem of the spectrum and power allocation for multiple vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) users in cellular vehicle-to-everything (C-V2X) communication, aiming to maximize the sum rate of V2I links while satisfying the low latency requirements of V2V links. However, channel state information (CSI) is hard to obtain accurately due to the mobility of vehicles. In addition, the effective sensing of state information among vehicles becomes difficult in an environment with complex and diverse information, which is detrimental to vehicles collaborating for resource allocation. Thus, we propose a framework of multi-agent deep reinforcement learning based on attention mechanism (AMARL) to improve the V2X communication performance. Specifically, for vehicle mobility, we model the problem as a multi-agent reinforcement learning process, where each V2V link is regarded an agent and all agents jointly intercommunicate with the environment. Each agent allocates spectrum and power through its deep Q network (DQN). To enhance effective intercommunication and the sense of collaboration among vehicles, we introduce an attention mechanism to focus on more relevant information, which in turn reduces the signaling overhead and optimizes their communication performance more explicitly. Experimental results show that the proposed AMARL-based approach can satisfy the requirements of a high rate for V2I links and low latency for V2V links. It also has an excellent adaptability to environmental change.

Keywords: vehicle-to-everything; resource allocation; attention mechanism; multi-agent reinforcement learning; low latency

MSC: 94-05

Citation: Ding, Y.; Huang, Y.; Tang, L.; Qin, X.; Jia, Z. Resource Allocation in V2X Communications Based on Multi-Agent Reinforcement Learning with Attention Mechanism. *Mathematics* **2022**, *10*, 3415. https://doi.org/10.3390/math10193415

Academic Editor: Vladimir M. Vishnevsky

Received: 2 September 2022
Accepted: 16 September 2022
Published: 20 September 2022

1. Introduction

Vehicle-to-everything (V2X) communications is one of the key technologies in future autonomous driving and intelligent transport systems, aiming to enhance user experience, improve road safety, and adapt to complex and diverse transmission environments [1,2]. Among them, vehicle-to-infrastructure (V2I) mainly satisfies the requirements of vehicle users for high throughput, such as video traffic offloading [3]. Vehicle-to-vehicle (V2V), which focuses on the requirements of low latency and high reliability between vehicles, has become a key technology for cooperative driving and improved road safety [4,5].

V2X communication supports various use cases by exchanging information between infrastructure, vehicles, and pedestrians through various wireless technologies. Some candidate wireless technologies have been proposed, including dedicated short-range communication (DSRC), cellular vehicular communication, and 5G vehicular communication. DSRC technology is based on the IEEE 802.11p standard [6], which supports short exchanges between DSRC devices. To implement DSRC technology, the US Federal Communications Commission (FCC) has allocated 75 MHz of spectrum in the 5.9 GHz band.

DSRC technology can be used to improve road safety, e.g., collision warning [7]. However, it faces problems such as a limited communication range, large channel access delay, and high deployment costs. The cellular network has the advantages of high coverage, high network capacity, and supports high mobility. It helps to solve the drawbacks of DSRC. The 3rd Generation Partnership Project (3GPP) has completed Releases 14 [8] and 15 [9], where LTE-based V2X services are one of the main features. In Rel-14, V2X mainly provides data transmission services for road safety. Rel-15 supports advanced V2X scenarios, such as vehicle platooning and remote driving. However, one of the main challenges of LET-V2X is the requirement that the resources used by vehicular users should not conflict with cellular users in dense vehicular traffic scenarios requires. To further enhance V2X communication services, 3GPP has been formulated in Release 16 [10] for 5G-based V2X communication (5G-V2X). The 5G wireless system incorporates various emerging technologies, such as massive MIMO and millimeter-wave communication. Each of these technologies will bring various challenges to 5G-V2X [11]. In order to satisfy the stringent requirement in V2X communication, the V2X technology is required to provide both V2I and V2V communication using a shared resource pool. In addition, there is an increasing number of vehicular communication users, which will lead to a severe shortage of wireless resources. Therefore, how to coordinate interference and optimize resource allocation are important challenges in V2X communication.

1.1. Related Work

Currently, approaches to solve V2X resource allocation fall into two main categories: traditional optimization theory [12–17] and machine learning [18–29]. In traditional methods, the design objectives and corresponding constraints are built into an optimization problem for resource allocation and interference coordination. In [12], a delay expression was obtained by queuing analysis of packets and then resource allocation was performed using slowly varying large-scale fading channel information to satisfy the requirements of V2X for high capacity and low delay. Similarly, the author in [13] introduced the latency violation probability (LVP) as a constraint, which was accurately characterized by utilizing effective capacity theory. In [14], a novel scheme was proposed to reduce the delay of V2V links, which equated the original problem to the maximum weighted independent set problem with associated weights (MWIS-AW), and suggested a greedy cellular V2V link selection algorithm to solve the MWIS-AW problem. To allocate wireless resources more intelligently and rationally, [15] proposed an adaptive strategy based on fuzzy logic, which can dynamically adjust the parameters in the fuzzy system to ensure the full utilization of resources and quality-of-service (QoS) according to the network state. In [16], an interference hypergraph (IHG) was constructed to model the interference relationship among different vehicle users, and a cluster coloring algorithm was used to achieve effective and efficient resource allocation. In [17], a graph partitioning approach was developed to partition the high interference V2V links into different clusters and modeled the spectrum sharing problem as a weighted three-dimensional matching problem to solve the performance–complexity tradeoffs. However, in these schemes, it is difficult to build an accurate model to obtain accurate channel state information (CSI) due to the mobility of the vehicles. In particular, traditional methods are hard to adapt the network environment when it becomes more complex.

Recently, machine learning methods have been extensively applied to wireless communications to address the challenges faced by traditional optimization methods [18,19]. Especially, reinforcement learning has made significant progress in wireless resource management by interfacing with the environment and sensing environment changes to make decisions accordingly. In [20], a hybrid architecture of centralized decision-making and distributed resource sharing is proposed. A neural network first compressed CSI to reduce the signaling overhead and feedback to the central processor at the base station (BS). Then, a deep Q-network was used to allocate resources and sent the decision results to all vehicles. In [21], a dual time-scale federal deep reinforcement learning algorithm was proposed to

solve the joint optimization problem of C-V2X transmission mode selection and resource allocation. In [22], the age of information (AoI) was considered to study the delay problem of V2V links. To cope with the variation of vehicle mobility and information arrival time, the original MDP was decomposed into a series of MDPs for V2V pairs. An LSTM-based DRL algorithm was proposed to solve the local observability and high-dimensional disasters of V2V pairs. The authors of [23] introduced a centralized resource allocation architecture, and the base station uses a double deep Q network (DDQN) to allocation resources intelligently based on partial CSI to reduce the signaling overhead. Unlike [20–23], Refs. [24–29] modeled the V2X resource allocation problem as a multi-agent reinforcement learning (MARL) problem, where each V2V link was considered as an agent. In [24], a fingerprint-based deep Q-network was proposed to handle the non-smoothness problem in multi-agent reinforcement learning [25]. A centralized training and distributed execution framework were constructed for resource allocation. In the literature [26], both V2I link and V2V link latencies were considered in order to reduce the overall V2X latency. Moreover, proximal policy optimization (PPO)-based multi-agent reinforcement learning was proposed to optimize the objectives. To adapt to the changing environment more effectively, [27] proposed meta-reinforcement learning for V2X resource allocation. Firstly, spectrum resources and power are allocated using DQN and deep deterministic policy gradient (DDPG), respectively. Then, meta-learning was introduced to enhance the adaptability of the allocation algorithm to the dynamic environment. In [28], the congestion problem of wireless resources was under consideration, multi-agent reinforcement learning (DIRAL) based on unique state representation was proposed, and the nonstationary problem was solved by designing a view-based location. In addition, considering the topological relationship of vehicle users, [29] proposed a graph neural network (GNN)-based reinforcement learning method to learn the low-dimensional features of V2V link states by GNN and use RL for spectrum allocation. Although, the RL method has achieved satisfactory results in the problem of V2X resource allocation. It still faces two problems: firstly, there are difficulties in making effective sensing between each agent; secondly, the process of interfacing the agent with the environment will indiscriminately receive state information from all other agents, which will lead to a high computational overhead and signaling overhead.

1.2. Contribution and Organization

In this paper, we consider the resource management in partial CSI cases to match the realistic situation. In addition, a multi-agent reinforcement learning algorithm is utilized for adaption to the dynamic vehicle environment. We regard the V2V link as an agent and make corresponding decisions based on local observations. Furthermore, the agents have competitive and cooperative relationships in a multi-agent environment. In the case of competitive relationships, the V2V links tend to be egoistic, which ultimately affects the communication performance of the whole system. Hence, under the cooperative relationship setting, we build a reinforcement learning architecture and design the reward function to be a common reward for all agents. Considering the information exchange between agents, inspired by [30,31], we introduce an attention mechanism [32] for information exchange between V2V links. Through the attention mechanism, each agent can focus on more relevant information and optimize itself more explicitly. The main contributions of this paper are summarized as follows:

- Due to the mobility of vehicular users, it is not easy to obtain CSI accurately. We propose the framework of MARL to adapt to the changing environment and use only partial CSI for wireless resource allocation to ensure the high rate of V2I links and low latency of V2V links.
- To make each agent more effective in acquiring the state information of other agents in the environment and to establish collaborative relationships, we propose an algorithm of multi-agent deep reinforcement learning with attention mechanism (AMARL) to enhance the sense of collaboration among agents. It also enables agents to obtain more useful information, reduce the signaling overhead, and allocate resources more clearly.

- Experimental results demonstrate that, compared to other baseline schemes, the proposed AMARL-based algorithm can satisfy the requirement of low latency for V2V links and significantly increase the total rate of V2I links. It also has better adaptability to environmental changes.

The remainder of this paper is organized as follows. Section 2 presents the system model and problem formulation. Section 3 presents the details of the proposed attention mechanism-based MARL algorithm for solving V2X resource allocation. The simulation results and analysis are presented in Section 4. Section 5 presents the conclusions.

2. System Model

As shown in Figure 1, we consider cellular V2X communications in an urban road traffic scenario, including a base station and multiple vehicle users. In particular, we focus on mode 4 in the cellular V2X architecture [33], in which each vehicle can choose its communication resources without relying on the base station for resource allocation. According to the different service requirements of V2X communications, the vehicle users are divided into V2I and V2V links. Specifically, V2I links support higher-throughput tasks while V2V links can provide secure and reliable information to vehicle users through information sharing. In this paper, we consider the uplink for V2I communication and assume that all vehicle users have a single antenna for their transceivers. Meanwhile, to improve spectrum utilization and to guarantee the high-rate requirements of the V2I link, we assume that each V2I is pre-allocated an orthogonal sub-band with a fixed transmit power and shares this sub-band resource with multiple V2V links. In addition, each V2V pair can only select one sub-band for communication.

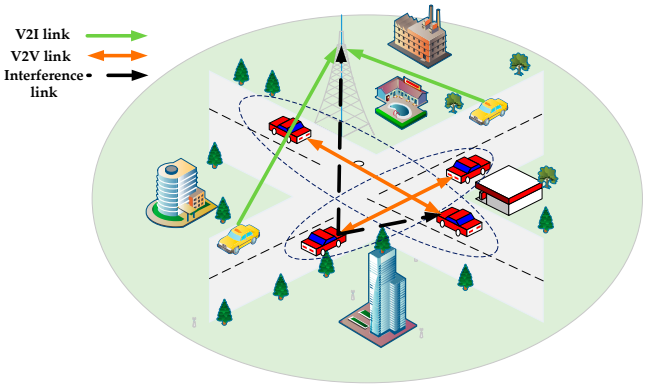


Figure 1. V2X communication scenarios.

We denote the V2I links and V2V links as the sets $\mathcal{M} = \{1, \cdots, M\}$ and $\mathcal{N} = \{1, \cdots, N\}$, respectively, where M and N denote the number of V2I and V2V, respectively. In addition, we assume that the number of sub-bands equals to the number of V2I links.

In this paper, the channel power gain considered includes the large-scale fading component and the small scale-fading component. The channel gain can be expressed as $g = \alpha\beta$, where α and β denote the large-scale fading and the small-scale fading, including the path loss and shadowing for each communication link, respectively. We define the channel power gains of the m-th. V2I link and the n-th V2V link on the m-th sub-band as $\hat{g}_{m,B}$ and $g_n[m]$, respectively. The interfering channel gains received at the receiver of the n-th V2V link from the transmitter of the m-th V2I link and the n'-th V2V link over the m-th sub-band are given by $\hat{g}_{m,n}$ and $g_{n',n}$. The interfering channel gain for the m-th V2I link from the n-th V2V link over the m-th sub-band is $g_{n,B}[m]$. For simplicity, the notations adopted in this paper are listed in Table 1.

Table 1. Key mathematical symbols.

Symbols	Definition
$\mathcal{M}, \mathcal{N}$	Set of V2I links and V2V links
M, N	Numbers of V2I links and V2V links
$\hat{g}_{m,B}$	Channel gain from the m-th V2I link to BS
$g_n[m]$	Channel gain between the n-th V2V link
$\hat{g}_{m,n}$	The interfering channel gain from m-th V2I link to n-th V2V link
$g_{n',n}$	The interfering channel gain from n'-th V2V link to n-th V2V link
$g_{n,B}[m]$	The interfering channel gain from n-th V2V link to m-th V2V link
$x_n[m]$	Indicator of the n-th V2V link reuse the spectrum of the m-th V2I link
$\gamma_m^{V2I}[m]$	The SINR of m-th V2I link
$\gamma_n^{V2V}[m]$	The SINR of n-th V2V link
σ^2, W	Noise power and bandwidth
$p_m^{V2I}, p_n^{V2V}[m]$	Transmit power of the m-th V2I link and the n-th V2V link
Δ_T	The coherence time of the channel
$\omega_{i,j}$	The attention weight of $V2V_i$ to $V2V_j$
R_t	Reward function
$Q_n(s, a, \theta)$	Q-network of the n-th V2V link
θ	Parameter of the Q-network
$Q_n(o', a'; \theta^{tar})$	Target Q-network of n-th V2V link
D_n	Mini-batch of experiences
ϵ	Exploration rate
γ	Discount factor

The received signal to interference plus noise (SINR) of the m-th V2I link and the n-th V2V link over the m-th sub-band can be expressed as:

$$\gamma_m^{V2I}[m] = \frac{\hat{g}_{m,B} \cdot p_m^{V2I}}{\sum_{n=1}^{N} x_n[m] \cdot g_{n,B}[m] \cdot p_n^{V2V}[m] + \sigma^2} \tag{1}$$

and:

$$\gamma_n^{V2V}[m] = \frac{g_n[m] \cdot p_n^{V2V}[m]}{I_n[m] + \sigma^2} \tag{2}$$

where p_m^{V2I} and $p_n^{V2V}[m]$ denote the transmit power of the m-th V2I link and the n-th V2V link at the m-th sub-band, σ^2 denotes the noise power, and:

$$I_n[m] = \hat{g}_{m,n} \cdot p_m^{V2I} + \sum_{n' \neq n}^{N} x_{n'}[m] \cdot g_{n',n}[m] \cdot p_{n'}^{V2V}[m] \tag{3}$$

denotes the total interference power of the n-th V2V link in the m-th sub-band. The binary variable $x_n[m] \in \{0, 1\}$ denotes the spectrum allocation indicator, if $x_n[m] = 1$ means the n-th V2V link uses the m-th sub-band. Otherwise, $x_n[m] = 0$. We assume that a V2V link only accesses one sub-band, $\sum_{m=1}^{M} x_n[m] \leq 1$ is satisfied. Then, the capacity of the m-th V2I link and the n-th V2V link is:

$$R_m^{V2I} = W \log\left(1 + \gamma_m^{V2I}[m]\right) \tag{4}$$

and:

$$R_n^{V2V}[m] = \sum_{m=1}^{M} x_n[m] \cdot W \log\left(1 + \gamma_n^{V2V}[m]\right) \tag{5}$$

where W is the bandwidth of the sub-band.

This paper aims to maximize the V2I link capacity to provide high-quality entertainment services while satisfying the low latency and high reliability requirements of V2V links to provide realistic and reliable information to vehicle users in road traffic. To satisfy the first requirement, the sum rate of V2I links needs to be maximized. To satisfy the second

requirement, we require V2V users to successfully transmit packets of size B in finite time T_{max} with the following probabilistic model:

$$\Pr\left\{\sum_{t=1}^{T_{max}} R_n^{V2V}[m,t] \geq \frac{B}{\Delta_T}\right\} \tag{6}$$

where Δ_T is the coherence time of the channel, and the index t is added in $R_n^{V2V}[m,t]$ to indicate the capacity of the n-th V2V link at different coherence time slots. Thus, the problem of V2X resource allocation can be formulated as an optimization problem as follows:

$$\max_{x,P^{V2V}} \sum_{m=1}^{M} R_m^{V2I} \tag{7}$$

$$s.t. \sum_{m=1}^{M} x_n[m] \leq 1 \tag{8}$$

$$p_n^{V2V}[m] \in P, \ \forall n, m \tag{9}$$

where P denotes the discrete power set of V2V link. Constraint (8) denotes that each V2V pair can occupy only one sub-band, and constraint (9) denotes the power condition is satisfied.

Problem (7) is a combinatorial optimization problem, and a limitation of traditional optimization methods is the high requirement for model accuracy. However, due to vehicle mobility, the environment is constantly changing, leading to uncertainty in the model parameters, and the complete CSI is difficult to obtain and solve by traditional methods. Therefore, we propose to address this problem through a deep reinforcement learning approach. In Section 4, we validate the effectiveness of the proposed method.

3. Resource Allocation Based on Multi-Agent Reinforcement Learning with Attention Mechanism Algorithm

In this section, we briefly introduce the basic concepts of attentional mechanism and multi-agent reinforcement learning and then describe how the algorithmic framework can be used to solve the problem of V2X resource allocation. Before presenting the algorithm in detail, we first introduce the three elements in reinforcement learning: the observation space, the action space, and the reward function.

3.1. Design of Three Elements

3.1.1. Observation Space

Due to the existence of vehicle mobility, it is more difficult to obtain a complete CSI. Therefore, we consider partial CSI as part of the observation space, which, on the one hand, is more in line with the real scenario; on the other hand, it is also beneficial to reduce the signaling overhead of CSI feedback. In mode 4, the vehicle performs wireless resource allocation by sensing channel measurements, in which it will inevitably receive interference information. Considering the need for low latency in V2V links, the state observation space of the V2V agent should also contain the remaining payload and time. Thus, the state of the V2V agent at the time t includes the received interference information, the remaining payload, and the remaining time.

We denote the observation space as: $O = \left\{o_t^1, \cdots, o_t^n, \cdots, o_t^N\right\}$, which is the set of all agents' states at moment t. o_t^n is the observation of the n-th agent at each time slot t. The remaining payload and remaining time are defined as L_t^n, U_t^n, respectively. Therefore, o_t^n can be expressed as:

$$o_t^n = \left\{\left\{I_{t-1}^n[m]\right\}_{m\in M}, L_t^n, U_t^n\right\} \tag{10}$$

3.1.2. Action Space

Based on the observed state, each V2V agent will make a decision on sub-band selection and power allocation. We define the action space for all V2V agents as $A = \{a_n\}_{n=1}^{N}$, where $a_n = \{x_n, p_n\}$ is the action space of the n-th V2V agent. x_n and p_n denote the set of possible sub-band selection and power allocation for the n-th V2V agent. Thus, the set of possible sub-band assignment decisions for the n-th V2V agent at the time slot t can be defined as:

$$x_t^n = \{x_t^n[1], \cdots, x_t^n[m], \cdots, x_t^n[M]\} \tag{11}$$

In problem (7), we carry out a discrete power allocation scheme [34]. The set of possible power selection of the n-th V2V agent at time slot t can be expressed as:

$$p_t^n \in \left\{0, \tfrac{1}{N-1}P_{max}, \tfrac{2}{N-1}P_{max}, \cdots, P_{max}\right\} \tag{12}$$

where N is the number of power levels.

3.1.3. Rewards Function

The design of the rewards function is closely related to the problem (7). Our objective is to maximize the total throughput of the V2I links while satisfying the latency and reliability requirements of the V2V links. In order to satisfy the requirement of the low latency of the V2V links, we set the following reward function:

$$G_t^n = \begin{cases} R_n^{V2V}(t), & L_t^n \geq 0 \\ c, & L_t^n < 0 \end{cases} \tag{13}$$

This means that we want the V2V link to complete the data transfer as quickly as possible. When there is a remaining load, the transmission is carried out at the effective rate of the V2V link until the load is fully transmitted. c is a hyperparameter, which is greater than the maximum possible V2V links rate, and the faster the remaining load is sent, the greater the reward. In addition, we want the transmission time to be as short as possible, which means that the probability of successful packet transmission within a given time constraint will increase. Therefore, the final reward function is set as follows:

$$R_t = c_1 \sum_{m=1}^{M} R_m^{V2I}(t) + c_2 \sum_{n=1}^{N} G_t^n - c_3(T_{max} - U_t^n) \tag{14}$$

where $\{c_i\}_{i=1,2,3}$ is a weighting factor, which reflects the degree of requirement for different QoS.

3.2. Algorithmic Framework

3.2.1. Overview of Attentional Mechanism

We consider that in the problem of V2X resource allocation, the interaction between V2V pairs affects their respective communication performance. If each V2V pair receives the state information of all other V2V pairs, it will lead to two problems. Firstly, mixing valuable and useless information would lead to problematic performance optimization; secondly, processing global information by V2V pair would require a large number of computational resources and a high signaling overhead, which is unacceptable. Therefore, to solve the above two problems, we introduced the attention mechanism based on reinforcement learning, which evaluates the importance of state information through attention weights and enables V2V pairs to obtain helpful information better.

We define the state information of the i-th V2V pair as $s_i (i \in \mathcal{N})$, and the corresponding query Q_i, key K_i, and value V_i, and then define several basic parameter matrices used to

describe the attention mechanism, namely the query matrix W^Q, the key matrix W^K, and the value matrix W^V. Thus, the attention weight of $V2V_i$ to $V2V_j$ is:

$$\omega_{i,j} = softmax\left(\frac{Q_i \cdot K_j^T}{\sqrt{d_k}}\right) \tag{15}$$

where d_k denotes the key dimension of each component.

The state information after passing the attention mechanism is then obtained by calculating a weighted sum of the values of the other V2V pairs, which is represented as:

$$s_i^A = \sum_{i \neq j} \omega_{i,j} \cdot V_i \tag{16}$$

3.2.2. Multi-Agent Reinforcement Learning

In multi-agent reinforcement learning, multiple agents are in the same environment. Each agent independently intersects with the environment to motivate it and uses the reward of feedback to improve its policy for higher rewards continuously. Furthermore, an agent's policy not only depends on its state and actions but also considers the states and actions of other agents, as shown in Figure 2.

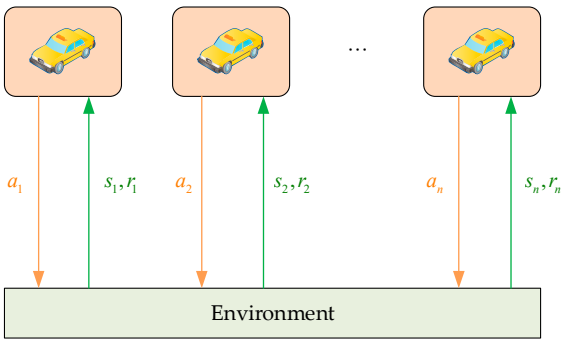


Figure 2. MARL framework.

3.2.3. AMARL Algorithm

In this section, we develop an attentional DRL-based algorithmic framework to solve the problem (7). As shown in Figure 3, we consider each V2V link as an agent body and model the resource management problem as an MDP, where all vehicles are in the same wireless environment. Each agent independently intersects with the environment to obtain its local observations and obtains information from other agents through the attention mechanism to allocate spectrum and power based on its observations.

To achieve the goal of maximizing the rate of V2I links and satisfying the low latency of V2V links, we construct an algorithmic framework with a Deep Q Network (DQN) as the backbone network and use a distributed architecture to solve the problem (7), where each agent has its Q network and optimizes its policy in this way. We consider the allocation of wireless resources within time and the introduction of an attention mechanism to sense changes in vehicle state information due to environmental changes.

With the introduction of the attention mechanism, the V2V links pay more attention to helpful information and integrate this information into its action value estimation function, i.e., the Q function, which can be expressed as:

$$Q_n(s, a, \theta) = f_n\left(add\left(s_n^A, s_n\right)\right) \tag{17}$$

The calculation process is shown in Figure 4, where $add\left(s_n^A, s_n\right) = s_n^A + s_n$, f_n is a three-layer multi-layer perceptron (MLP), s^A is the output state of the agent after the attention network, and θ is a parameter of the network.

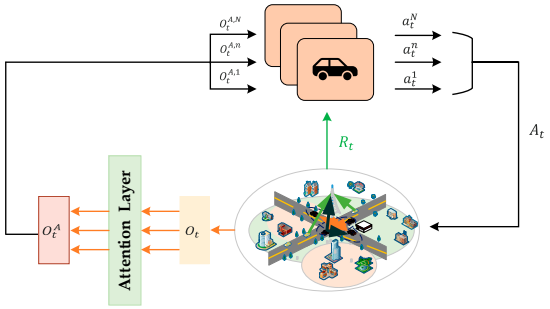


Figure 3. The basic AMARL for V2X communications.

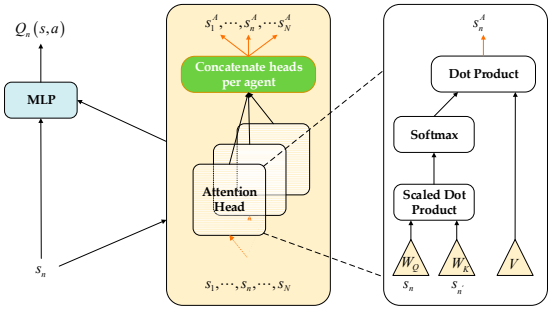


Figure 4. Calculating the Q value for agent n.

To obtain the optimal policy π, the optimal action value function is defined:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{18}$$

From the Bellman optimality equation [35], Equation (16) can be written as

$$Q^*(s_t,a_t) = \mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,a_t)}\left[R_t + \gamma \max_{a\epsilon A} Q^*(s_{t+1},a)\middle| s_t = s, a_t = a\right] \tag{19}$$

where γ is the discount factor. From the Monte Carlo approximation, (17) can be transformed into:

$$Q^*(s_t,a_t) \approx r_t + \gamma \max_{a\in A} Q^*(s_{t+1},a) \tag{20}$$

approximating the value of Q in (20) with the Q network yields:

$$Q(s_t,a_t;\theta) \approx r_t + \gamma \max_{a\in A} Q(s_{t+1},a;\theta) \tag{21}$$

where the left-hand side of Equation (19) is the prediction of the Q network at moment t and the TD target [36] on the right-hand side is the prediction of the Q network at moment $t+1$, denoted as $y_t = r_t + \gamma \max_{a\in A} Q(s_{t+1},a;\theta)$. Thus, the loss function can be defined as:

$$loss = [Q(s_t,a_t;\theta) - y_t]^2 \tag{22}$$

The training of DQN can be divided into two parts: collecting the training data and updating the parameters θ.

(1) Collecting training data:

The n-th V2V link needs to intersect with the environment using some kind of strategy π, which we for call a behavioral policy. The ϵ-greed policy is generally used [37]:

$$a = \begin{cases} \arg\max_{a\in A} Q_n(o',a;\theta), with\ probability\ 1-\epsilon \\ random\ action, with\ probability\ \epsilon \end{cases} \tag{23}$$

the V2V link performs an action that leads to a change in the environment, which we refer to as a the trajectory of episode, is written as: $o_1^n, a_1^n, r_1^n, \cdots, o_t^n, a_t^n, r_t^n \cdots$, and is stored in an array as a four-tuple $(o_t^n, a_t^n, r_t^n, o_{t+1}^n)$, called the experience replay array D.

(2) Updating parameters:

A mini-batch of experiences D_n are uniformly sampled from the experience replay array D to update parameter θ using stochastic gradient descent. The TD algorithm is used to train the DQN network; however, maximization in the TD algorithm leads to an overestimation problem, where the TD target overestimates the true value. To alleviate this problem, a target network [38] is used to calculate the TD target, i.e.,

$$y_t{}' = r_t + \gamma \max_{a'} Q_n\left(o', a'; \theta^{tar}\right) \tag{24}$$

Therefore, the loss function is:

$$L_n(\theta) = \frac{1}{2D_n} \sum_{t \in D_n} \left[y_t{}' - Q_n(o, a, \theta)\right]^2 \tag{25}$$

Notation:

$$\delta_t = y_t{}' - Q_n(o, a, \theta) \tag{26}$$

is the TD error. Perform gradient descent to update the network parameters:

$$\theta \leftarrow \theta - \alpha \cdot \sum_{t \in D_n} \delta_t \cdot \nabla_\theta \, Q_n^\pi(o, a, \theta) \tag{27}$$

where θ^{tar} is the target network parameter, which is periodically updated by the Q-network parameter θ to improve the stability of the network. The training process is summarized in Algorithm 1.

Algorithm 1 Training Process

1: Input: V2X environment simulator, Attention network model, DQN model, payload size, and maximum tolerant latency
2: Output: AMARL network's weight
3: Initialize: experience replay array, the parameters of DQN and target DQN
4: **for** each episode $i = 1, 2, \cdots$ **do**
5: Update environment;
6: Reset remaining payload L_t^n and remaining time U_t^n;
7: **for** each step $t = 1, 2, \cdots$ **do**
8: Observed state of all V2V agents: $o_t = \{o_t^n\}_{n=1,\ldots}$;
9: Through the attention network: $o_t^A = \left\{o_t^{A,n}\right\}_{n=1,\ldots}$;
10: **for** each V2V agent $n = 1, 2, \cdots$ **do**
11: Based on add $\left(o_t^{A,n}, o_t\right)$ select action a_t^n according to the ϵ-greed policy;
12: **end for**
13: All agents take actions and gain shared reward R_t;
14: Update environment;
15: **for** each V2V agent $n = 1, 2, \cdots$ **do**
16: Gain the next moment of observation: o_{t+1}^n;
17: Store $\left(o_t^n, a_t^n, r_t^n, o_{t+1}^n\right)$ in the experience replay array;
18: **end for**
19: **end for**
20: **for** each V2V agent $n = 1, 2, \cdots$ **do**
21: Sample a mini-batch experiences D_n from experience replay array D;
22: Update DQN parameter θ according to (25);
23: Update the target DQN every k steps: $\theta^{tar} = \theta$;
24: **end for**
25: **end for**

In the test phase, at each time step t, each V2V agent compiles the observed states. Then, it selects an action with the maximum Q value based on the trained Q-network. After that, all V2V links determine the power and sub-band for transmission by the selected actions. The testing procedure is summarized in Algorithm 2.

Algorithm 2 Testing Process

1: Input: V2X environment simulator, AMARL network model
2: Output: All V2V agents actions
3: Start: Load AMARL network model, Start V2X environment simulator
4: **for** each episode $i = 1, 2, \cdots$ **do**
5: Update environment;
6: Reset remaining payload L_t^n and remaining time U_t^n;
7: **for** each step $t = 1, 2, \cdots$ **do**
8: Observed state of all V2V agents: $o_t = \{o_t^n\}_{n=1,\ldots}$;
9: Through the attention network: $o_t^A = \left\{o_t^{A,n}\right\}_{n=1,\cdots}$;
10: **for** each V2V agent $n = 1, 2, \cdots$ **do**
11: Compile the state observation space $o_t^{A,n}$ and select the action with the maximum Q value based on the trained Q network;
12: **end for**
13: **end for**
14: **end for**

4. Simulation Results

In this section, we verify the feasibility of the proposed algorithm in V2X resource allocation through simulation experiments. We follow the city case simulation in 3GPP TR36.885 [39] (including density, speed, vehicle channel, V2V data traffic, etc.) and follow the set values of the main parameters in [24] to train the model. According to [39,40], we generate V2X communication scenarios and datasets by Python. The main simulation parameters are given in Table 2, and the channel models for the V2V link and V2I link are given in Table 3.

Table 2. The main simulation parameters.

Parameters	Values
Carrier frequency	2 GHz
Sub-channel bandwidth	1 MHz
BS antenna height	25 m
BS antenna gain	8 dBi
BS receiver noise figure	5 dB
Vehicle antenna height	1.5 m
Vehicle receiver gain	3 dBi
Vehicle receiver noise figure	9 dB
Vehicle speed	[10, 15] m/s
V2I transmission power	35 dBm
V2V Maximum transmission power	33 dBm
Noise power σ^2	-114 dBm
Maximum tolerant latency of V2V links	100 ms
V2V payload size B	$[1, 2, \cdots, 6] \times 1060$ bytes
Number of V2I links	4
Number of V2V links	4
Discount factor γ	0.9
Reward weights $\{c_i\}_{i=1,2,3}$	{0.1, 0.9, 1.0}
Power levels N_p	5

Table 3. The channel models for the V2V link and V2I link.

Parameters	V2I Link	V2V Link
Path loss model	$128.1 + 37 \log_{10} d$, d in km	LOS in WINNER + B1 Manhattan [40]
Shadowing distribute	Log-normal	Log-normal
Shadowing standard deviation	8 dB	3 dB
Decorrelation distance	50 m	10 m
Fast fading	Rayleigh fading	Rayleigh fading
Fast fading updata	Every 1 ms	Every 1 ms

In building the DQN for each agent, we constructed three fully connected layers containing 250, 180, and 100 neurons, respectively. The activation function of the hidden layer in the DQN used the ReLu $f(x) = \max(0, x)$, the RMSProp optimizer was used to update the network parameters, and the learning rate $\alpha = 0.001$. In the training phase, similar to [24], we fix the payload of V2V pairs to be 2×1060 bytes, train a total of 3000 episodes of Q-network for each agent, and the exploration rate ϵ is linearly annealed from 1 to 0.2. In the testing phase, we vary the payload and speed of V2V pairs to verify the adaptability of the proposed scheme to the environment.

In order to verify the validity of the proposed method, we compared it with the following four methods:

(1) Meta-reinforcement learning [27]: In this scheme, DQN is used to solve the problem of spectrum allocation, deep deterministic policy gradient (DDPG) is used to solve the problem of continuous power allocation, and meta-learning is introduced to make the agent adapt to the changes in the environment.
(2) Proposed RL (no attention): This scheme does not incorporate an attention mechanism, and the agent will obtain the state information of other agents without any difference and then allocate wireless resources.
(3) Brute-Force: This scheme is implemented in a centralized manner and requires accurate CSI. It focuses only on improving the performance of V2Vs, ignoring the need for V2I links, and performs an exhaustive search of the action space of all V2V pairs to maximize V2Vs and rates.
(4) Random: randomizes spectrum and power allocation.

4.1. Impact of Payload Size on Network Performance

Figure 5 shows the change in the sum rate of the V2I links, and the probability of successful transmission of the V2V links as the payload changes. In particular, based on the maximum V2V links transmission power of 23 dBm set in [24], we use this power as a lower limit for this paper's transmission power. As can be seen from Figure 5, the sum rate of the V2I link and the probability of successful transmission of the V2V link decrease for all schemes (except Brute-Force) as the V2V links payload increases. This is because, when the payload increases, the V2V links require more transmission time and higher transmission power, which causes more interference in the V2I and V2V links, resulting in a decreased communication performance. Compared to the meta-reinforcement learning scheme, Figure 5a shows that the proposed scheme maintains the higher sum rate of the V2I links as the payload increases and is close to the Brute-Force scheme. Even when the transmission power of the V2V links is set to a minimum of 23 dBm, the proposed scheme still has a much better V2I links sum rate than the meta-reinforcement learning scheme. In Figure 5b, the successful transmission probability of V2V links for different methods are compared. The performance of the proposed method is close to that of the meta-reinforcement learning method using full CSI when partial CSI is utilized. This indicates that the proposed algorithm can achieve the expected requirements of V2V link delay with a low signaling overhead. Figure 5 also shows the robustness of the proposed algorithm to the variation of the payload of V2V links.

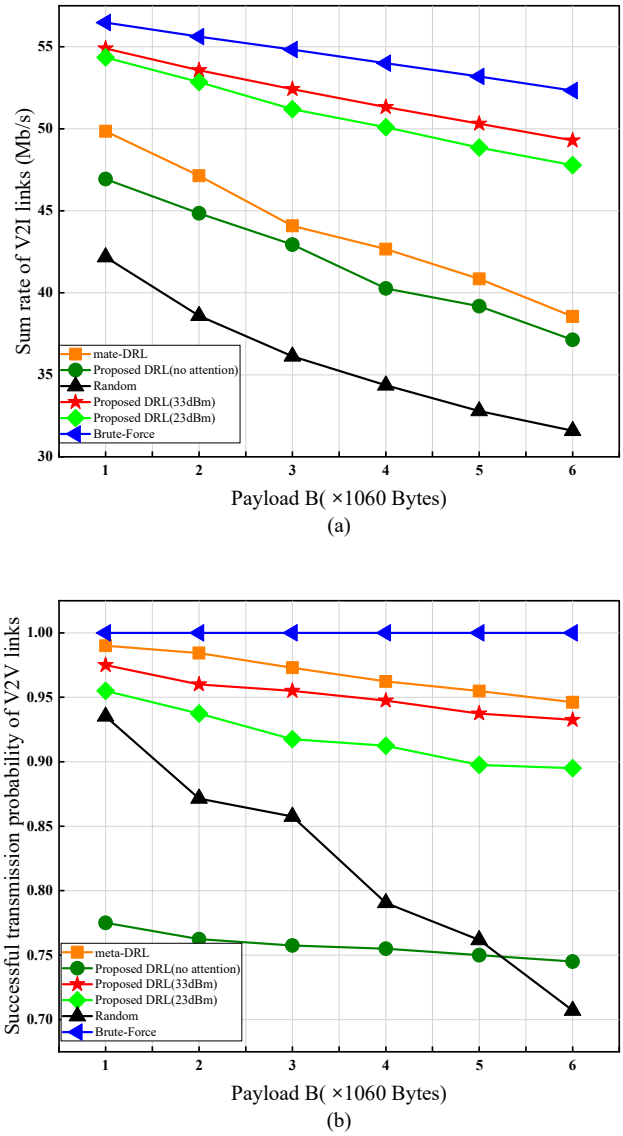


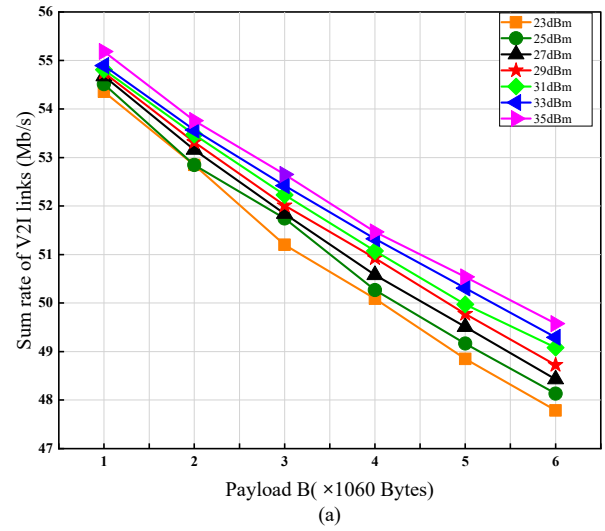
Figure 5. The performance for different payload sizes: (**a**) Sum rate of V2I links. (**b**) Successful transmission probability of V2V links.

Furthermore, we observe the proposed algorithm's performance before and after the introduction of the attention mechanism. The communication performance is substantially improved after the attention mechanism's introduction. Before the attention mechanism, V2V links indiscriminately obtained the state information of other V2V links, enhancing the interference level and increasing the signaling overhead. Moreover, with the introduction of the attention mechanism, a collaborative relationship is built between V2V links, allowing better use of information from other V2V links for more effective interference coordination, thus improving the communication performance.

4.2. Impact of V2V Links Transmission Power on Network Performance

In this subsection, we investigate the impact of the V2V links' power variations on the network performance to find a low-power design solution that satisfies the performance requirements. As shown in Figure 6, we set the maximum transmission power of the V2V links to {23, 25, 27, 29, 31, 33, 35} dBm. As the payload increases, the performance at all set powers decreases. Figure 6a shows that with the same load, the sum rate of the V2I links increases as the transmission power of the V2V links increases, and the performance at all powers is relatively similar. Similarly, Figure 6b shows that the probability of successful

transmission of the V2V links also increases with increasing power, which is due to the fact that as the power of the V2V link increases, the rate of the V2V links becomes larger as the transmission time decreases. In addition, we found that when the power of the V2V links is set to 35 dB, the probability of successful transmission of the V2V links decreases by 5.25% with the payload increase, although the network performance will still improve. Moreover, when the maximum power is 33 dBm, the decline in the successful transmission probability is 4.25% and approaches the performance of a maximum power of 35 dBm. Compared with other power settings, the performance of a maximum power of 33 dBm still has an advantage. This provides some reference for practical engineering design. If only the high throughput of the V2I link is required, setting the maximum power of the V2V links to 23 dBm is sufficient and reduces power consumption.



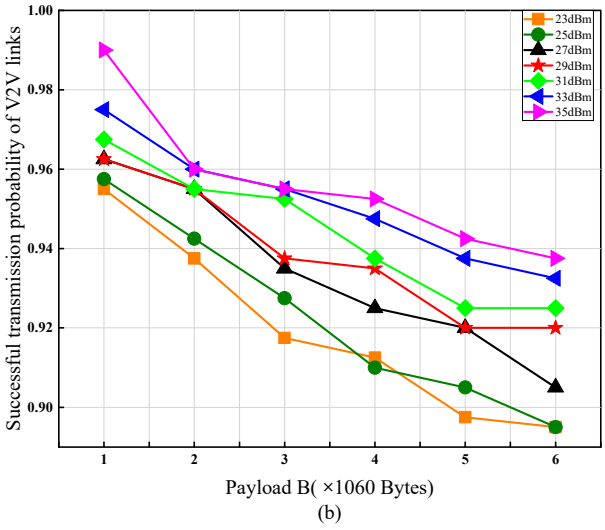


Figure 6. The performance comparisons for different V2V links transmission power: (**a**) Sum rate of V2I links. (**b**) Successful transmission probability of V2V links.

4.3. Impact of Vehicle Velocity on Network Performance

To further investigate the adaptability of the proposed algorithm to environmental changes, we investigate the effect of different vehicle speeds on the network performance. In the training phase, the speed was fixed at [10, 15] m/s to verify the robustness of the proposed algorithm. As shown in Figure 7, the performance of the proposed algorithm decreases with increasing speed for the same payload. This is because the environment

changes more significantly as the vehicle speed increases, increasing the difficulty of obtaining state information and the uncertainty of the state information. However, the proposed scheme can still maintain a high throughput of the V2I links, and the probability of successful transmission of the V2V links, which indicates that the proposed algorithm can adapt to the changes in the environment.

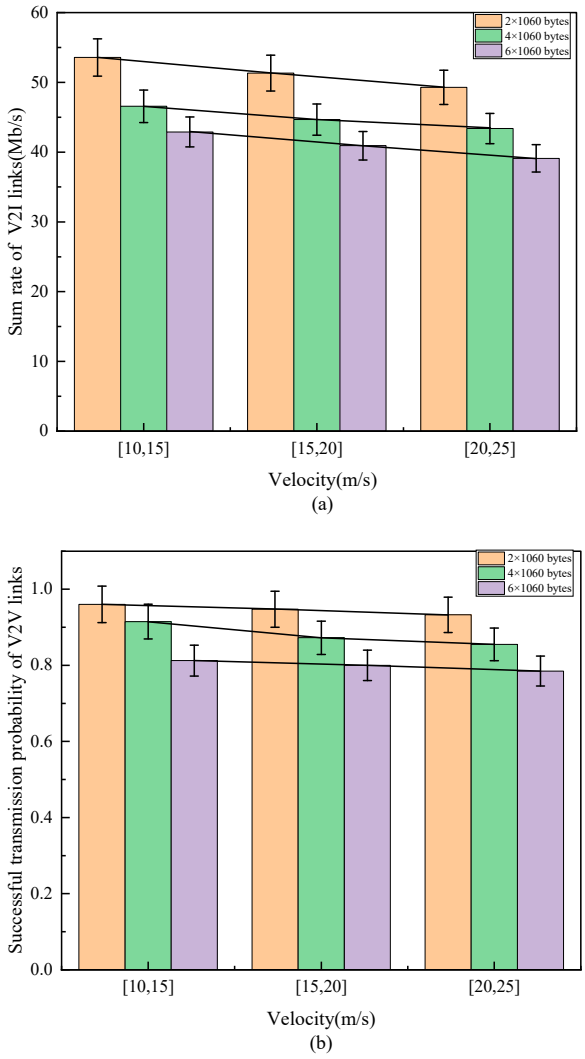


Figure 7. The performance comparison different velocity: (**a**) Sum rate of V2I links. (**b**) Successful transmission probability of V2V links.

Further, we investigated the effectiveness of the proposed AMARL algorithm. As shown in Figure 8, we fixed a payload of 2 × 1060 Bytes and compared the network performance of the AMARL algorithm with the MARL algorithm (no attention). Figure 8a shows that the sum rate of the V2I links using the AMARL algorithm is higher than that of the MARL algorithm in a low-speed environment. As the speed increases, the proposed algorithm is slightly higher than the MARL algorithm. For practical design reasons, the proposed algorithm will be chosen over the MARL algorithm in low-speed environments where higher throughput of the V2I links is required. In high-speed environments, the MARL algorithm may be better; its network performance can satisfy the throughput requirements of some V2I links with a lower computational overhead than the proposed algorithm. Overall, the network performance of the proposed algorithm is better than the MARL algorithm.

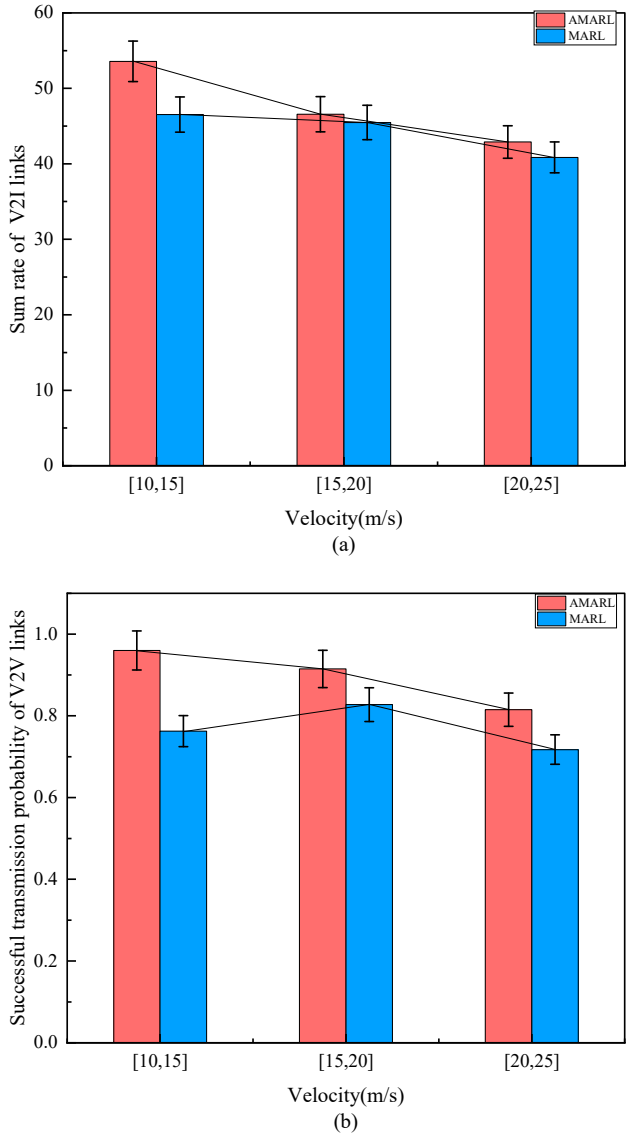


Figure 8. The performance comparison between AMARL and MARL: (**a**) Sum rate of V2I links. (**b**) Successful transmission probability of V2V links.

Figure 8b shows the effect of the vehicle speed variation on the successful transmission probability of the V2V link. It can be seen from the figure that the proposed algorithm outperforms the MARL algorithm. Even at the highest vehicle speed, the minimum successful transmission probability of the proposed algorithm is close to the highest successful transmission probability of the MARL algorithm. This is due to the introduction of the attention mechanism. Specifically, introducing the attention mechanism will promote the cooperative relationship of V2V links and reduce unnecessary communication interference by obtaining valid information, thus improving the throughput of V2V links and reducing the data transmission time.

The demonstration of the network performance metrics in Figure 8 verifies that the proposed approach can adapt to environmental changes and demonstrates the attention mechanism's effectiveness.

5. Conclusions

In this paper, we propose an attention-based multi-agent reinforcement learning algorithm for spectrum and power allocation of V2X, aiming to satisfy the requirements of high throughput for V2I links and low latency for V2V links. Meanwhile, we used partial CSI for

training to reduce the signaling overhead. The attention mechanism's introduction enables more efficient information exchange between V2V links and more explicit optimization of their own policies. The simulation results demonstrate the effectiveness of the proposed scheme, and our model can achieve the expected network performance and adapt better to environmental changes. We also explored the impact of power variation on network performance, which provides a reference for practical engineering design. However, our work also has shortcomings. We did not further consider effective interactions between vehicles and the environment. In this way, it may be impossible to ensure that the strategies trained by reinforcement learning satisfy the practical needs. Therefore, in future work, it is hoped that the process of vehicle–environment intercommunication will consider the expert knowledge of the environment (e.g., Channel Knowledge Map (CKM) [41]). CKM is a site-specific database tagged with the transmitter/receiver locations, which contains useful CSI to help enhance environmental awareness and avoid complex real-time CSI acquisition. In addition, the use of the proposed scheme for MIMO-V2X is a worthwhile research direction in order to further satisfy the high spectral efficiency gain and high data rate requirements.

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

Funding: This paper is supported by the following two project funds: "The ICP Household Scheduling Analysis Service Project for Xinjiang Mobile Communications Group". "Research and Development of Key Technologies and Application Demonstration of Integrated Food Data Supervision" Platform in Xinjiang Region. 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. Ko, S.-W.; Chae, H.; Han, K.; Lee, S.; Seo, D.-W.; Huang, K. V2X-Based Vehicular Positioning: Opportunities, Challenges, and Future Directions. *IEEE Wirel. Commun.* **2021**, *28*, 144–151. [CrossRef]
2. Rahim, N.A.; Liu, Z.; Lee, H.; Khyam, M.O.; He, J.; Pesch, D.; Moessner, K.; Saad, W.; Poor, H.V. 6G for Vehicle-to-Everything (V2X) Communications: Enabling Technologies, Challenges, and Opportunities. *Proc. IEEE* **2022**, *110*, 712–734. [CrossRef]
3. Pervez, F.; Yang, C.; Zhao, L. Dynamic Resource Management to Enhance Video Streaming Experience in a C-V2X Network. In Proceedings of the 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), online, 18 November–16 December 2020; pp. 1–5. [CrossRef]
4. Coll-Perales, B.; Schulte-Tigges, J.; Rondinone, M.; Gozalvez, J.; Reke, M.; Matheis, D.; Walter, T. Prototyping and Evaluation of Infrastructure-Assisted Transition of Control for Cooperative Automated Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 6720–6736. [CrossRef]
5. Eiermann, L.; Wirthmuller, F.; Massow, K.; Breuel, G.; Radusch, I. Driver Assistance for Safe and Comfortable On-Ramp Merging Using Environment Models Extended through V2X Communication and Role-Based Behavior Predictions. In Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 3–5 September 2020; pp. 263–270. [CrossRef]
6. Zhou, H.; Xu, W.; Chen, J.; Wang, W. Evolutionary V2X Technologies Toward the Internet of Vehicles: Challenges and Opportunities. *Proc. IEEE* **2020**, *108*, 308–323. [CrossRef]
7. Jiang, D.; Taliwal, V.; Meier, A.; Holfelder, W.; Herrtwich, R. Design of 5.9 ghz dsrc-based vehicular safety communication. *IEEE Wirel. Commun.* **2006**, *13*, 36–43. [CrossRef]
8. *3GPP TR 21.914*; Digital Cellular Telecommunications System (Phase2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; 5G.; Version 14.0.0 Release 14. ETSI: Valbonne, France, 2018.
9. *3GPP TR 21.915*; Study on New Radio (NR) Access Technology; Version 1.1.0 Release 15. ETSI: Valbonne, France, 2018.
10. *TS 23.287*; Architecture Enhancements for 5G System (5GS) to Support Vehicle-To-Everything (V2X) Services, 3GPP, V16.4.0 (Release 16). ETSI: Valbonne, France, 2020.

11. 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]
12. 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]
13. Guo, C.; Liang, L.; Li, G.Y. Resource Allocation for Low-Latency Vehicular Communications: An Effective Capacity Perspective. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 905–917. [CrossRef]
14. 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]
15. Zhang, M.; Dou, Y.; Chong, P.H.J.; Chan, H.C.B.; Seet, B.-C. Fuzzy Logic-Based Resource Allocation Algorithm for V2X Communications in 5G Cellular Networks. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 2501–2513. [CrossRef]
16. Chen, C.; Wang, B.; Zhang, R. Interference Hypergraph-Based Resource Allocation (IHG-RA) for NOMA-Integrated V2X Networks. *IEEE Internet Things J.* **2019**, *6*, 161–170. [CrossRef]
17. Liang, L.; Xie, S.; Li, G.Y.; Ding, Z.; Yu, X. Graph-Based Resource Sharing in Vehicular Communication. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 4579–4592. [CrossRef]
18. Zappone, A.; Di Renzo, M.; Debbah, M. Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both? *IEEE Trans. Commun.* **2019**, *67*, 7331–7376. [CrossRef]
19. Wang, J.; Jiang, C.; Zhang, H.; Ren, Y.; Chen, K.-C.; Hanzo, L. Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks. *IEEE Commun. Surv. Tutorials* **2020**, *22*, 1472–1514. [CrossRef]
20. 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]
21. 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]
22. Chen, X.; Wu, C.; Chen, T.; Zhang, H.; Liu, Z.; Zhang, Y.; Bennis, M. Age of Information Aware Radio Resource Management in Vehicular Networks: A Proactive Deep Reinforcement Learning Perspective. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 2268–2281. [CrossRef]
23. Fu, J.; Qin, X.; Huang, Y.; Tang, L.; Liu, Y. Deep Reinforcement Learning-Based Resource Allocation for Cellular Vehicular Network Mode 3 with Underlay Approach. *Sensors* **2022**, *22*, 1874. [CrossRef]
24. 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]
25. Foerster, J.; Nardelli, N.; Farquhar, G.; Afouras, T.; Torr, P.H.S.; Kohli, P.; Whiteson, S. Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning. In Proceedings of the International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1146–1155.
26. Wang, R.; Jiang, X.; Zhou, Y.; Li, Z.; Wu, D.; Tang, T.; Fedotov, A.; Badenko, V. Multi-agent reinforcement learning for edge information sharing in vehicular networks. *Digit. Commun. Netw.* **2022**, *8*, 267–277. [CrossRef]
27. Yuan, Y.; Zheng, G.; Wong, K.-K.; Ben Letaief, K. Meta-Reinforcement Learning Based Resource Allocation for Dynamic V2X Communications. *IEEE Trans. Veh. Technol.* **2021**, *70*, 8964–8977. [CrossRef]
28. Gündoğan, A.; Gürsu, H.M.; Pauli, V.; Kellerer, W. Distributed resource allocation with multi-agent deep reinforcement learning for 5G-V2V communication. In Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, Online, 11–14 October 2020; pp. 357–362.
29. 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. [CrossRef]
30. He, Y.; Wang, Y.; Yu, F.R.; Lin, Q.; Li, J.; Leung, V.C.M. Efficient Resource Allocation for Multi-Beam Satellite-Terrestrial Vehicular Networks: A Multi-Agent Actor-Critic Method With Attention Mechanism. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 2727–2738. [CrossRef]
31. Iqbal, S.; Sha, F. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2961–2970.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Long Beach, CA, USA, 4–9 December 2017.
33. Gonzalez-Martin, M.; Sepulcre, M.; Molina-Masegosa, R.; Gozalvez, J. Analytical Models of the Performance of C-V2X Mode 4 Vehicular Communications. *IEEE Trans. Veh. Technol.* **2018**, *68*, 1155–1166. [CrossRef]
34. Nguyen, H.-H.; Hwang, W.-J. Distributed Scheduling and Discrete Power Control for Energy Efficiency in Multi-Cell Networks. *IEEE Commun. Lett.* **2015**, *19*, 2198–2201. [CrossRef]
35. Otterlo, M.; Wiering, M. Reinforcement Learning and Markov Decision Processes. In *Reinforcement Learning*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 3–42.
36. Watkins, C.J.C.H. Learning from Delayed Rewards. Ph.D. Thesis, King's College, London, UK, 1989.
37. Tokic, M.; Palm, G. Value-difference based exploration: Adaptive control between epsilon-greedy and softmax. In *Annual Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 335–346.

38. 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]
39. 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: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2934 (accessed on 1 September 2022).
40. Kyösti, P.; Meinilä, J.; Hentila, L.; Zhao, X.; Jämsä, T.; Schneider, C.; Narandzic, M.; Milojevic, M.; Hong, A.; Ylitalo, J.; et al. IST-4-027756 WINNER II D1.1.2 v1.2 WINNER II channel models. *Inf. Soc. Technol.* **2008**, *11*, 1–82.
41. Zeng, Y.; Xu, X. Toward Environment-Aware 6G Communications via Channel Knowledge Map. *IEEE Wirel. Commun.* **2021**, *28*, 84–91. [CrossRef]