MARVEL: Multi-Agent Reinforcement Learning for VANET Delay Minimization

Chengyue Lu[1], Zihan Wang[1], Wenbo Ding[1,*], Gang Li[2], Sicong Liu[3], Ling Cheng[4]

[1] Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, 518055, China
[2] Department of Electronic Engineering, Tsinghua University, 100084, China
[3] Department of Information and Communication Engineering, Xiamen University, 361005, China
[4] School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, 2000, South Africa
[*] The corresponding author, email: ding.wenbo@sz.tsinghua.edu.cn

Abstract: In urban Vehicular Ad hoc Networks (VANETs), high mobility of vehicular environment and frequently changed network topology call for a low delay end-to-end routing algorithm. In this paper, we propose a Multi-Agent Reinforcement Learning (MARL) based decentralized routing scheme, where the inherent similarity between the routing problem in VANET and the MARL problem is exploited. The proposed routing scheme models the interaction between vehicles and the environment as a multi-agent problem in which each vehicle autonomously establishes the communication channel with a neighbor device regardless of the global information. Simulation performed in the 3GPP Manhattan mobility model demonstrates that our proposed decentralized routing algorithm achieves less than 45.8 ms average latency and high stability of 0.05 % averaging failure rate with varying vehicle capacities.

Keywords: VANET; multi-agent RL; delay minimization; routing algorithm

I. INTRODUCTION

Rapid urbanization is resulting in a growing vehicle population accompany with an increase in traffic accidents and deterioration of traffic congestion. Under this circumstance, the Intelligent Transportation System (ITS) and Vehicle-to-Everything (V2X) communications have been proposed aiming to relieve traffic pressure [1]. As an essential ITS component, the Vehicular Ad hoc Network (VANET) is committed to building a self-organizing, easy to deploy, low cost, and open-structure vehicle communication network [2, 3], which has demonstrated a great potential of tackling the above issues.

However, due to the nature of highly dynamic topology, strict delay requirements, high node mobility, etc., VANET usually suffers from data packet dropouts compared to the traditional Mobile Ad hoc Network (MANET) [2, 4, 5]. Consequently, the routing protocols for MANET, such as ad hoc on-demand distance vector (AODV) [6], Dynamic Source Routing (DSR) [7], optimized link-state protocol (OLSR) [8] and so on, will have significant performance degradation in VANET, especially under highly dynamic vehicle communication scenarios [9]. To this end, many efforts have been carried out in routing protocol design for VANETs to achieve better performance [10–12]. In general, the VANET routing protocols can be divided into two types, the topology-based and the location-based ones. The topology-based protocols are mainly imitated from MANET but customized for VANET scenarios, for example, Direction AODV (DAODV) by Abedi *et al.* [13], and the Receive On the Most Stable Group-Path (ROMSGP) scheme by Taleb *et al.* [14]. Nevertheless, in the context of vehicular

Received: Nov. 22, 2020
Revised: Mar. 15, 2021
Editor: Shan Zhang

communications, especially in traffic safety applications, the high cost of routing establishment usually causes intolerable delays. The location-based ones have emerged to be the popular protocols for VANET due to the additional location and velocity information from the Global Positioning System (GPS), of which classical algorithms such as the Greedy Perimeter Stateless Routing (GPSR) by Karp *et al.* [15], the Geographic Source Routing (GSR) by Lochert *et al.* [16] and the Anchor-based Street and Traffic-Aware touting (A-STAR) by Liu *et al.* [17] are proposed. However, the location-based protocols often experience connection failure in complex road conditions, especially in crossroads.

The traditional routing scheme has high convergence and short response time, whereas it is not practical to design a universal routing protocol. Different VANET routing protocols are required in various deployments. For the large-scale VANET, the network's dynamic changes are more frequent, and its protocol should be highly adaptable to topology changes and have good scalability. For VANETs with fast node movement speeds and strict delay requirements, their protocols must guarantee low delays. With the development of Machine Learning (ML) and Deep Learning (DL), Tang *et al.* [18] began their research on ML-based mobility prediction in delay-minimization routing. The ML-based algorithm can achieve relatively low average latency in different traffic flow environments. Then, the Reinforcement Learning (RL) based routing schemes were proposed by C. Wu *et al.* and F. Li *et al.* [19, 20], where the RL agent training is to obtain optimal responses after observing data from the environment. Because the environment modeling by RL is close to the real world, the RL-based routing algorithms achieve better performance compared with the traditional routing protocols. However, in resource management, single-agent RL is hard to provide a distributed perspective on identifying the resource requirement of each agent [21]. Therefore, combine the RL and multi-agent problem, the distributed perspective of Multi-Agent Reinforcement Learning (MARL) makes it more adaptive to the potential decentralized applications such as VANET.

To this end, this paper focuses on designing a unicast routing algorithm with minimum delay under the framework of MARL, where the main idea is modeling the decision of router selection for each vehicle as a Markov Decision Process (MDP). Vehicles learn the selection strategy based on the MDP model with an online distributed learning algorithm. The improved versions of Deep Q-Network (DQN) are adopted to solve the multi-agent problem. Specifically, the model design is based on Independent Q-Learning (IQL) [22], and the converging stability of the proposed algorithm is improved by Deep Double Q-learning Network (DDQN) [23] and the dueling network [24]. By modeling the routing problem into the multi-agent problem, all the vehicles and infrastructures are considered as agents that can automatically find an optimal path to minimize communication latency.

The rest of this paper is organized as follows. Section II defines the network model and provides problem formulation. In Section III, we present the modeling and implementation of MARL based routing algorithm. Then, the simulation results are shown in Section IV. Finally, concluding remarks are given in Section V.

II. SYSTEM MODEL

2.1 Network Model

The network model of a communication system including V2I and V2V links in an urban scenario is considered, which simultaneously provides the high-speed communication and the periodic safety messages, as introduced in 3GPP Release 15 for cellular V2X enhancement [25]. In this model, M vehicles and N BS/RSUs are planted in the simulation environment. All the BS and RSUs are connected via a wired link. Each vehicle periodically sends packets that not only containing application messages but also include location and velocity information to its neighbors. Moreover, we assume that all the vehicles use transceivers with an identical single antenna and all the packet to be transmitted has the fixed size. When a vehicle transmits packets, it chooses a channel different from the channel occupied by the periodic safety messages to avoid interference. Since the vehicle can select any nearby vehicle or RSU as a router, if one channel is blocked due to excessive interference or distant range, the communication will be re-established via another device.

The uplink transmission rate R_m^y of vehicle m to

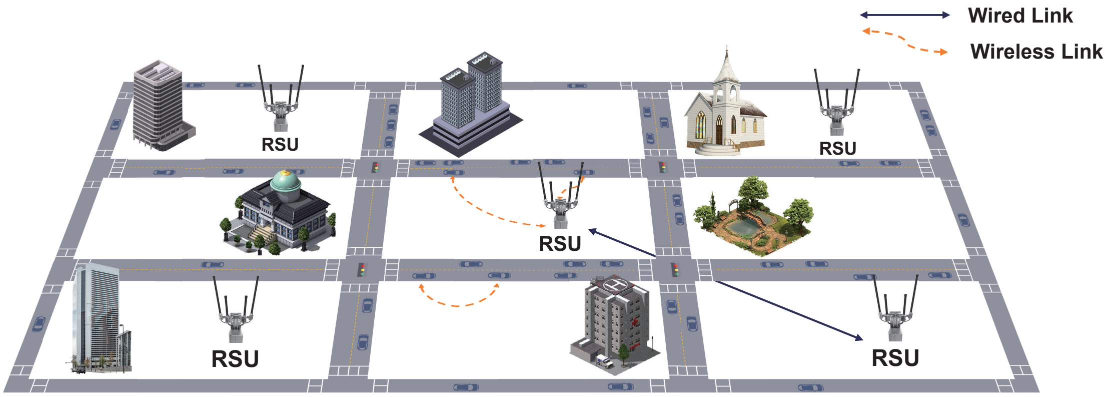


Figure 1. *System overview of the urban VANET.*

RSU y and the downlink transmission rate R_y^n of RSU y to vehicle n for V2I communication are defined by

$$R_{\mathrm{V2I}} = R_m^y = W_m \cdot \log_2\left(1 + \frac{P_V}{\sigma^2 + I_c}\alpha_{m,y}h_{m,y}\right), \tag{1}$$

$$R_{\mathrm{I2V}} = R_y^n = W_n \cdot \log_2\left(1 + \frac{P_I}{\sigma^2 + I_c}\alpha_{y,n}h_{y,n}\right), \tag{2}$$

respectively, where W_m and W_n are the bandwidth allocated for uplink and downlink. P_V denotes the transmission power of vehicles; P_I denotes the transmission power of RSUs. $h_{m,y}$ and $h_{y,n}$ represents the small-scale fading power component, which are related to the frequency and assumed to be an exponential distribution of the unit mean. $\alpha_{m,y}$ and $\alpha_{y,n}$ are the large-scale fading effect which consists of path loss and shadowing. Additionally, σ^2 represents the power of Additive White Gaussian Noise (AWGN) and I_c is the interference in the selected channel.

The transmission rate R_m^n between vehicle m and n is described by

$$R_{\mathrm{V2V}} = R_m^n = W_V \cdot \log_2\left(1 + \frac{P_V}{\sigma^2 + I_c}\alpha_{m,n}h_{m,n}\right), \tag{3}$$

where W_V is the V2V communication bandwidth, and $h_{m,n}$ and $\alpha_{m,n}$ are the small-scale fading and the large-scale fading effects respectively.

2.2 Problem Formulation

In this paper, the urban case evaluation method is shown in Figure 1 We designed a request delivery task, where each vehicle will send a packet with a fixed size to the target vehicle, and the router for this transmission can be selected from other vehicles, Base Station and Road Side Units (BS/RSU). Each vehicle and RSU are indexed to $V_m \in \mathcal{V}$ and $I_n \in \mathcal{I}$, respectively. We denote the number of packets by X and their sizes by Z.

The t^{th}-hop delay $T_{t,x,y}$ can be calculated by

$$T_{t,x,y} = \begin{cases} \frac{Z}{R_{\mathrm{V2I}_{t,x,y}}}, (x \in \mathcal{V}, y \in \mathcal{I}) \\ \frac{Z}{R_{\mathrm{I2V}_{t,x,y}}}, (x \in \mathcal{I}, y \in \mathcal{V}) \\ \frac{Z}{R_{\mathrm{V2V}_{t,x,y}}}, (x \in \mathcal{V}, y \in \mathcal{V}) \\ \frac{Z}{R_{\mathrm{I2I}}}, (x \in \mathcal{I}, y \in \mathcal{I}) \end{cases}, \tag{4}$$

where R_{I2I} is the wired transmission rate between RSUs, and other notations are in consistent with Section 2.1.

The goal of the routing algorithm is to minimize the overall vehicle service latency that the process can be expressed as the following optimization problem,

$$\begin{array}{rl} (\mathcal{P}_1): \underset{\alpha_{t,x,y}}{\text{minimize}} & \sum_{t=0}^{t_{\max}} \sum_{y:y\neq x} \alpha_{t,x,y}T_{t,x,y} \\ \text{subject to} & \sum_{y:y\neq x} \alpha_{t-1,x,y} = \sum_{x:x\neq y} \alpha_{t,y,x} \\ & \sum_{y:y\neq S} \alpha_{0,S,y} = 1 \\ & \sum_{t=0}^{t_{\max}} \sum_{x:x\neq D} \alpha_{t_{\max},x,D} = 1 \\ & \alpha_{t,x,y} \in \{0,1\} \\ & x, y \in \{\mathcal{V}, \mathcal{I}\} \\ & t \in \{0, 1, \cdot\cdot\cdot, t_{\max}\} \end{array}, \tag{5}$$

where $\alpha_{t,x,y}$ is a binary variable. If $\alpha_{t,x,y} = 1$, it means that the t^{th}-hop vehicle x will send a packet to vehicle y. The first constraint denotes all the packets received by vehicle x in $(t-1)^{th}$-hop will be sent out at the t^{th} hop. The last two constraints indicate that, at the starting time, source vehicle S must choose a device as the router, and one packet is sent to the destination vehicle D only once throughout the simulation.

Typically, the problem $\mathcal{P}_1$ is NP-hard as it is an integer programming problem where vehicles' location and the fading effect are updated in each step. To solve this problem, we propose an MARL based routing scheme.

III. MARL BASED ROUTING SCHEME

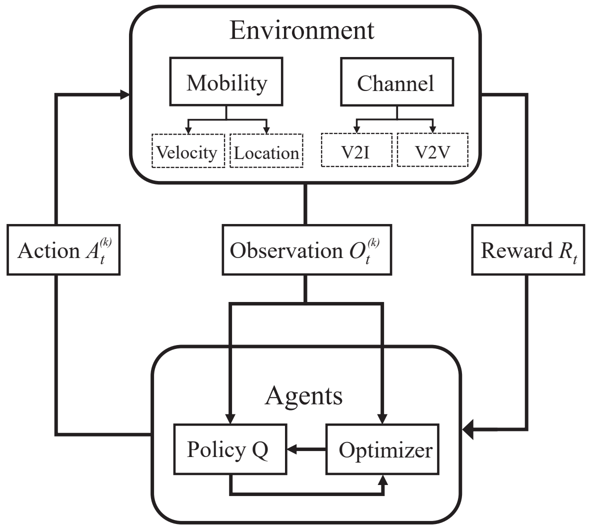


Figure 2. *Structure of agents' interaction in the routing environment.*

The proposed MARL scheme is to address the V2X routing problem illustrated in Figure 1. First, the observation space was designed to accommodate the environment. Then, followed by the standard action space configuration, our carefully designed reward policy is proposed to respond promptly. Here, the interaction between agents and the routing environment is depicted in Figure 2.

After modeling the routing problem into the RL interface, we combine the two state-of-the-art methods, Dueling DQN and DDQN [23, 24], to solve the overestimation problem and speed up the convergence in the nature DQN [26]. Since it is inapplicable to deploy centralized RL in the VENET, we introduce the multi-agent strategy to fit the real scenario better. At the end of this section, the implementation of the MARL-based routing algorithm is elaborated.

3.1 Observation Space

We define the states based on three factors, overall latency T, safety messages, and the local channel information. The safety messages are obtained from neighbor vehicles, which contain position P_k, speed V_k, and direction D_k. And the local channel information includes channel interferences from other V2V transmitters $\alpha_{k,k'}$, and from the RSUs $\alpha_{k,y}$ (for all $k \neq k'$, $k \in \mathcal{V}$ and $y \in \mathcal{I}$). As a result, the observation space for the vehicle k is integrated as

$$O_t^{(k)} = \left\{ \begin{array}{l} T, \{P_x, P_y, V_x, D_x\}_{x\in\mathcal{V}, y\in\mathcal{I}}, \\ \{\alpha_{k,x}, \alpha_{k,y}\}_{x\in\mathcal{V}, x\neq k, y\in\mathcal{I}} \end{array} \right\}. \tag{6}$$

3.2 Action Space

A vehicle that carries a packet needs to select another vehicle or RSU as the router. Therefore, the action function can be designed as

$$A_t^{(k)} = \{x \text{ or } y\}_{x\in\mathcal{V}, x\neq k, y\in\mathcal{I}} \,. \tag{7}$$

3.3 Reward Design

The design of the reward function will directly affect the performance of the model. If we simply design the reward function as a time-dependent reward, when the destination vehicle receives the packet. The delayed reward will cause two problems. One is the huge time consumption in exploratory work, and the other is the difficulty in addressing the credit assignment problem.

To achieve an accurate and real-time reward function, we design the reward function with three phases:

$$R_t = \begin{cases} \alpha - kT, & \text{if } y = D, \\ \beta(dis(P_D, P_x) & \text{if } y \neq D \text{ and} \\ -dis(P_D, P_y)), & dis(P_x, P_y) \leq R, \\ -\lambda, & \text{if } dis(P_x, P_y) > R, \end{cases}, \tag{8}$$

where α, β, λ and k are parameters set manually, R represents the transmission range and $dis()$ is the operator of calculating the Euclidean distance. The first phase means the agent will get a time-dependent reward when the packet transmits to the destination vehicle. The second phase indicates that if the t^{th}-hop is successful, the reward will depend on how much the distance change before and after transmission and the third one shows the penalty of transmission failure.

3.4 Learning Algorithm

In the proposed learning algorithm, we firstly use DQN to feed the RL model. Then, two improved methods, DDQN and Dueling DQN are adopted to boost up convergence. Finally, the proposed MARL based routing algorithm by incorporating the RL and multi-agent model is described.

DQN is a combination of Q-Learning and Deep Neural Network (DNN), which can learn control strategies directly from high-dimensional raw data. In classic Q-Learning, the Q values of each state-action pair are store in Q-table. However, the high dimensional state and action space in the real application make the Q-table update problem can only be approximately addressed by converting into a function fitting problem. In the following formula, the Q function approximates the optimal Q value by updating the parameter

$$Q(s, a; \theta) \approx Q^*(s, a), \tag{9}$$

where s and α denote the state and action respectively, and θ is the parameter matrix that updates in each learning step. Moreover, Q^* is the best Q function. The input of DNN is the original data (state), and the output is value evaluation (Q value) corresponding to each action. The function of DNN in DQN is to fit the Q-table in the high-dimensional continuous space. And optimized by supervised learning. DQN determines the loss function based on the Q-Learning updating formula:

$$\begin{aligned} &Q^*(s, a) = Q(s, a) \\ &+\alpha\left(r + \gamma \max_{a'} Q\left(s', a'\right) - Q(s, a)\right) \end{aligned}, \tag{10}$$

where λ is the discount rate, s' and α' is the next state and the next action respectively. And the DQN's loss function is:

$$L(\theta) = E\left[\left(Q_{\text{target}}(s, a) - Q(s, a; \theta)\right)^2\right], \tag{11}$$

$$Q_{\text{target}}(s, a) = r + \gamma \max_{a'} Q\left(s', a'; \theta'\right), \tag{12}$$

where θ' is the periodical updating target Q network weights. The introduction of the target Q net not only reduce the correlation between the primary Q value and the target Q value simultaneously but also improved the algorithm stability. Another improvement of DQN from Q-Learning is the experience replay, mainly used to solve correlation and non-static distribution problems. Specifically, it stores the transfer samples (s, a, r, s') obtained from the interaction between each time step agent and the environment into the playback memory unit and then randomly fetch samples for training.

DDQN is proposed by Hasselt *et al.* [23], who proved that classic DQN tends to overestimate the Q value of action, and the estimation error will be magnified with the increasing of action space. If the overestimation is non-uniform, it will cause the Q value of a suboptimal action to exceed the optimal Q value, and the optimal policy will never be found. The DDQN modify the generation mode of the target Q value to be learned as the following equation:

$$\begin{aligned} &Q_{\text{target}}(s, a) = r \\ &+\gamma Q\left(s', \arg\max_{a'}\left(Q\left(s', a'; \theta\right)\right), \theta^{-}\right) \end{aligned}. \tag{13}$$

The next state's optimal action is found within the primary Q net, and then resorting target Q Net to find the Q value of the action to form the target Q value. Therefore, the maximum value in the target Q net is unnecessary, which avoids selecting the overestimated suboptimal action.

Dueling DQN is a competitive network model of DQN proposed by Wang *et al.* [24]. The network divides the abstract features extracted by DNN into two branches, one branch represents the state value function $V(s; \theta)$, and another represents the action advantage function $A(s, a; \theta)$ of the dependent state and the added value of selecting an action. Eventually, these two branches are combined to yield the Q value for each action as the following equation:

$$\begin{aligned} &Q(s, a; \theta) = V(s; \theta) \\ &+\left(A(s, a; \theta) - \tfrac{1}{|A|}\sum_{a'} A\left(s, a'; \theta\right)\right) \end{aligned}. \tag{14}$$

In this way, each action's dominant function in this state can be independent of the order. The range of Q value can be reduced, redundant degrees of freedom can be removed, and the algorithm's stability gets improved.

The multi-agent model design is based on the IQL [22], which addressed the difficulty of deploying centralized RL algorithms in VANET routing. To be specific, all vehicles share the same Q network parameters in centralized RL model, so that the input dimension is too large to converge the model. Also, the input design makes it impossible for heterogeneous updating. Therefore, in the real application, the centralized large scale algorithm is not ready to deploy [27]. However, in the multi-agent setting, our model assumes only one packet will be transferred from one source vehicle to one destination vehicle at each iteration in the learning stage. The vehicle carrying the packet will choose a neighborhood within the transmission range as a router until the destination vehicle completes receiving the packet. Otherwise, links will break when the distance between the sender and receiver exceeds the transmission distance. In our algorithm, each training step sender k observes the current state and fed it to the primary Q net to choose the best action which achieves the highest Q value with the ϵ-greedy strategy. The ϵ-greedy strategy is a soft policy, meaning that the best action will be chosen with probability $1-\epsilon$ and probability ϵ for a random action. After interacting with the environment, the current state, action, reward, and the next state are stored in the replay memory as a transition tuple, i.e., $\left(O_t^{(k)}, A_t^{(k)}, R_t, O_{t+1}^{(k)}\right)$. Then, a mini-batch of experiences will be sampled randomly to update the primary Q net parameters with the Adaptive Moment Estimation (Adam) [28] gradient-descent method. Many optimizer algorithms are evaluated and compared in [29], which concludes that Adam is the best choice overall. At the end of each episode, the target Q net parameters will be updated as a primary Q net copy. In summary, the whole process of our training procedure is shown in Algorithm 1.

Algorithm 1. *The proposed MARL-based routing algorithm.*

```
Start environment simulator and generating vehicles
Initialize network parameters randomly for all agents
for each episode do
    Update large-scale fading and vehicles information
    for each step do
        Observe $O_t^{(k)}$
        Choose an action $A_t^{(k)}$ according to the $\epsilon$-greedy strategy
        Act $A_t^{(k)}$ and observe the reward $R_t$
        Update channel small-scale fading
        Observe $O_{t+1}^{(k)}$
        Store $\left(O_t^{(k)}, A_t^{(k)}, R_t, O_{t+1}^{(k)}\right)$ into reply memory $M$
        Compute remain packet size $Z$
        if $Z \leq 0$ then
            Change sender k
        end if
        Select mini-batches samples from $M$
        Calculate the loss L, defined in Equation (11–14)
        Update primary $Q$ net parameters with $\nabla L$ using Adam optimizer
    end for
    Update target $Q$ net parameters
end for
```

3.5 Algorithm Complexity

According to algorithms 6.3 and 6.4 in [30], the time complexity order of DNN is O(n^2) both in forward and backward propagation. Since the input space of the algorithm is K for MARL and K^2 for the centralized RL method. The time complexity order of the MARL method is O(K^2) and centralized RL is O(K^4) in one hop. Therefore, the MARL method is less time-consuming than the centralized RL methods and this method is suitable for delay-sensitive applications.

IV. SIMULATION RESULTS

To evaluate the MARL-based VANET routing algorithm's performance, we built a simulation environment based on the Manhattan case defined in Annex A of 3GPP TR 36.885 [31]. The annex regulating vehicle User Equipment (UE) drop, RSU deployment, and mobility model, which contains the specification of road grid, simulation area size, vehicle density, and the magnitude of vehicle speed. Simulation parameters setting, which refers to Liang *et al.*'s [32] works, are summarized in Table 1, and channel models for two types of wireless links are listed in Table 2.

Table 1. *Simulation parameters [31].*

Parameter	Value
Number of RSUs N	5
Bandwidth	4 MHz
BS antenna height	25 m
BS antenna gain	8 dBi
BS receiver noise figure	5 dB
Vehicle antenna height	1.5 m
Vehicle antenna gain	3 dBi
Vehicle receiver noise figure	9 dB
Absolute vehicle speed v	[36-54] km/h
Vehicle drop and mobility model	Urban case of A.1.2 in [31]*
Vehicle sender transmission power P^C	10 dBm
BS sender transmission power P^D	23 dBm
Noise power σ^2	-114 dBm
Packet size	1 Mbit

* We shrink the height and width of the simulation area by a factor of 2.

Our network in the routing scheme has three fully connected hidden layers, and the number of neurons in each layer is 2048, 512, and 128, respectively. We utilize the rectified linear unit (ReLU), $f(x) = \max(0, x)$, as the activation function. The ϵ-greedy policy is used to avoid overfitting during evaluation [26], and Adam [28] optimizer is used for training with a learning rate of 0.001. The small-scale fading effect is changed in 1 ms, and the large-scale fading effect and the position of vehicles are updated every 100 ms.

Moreover, we compare the proposed RL-based routing algorithms with three classic centralized routing protocols AODV [6], DAODV [13] and GPSR [15]. The evaluation of these five methods is performed by 10000 simulations in each different vehicle number setting. Figure 3 shows the average delays of centralized RL algorithm, AODV protocol, DAODV protocol, GPSR protocol, and our proposed MARL-based routing scheme in different vehicle densities. It is worth noting that the centralized RL algorithm has an average delay of over 100 ms when the number of vehicles reaches 50. Compared to the rest algorithms, the proposed MARL-based scheme achieves both lower latency and higher stability. It can be observed that the MARL-based scheme keeps a steady 45.8 ms average delay regardless of the traffic volume variations. In contrast, the average delay of AODV, DAODV, and GPSR protocol rises with the increase in vehicles. From the simulation results, the MARL-based routing scheme outperforms 6 %-14 % of AODV (3.9 ms-7.5 ms), 4 %-10 % of the DAODV (2.8 ms-5.2 ms), and 1 %-12 % of the GPSR (0.5 ms-6.0 ms).

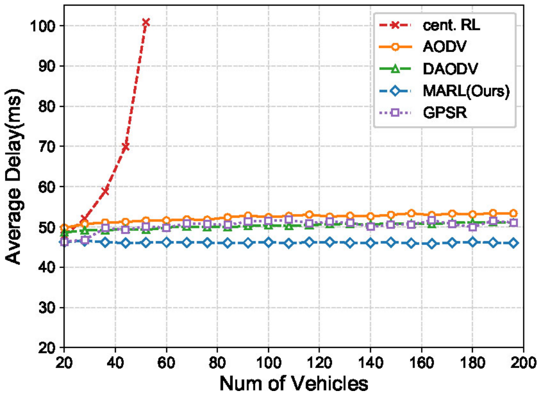


Figure 3. *Average delay of the centralized RL algorithm, DAODV protocol, AODV protocol, GPSR protocol, and our proposed MARL scheme with varying number of vehicles.*

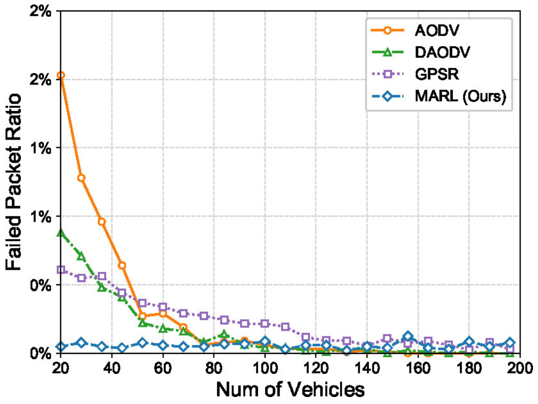


Figure 4. *Failed packet ratio with the various number of vehicles.*

Figure 4 illustrates the ratio of dropout packets in different vehicle capacities under three routing schemes. The centralized RL algorithm suffers an intolerable dropout rate hence not presented in the figure. The proposed MARL routing scheme achieves

Table 2. *Simulation parameters [31, 33].*

Parameter	V2I and I2V link	V2V link
Path loss model	128.1 + 37.6log10 d, d in km	LOS in WINNER + B1 Manhattan [33]
Shadowing distribution	Log-normal	Log-normal
Shadowing standard deviation ξ	8 dB	3 dB
Decorrelation distance	50 m	10 m
Path loss and shadowing update	A.1.4 in [31] every 100 ms	A.1.4 in [31] every 100 ms
Fast fading	Rayleigh fading	Rayleigh fading
Fast fading update	Every 1 ms	Every 1 ms

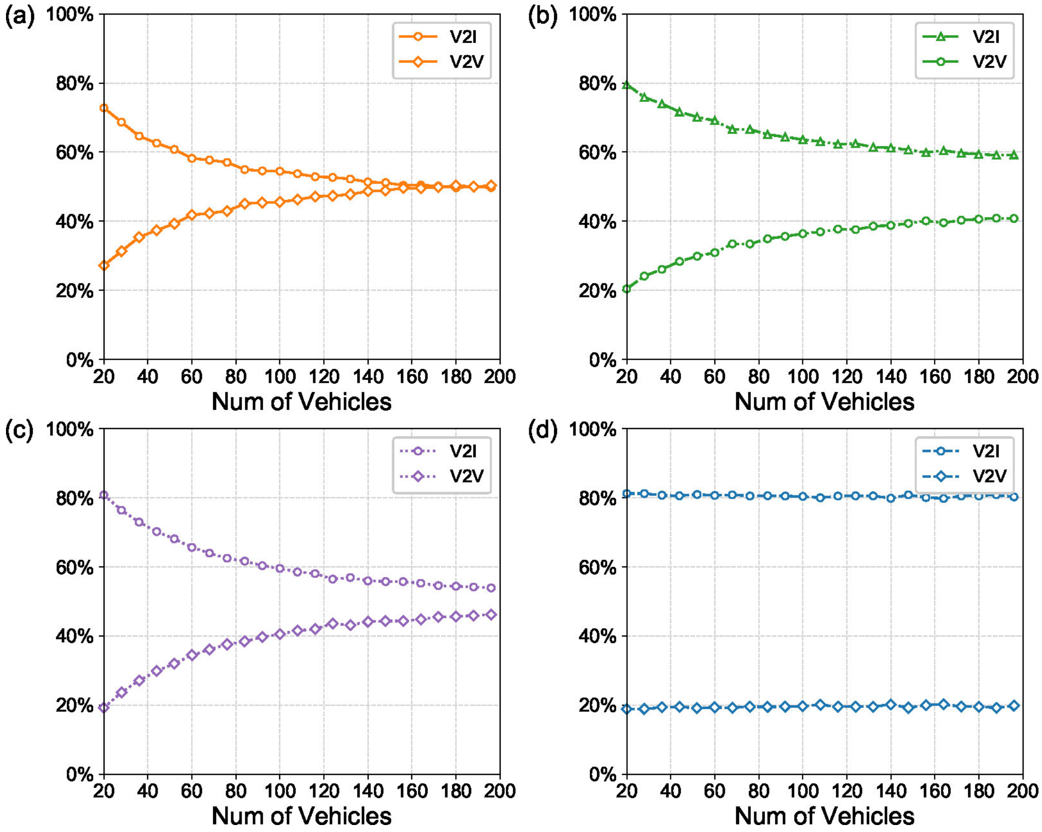


Figure 5. *Communication type distribution of three algorithms (a. AODV, b. DAODV, c. GPSR, and, d. MARL-based algorithm) with a different number of vehicles setting.*

an overall failed packet ratio of 0.05 %. Our method performs an order of magnitude less dropout rate than the other two algorithms in low vehicle density (less than 80 vehicles in the simulation environment). We believed that in a sparse vehicle distribution, the link between vehicles has a higher probability of breaking; however, the proposed MARL-base method learns mobility patterns which enabled better target estimation.

To further explain the reason that the proposed MARL-based routing scheme outperforms the AODV, DAODV, and GPSR protocols in delay and stability, we analyze the portion of V2I and V2V links in three methods under different vehicles number settings, as depicted in Figure 5. The ratio of high-performance V2I channel usage in the MARL-based algorithm is higher than the others, and keeps relatively consistent with the increase of participate vehicles. This indicates that the MARL-based algorithm tends to maximize the usage of high-quality communication resources, therefore exhibiting low delay and high stability despite the increase of vehicles.

V. CONCLUTION

This paper has proposed a decentralized routing algorithm for VANET to minimize the network delay under the MARL framework. An innovative input state and reward function were carefully designed for DQN to reduce the overall latency and the dropout rate, especially in low vehicle density. The proposed method was compared with AODV and DAODV protocols in the 3GPP Manhattan mobility model. The simulation results show that the routing delay of the proposed algorithm is robust to the vehicle amount change, and it has a distinctive low failure rate in sparse vehicle distribution scene compared with others. To summarize, this work presented an efficient and intelligent routing approach for vehicle communication, which expect to empower a variety of low latency services to the vehicle network, especially to notify the traffic emergency.

ACKNOWLEDGEMENT

The authors wish to thank Dr. Le Liang, who gave valuable suggestions that helped design the RL model and simulation environment. The authors also would like to thank the support from Mr. Jin Sima, for the guidance of mathematical derivation. This work is supported by the National Science Foundation of China under grant No. 61901403, 61790551, and 61925106, Youth Innovation Fund of Xiamen No. 3502Z20206039 and Tsinghua-Foshan Innovation Special Fund (TFISF) No. 2020THFS0109.

References

[1] O. Andrisano, R. Verdone, *et al.*, "Intelligent transportation systems: the role of third generation mobile radio networks," *IEEE Communications Magazine*, vol. 38, no. 9, 2000, pp. 144–151.

[2] H. Hartenstein and L. Laberteaux, "A tutorial survey on vehicular ad hoc networks," *IEEE Communications magazine*, vol. 46, no. 6, 2008, pp. 164–171.

[3] O. Kaiwartya, A. H. Abdullah, *et al.*, "Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects," *IEEE Access*, vol. 4, 2016, pp. 5356–5373.

[4] L. Liang, H. Peng, *et al.*, "Vehicular communications: A physical layer perspective," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, 2017, pp. 10 647–10 659.

[5] H. Peng, L. Liang, *et al.*, "Vehicular communications: A network layer perspective," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, 2018, pp. 1064–1078.

[6] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proceedings WMCSA'99. Second IEEE Workshop on Mobile Computing Systems and Applications*. IEEE, 1999, pp. 90–100.

[7] D. B. Johnson, D. A. Maltz, *et al.*, "Dsr: The dynamic source routing protocol for multi-hop wireless ad hoc networks," *Ad hoc networking*, vol. 5, no. 1, 2001, pp. 139–172.

[8] T. Clausen, P. Jacquet, *et al.*, "Optimized link state routing protocol (olsr)," 2003.

[9] M. H. Eiza, Q. Ni, *et al.*, "Investigation of routing reliability of vehicular ad hoc networks," *EURASIP journal on wireless communications and networking*, vol. 2013, no. 1, 2013, pp. 1–15.

[10] D. Tian, K. Zheng, *et al.*, "A microbial inspired routing protocol for vanets," *IEEE Internet of Things Journal*, vol. 5, no. 4, 2017, pp. 2293–2303.

[11] N. Alsharif and X. Shen, "i car-ii: Infrastructure-based connectivity aware routing in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, 2016, pp. 4231–4244.

[12] J. He, L. Cai, *et al.*, "Delay analysis and routing for two-dimensional vanets using carry-and-forward mechanism," *IEEE Transactions on Mobile Computing*, vol. 16, no. 7, 2016, pp. 1830–1841.

[13] O. Abedi, M. Fathy, *et al.*, "Enhancing aodv routing protocol using mobility parameters in vanet," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*. IEEE, 2008, pp. 229–235.

[14] T. Taleb, E. Sakhaee, *et al.*, "A stable routing protocol to support its services in vanet networks," *IEEE Transactions on Vehicular technology*, vol. 56, no. 6, 2007, pp. 3337–3347.

[15] B. Karp and H.-T. Kung, "Gpsr: Greedy perimeter stateless routing for wireless networks," in *Proceedings of the 6th annual international conference on Mobile computing and networking*, 2000, pp. 243–254.

[16] C. Lochert, H. Hartenstein, *et al.*, "A routing strategy for vehicular ad hoc networks in city environments," in *IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No. 03TH8683)*. IEEE, 2003, pp. 156–161.

[17] G. Liu, B.-S. Lee, *et al.*, "A routing strategy for metropolis vehicular communications," in *International conference on information networking*. Springer, 2004, pp. 134–143.

[18] Y. Tang, N. Cheng, *et al.*, "Delay-minimization routing for heterogeneous vanets with machine learning based mobility prediction," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, 2019, pp. 3967–3979.

[19] C. Wu, K. Kumekawa, *et al.*, "Distributed reinforcement learning approach for vehicular ad hoc networks," *IEICE*

transactions on communications, vol. 93, no. 6, 2010, pp. 1431–1442.

[20] F. Li, X. Song, *et al.*, "Hierarchical routing for vehicular ad hoc networks via reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, 2018, pp. 1852–1865.

[21] L. Buşoniu, R. Babuška, *et al.*, "Multi-agent reinforcement learning: An overview," *Innovations in Multi-agent Systems and Applications-1*, 2010, pp. 183–221.

[22] J. Foerster, N. Nardelli, *et al.*, "Stabilising experience replay for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1146–1155.

[23] H. Van Hasselt, A. Guez, *et al.*, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[24] Z. Wang, T. Schaul, *et al.*, "Dueling network architectures for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1995–2003.

[25] "Technical specification group radio access network; Study enhancement 3GPP Support for 5G V2X Services; (Release 15)," Document 3GPP TR 22.886 V15.1.0, 3rd Generation Partnership Project, Mar. 2017.

[26] V. Mnih, K. Kavukcuoglu, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, 2015, pp. 529–533.

[27] B. L. Golden, S. Raghavan, *et al.*, *The vehicle routing problem: latest advances and new challenges*. Springer Science & Business Media, 2008, vol. 43.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[30] I. Goodfellow, Y. Bengio, *et al.*, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

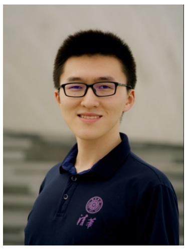
[31] "Technical specification group radio access network; Study LTE-Based V2X Services; (Release 14)," Document 3GPP TR 36.885 V14.0.0, 3rd Generation Partnership Project, Jun. 2016.

[32] L. Liang, H. Ye, *et al.*, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, 2019, pp. 2282–2292.

[33] "WF on SLS evaluation assumptions for EV2X, document R1-165704," 3GPP TSG RAN WG1 Meeting #85, May 2016.

Biographies

Chengyue Lu received the BS degrees from School of Electrical and Information Engineering, Hunan University of Technology in 2020. He is now a visiting student in Data Science and Information Technology at Smart Sensing and Robotics (SSR) group, Tsinghua University. His research interests include machine learning, Vehicle to Everything (V2X), Internet of Things (IoTs), and robotics.



Zihan Wang received the dual BEng. degrees (1st class Hons.) from School of Telecommunications Engineering, Xidian University and Edinburgh Centre for Robotics, Heriot-Watt University, respectively in 2019. He is currently pursuing his MS degree in Data Science and Information Technology at Smart Sensing and Robotics (SSR) group, Tsinghua University. His research interests include self-powered sensors, Internet of Things (IoTs), and robotics.

Wenbo Ding received the BS and PhD degrees (Hons.) from Tsinghua University in 2011 and 2016, respectively. He worked as a postdoctoral research fellow at Georgia Tech under the supervision of Professor Z. L. Wang from 2016 to 2019. He is now a tenure-track assistant professor and PhD supervisor at Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, where he leads the Smart Sensing and Robotics (SSR) group. His research interests are diverse and interdisciplinary, which include self-powered sensors, energy harvesting, and wearable devices for health and soft robotics with the help of signal processing, machine learning, and mobile computing. He has received many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva and the IEEE Scott Helt Memorial Award.

Gang Li received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2002 and 2007, respectively. Since July 2007, he has been with the Faculty of Tsinghua University, where he is a Professor with the Department of Electronic Engineering. From 2012 to 2014, he visited The Ohio State University, Columbus, OH, USA, and Syracuse University, Syracuse, NY, USA. He has authored or coauthored more than 160 journal and conference papers. He is the Author of the book, Advanced Sparsity-Driven Models and Methods for Radar Applications. His research interests include radar signal processing, distributed signal processing, sparse signal processing, remote sensing, and information fusion.

Sicong Liu received his B.S.E. and PhD degree (with highest honor) both in electronic engineering from Tsinghua University, Beijing, China in 2012 and 2017. He was a visiting scholar in City University of Hong Kong, China. He served as a senior engineer in Huawei Technologies.

Currently, he is a specially-appointed associate researcher and an assistant professor in Xiamen University, China. His research interests include sparse signal processing, machine learning, and wireless communications.

Ling Cheng received the degree B. Eng. Electronics and Information (cum laude) from Huazhong University of Science and Technology (HUST) in 1995, M. Ing. Electrical and Electronics (cum laude) in 2005, and D. Ing. Electrical and Electronics in 2011 from University of Johannesburg (UJ). His research interests are in Telecommunications and Artificial Intelligence. In 2010, he joined University of the Witwatersrand where he was promoted to Full Professor in 2019. He serves as the associate editor of three journals. He has published more than one hundred research papers in journals and conference proceedings. He has been a visiting professor at five universities and the principal advisor for over forty full research post-graduate students. He was awarded the Chancellor's medals in 2005, 2019 and the National Research Foundation ratings in 2014, 2020. The IEEE ISPLC 2015 best student paper award was made to his Ph.D. student in Austin. He is a senior member of IEEE and the vice-chair of IEEE South African Information Theory Chapter.