This article has been accepted for publication in a future proceedings of this symposium, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/pimrc54779.2022.9978038.10270247, 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)

Intelligent Energy Efficient Resource Allocation for URLLC Services in IoV Networks

Rana Muhammad Sohaib[*], Oluwakayode Onireti[*], Yusuf Sambo[*], Rafiq Swash[#], and Muhammad Imran[*],
[*]James Watt School of Engineering, University of Glasgow, G12 8QQ, Glasgow, UK
[#]AIDRIVERS LTD, 1390 Uxbridge Road, UB10 0NE London, UK
Email: [*]{ranamuhammad.sohaib, Oluwakayode.Onireti, Yusuf.Sambo, Muhammad.Imran}@glasgow.ac.uk and
[#]swash@aidrivers.ai

***Abstract*—Internet of vehicles (IoV) has been developed as a promising technology to improve road safety. However, resource management can be challenging in a congested traffic environment, which can affect the energy efficiency (EE) and spectrum efficiency (SE) in IoV networks. In this paper, we present a novel intelligent resource allocation approach based on deep reinforcement learning to maximize the weighted composite efficiency that incorporates the EE and SE metric subject to latency and reliability constraints of vehicle-to-vehicle (V2V) users. We employ Thompson sampling with double deep Q network to transform the objective function. Moreover, we present a probability-based learning approach to meet the quality of service requirements and to increase the learning ability of the proposed model. The simulation results indicate that the proposed approach maximizes the composite efficiency while satisfying the latency and reliability constraints of V2V users.**

***Index Terms*—5G, V2X, DRL, EE, SE, IoV.**

I. Introduction

In recent years, the internet of vehicles (IoV) networks has emerged as an important element in intelligent transport systems (ITS) by supporting wireless transmission services among vehicles [1], [2]. IoV network enables intelligent vehicle-to-vehicle (V2V) communication by providing internet services between vehicles [3]. Future cellular technology needs to coordinate the vehicle-to-everything (V2X) communications in an organized manner to enhance road safety, which involves communications among V2V users and vehicle-to-infrastructure (V2I) users [4]. Whereas, V2V users must meet the ultra-reliable and low latency communications (URLLC) requirements [5]. V2X communication can help to improve traffic efficiency and road safety by gathering real-time traffic information. The information services in the IoV environment demand endless continuous access to the network, where V2I pairs need to meet the capacity requirements by supporting high transmission rate. Moreover V2V pairs must meet the URLLC requirements in order to communicate safety-critical messages [6], [7]. Since a dedicated spectrum is assigned in V2X communications, and there will be interference among V2I and V2V pairs due to sharing of single resource block among vehicular users [8]. Many resource management schemes have been presented recently in order to meet the spectral efficiency (SE) and energy efficiency (EE) requirements of V2V and V2I pairs in IoV networks. In [9], the authors look into the problem of maximizing V2V pairs reuse. They utilize the upper bound of the outage probability to improve the SE.

In [10], the authors presents the centralized resource allocation approach where they transform the URLLC requirements of V2V users into optimization limitations, and present a heuristic approach to address the optimization problem. Large transmission overhead is a problem in a centralized approach and it increases significantly as the network size increases [11]. In [11], the authors proposed a decentralized scheme to allocate the frequency bands efficiently to V2V users by classifying them based on their data rate and positions. The conventional optimization schemes based on iterative algorithms are not effective to manage the resources due to the increase of network size in IoV networks. In particular, they cannot make rapid decisions in fast varying channel state conditions of vehicular networks and thus cannot address the URLLC requirements of V2V users. Machine learning (ML) based approach can be effective in addressing these challenges [12]. Deep reinforcement learning (DRL) has shown potential to address more difficult decision-making problems as required in IoV networks [13].

Authors in [14] present a DRL-based decentralized approach to maximize the SE in unicast and broadcast environments while considering the URLLC requirements of V2V pairs. In [15], the authors propose a deep Q-network (DQN) based approach to manage the resources. To optimize the communication mode selection in V2X communications, authors in [16] proposed a DRL-based semi-decentralized approach to increase the reliability of V2V users. In [17], the authors present a reinforcement learning (RL) based algorithm to manage the network load by classifying the vehicles into clusters. Authors in [18] describe case studies based on reinforcement learning to improve resource management in the V2X environment. In [19], the authors proposed a DQN-based centralized spectrum allocation framework, where the agent aims to maximize the SE of V2V and V2I users subject to the interference of priority users. Authors in [20] present a DRL-based real-time energy-aware offloading resource management framework to reduce the latency and energy consumption of vehicular users with mobile edge computing (MEC). Work in [21] presents a DRL-based deep deterministic policy gradient (DDPG) approach to improve the EE in complex multi-user V2V communication.

This article has been accepted for publication in a future proceedings of this symposium, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/pimrc54779.2022.9978038.10270247, 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)

In [22], the authors present a multi-agent DRL approach to maximize the total sum rate of V2V and V2I users subject to URLLC requirements of V2V users. Authors in [23] proposed an approach based on proximal policy optimization (PPO) and DRL to manage the resources, where V2V pair acting as an agent learn the optimal policy by interacting with the environment and taking appropriate action to select the optimal power and sub-channel. However, most studies are based on actor-critic (AC), q-learning and DQN methods. All these methods can manage the resources intelligently but also have some limitations. The q-learning approach fails to converge rapidly. Moreover, the DQN approach may perform poorly due to the overestimation of Q-value, while AC converges to a local optimum and has high variance.

A. Contributions

In this work, we propose an efficient DRL-based DDQN approach with Thompson sampling to manage the radio resources intelligently in the IoV environment. We formulate to optimize the SE and EE of V2V and V2I users, while ensuring the URLLC requirements of V2V users. The main contributions are listed below.

- We propose a decentralized DRL-based transmit power control and resource block allocation approach where we try to optimize the weighted composite efficiency that incorporate both the EE and SE metrics. We aim to increase the composite efficiency of V2I and V2V users subject to URLLC requirements of V2V users in IoV networks.
- An efficient learning approach based on Thompson sampling named probability-based learning (PBL) is proposed to manage the resources intelligently. If the existing or new vehicular user experience poor performance then it can learn from the master user. This help to improve the overall system performance. The user that has the highest probability with the learning user is chosen as the master user.
- Simulation results shows that our proposed approach performs better in the dynamic IoV environment.

The remainder of this paper is organized as follows. In Section II, we present the system model and describe the optimization problem. In Section III, we describe the details of our proposed PBL approach to model the optimization problem. Simulation results are provided in Section IV. Lastly, we conclude the paper in Section V.

II. System Model

We consider the unicast communication scenario in the IoV network as shown in Fig. 1, where we have two kinds of communications: V2V and V2I communication. V2I communication indicates transmission between BS and vehicular users, and V2V communication is the direct communication between vehicular users. We assume that the orthogonal frequency division multiple access (OFDMA) scheme is adopted by the network to support vehicular users. There are M V2I users, denoted as $\mathcal{M} = \{1, 2, .., m, .., M\}$, and K V2V users, denoted as $\mathcal{K} = \{1, 2, .., k, .., K\}$, and orthogonal resource blocks are allocated to the V2I users. We assume that each V2V user can reuse only one resource block, and each resource block can be used by multiple V2V users because the interference is more manageable at the BS level. Hence, the resource block allocation element $s_{k,m}$ is such that

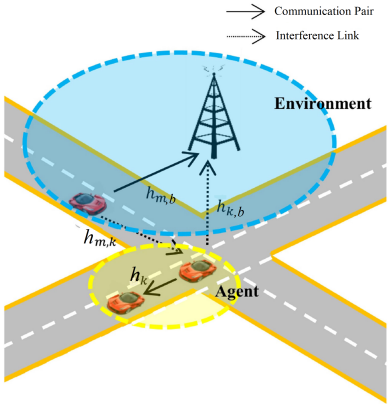

Fig. 1: System model of the unicast-based IoV networks.

$$s_{k,m} = \begin{cases} 1, & \text{If the } k^{th} \text{ pair reuses the } m^{th} \text{ pair resource} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The capacity of the m^{th} V2I pair can be represented as:

$$R_m = B \log_2 \left(1 + \frac{P_m h_{m,b}}{\sum_{k=1}^{K} s_{k,m} P_k h_{k,b} + \sigma^2} \right) \tag{2}$$

where B indicates bandwidth while the transmission power of m^{th} V2I pair and k^{th} V2V pair are P_m and P_k, respectively. Channel gain from m^{th} V2I pair to BS and from k^{th} V2V pair to BS is indicated by $h_{m,b}$ and $h_{k,b}$, respectively. The noise power is expressed by σ^2. The capacity of k^{th} V2V pair can be formulated as:

$$R_k = B \log_2 \left(1 + \frac{P_k h_k}{\zeta_k + \sigma^2} \right) \tag{3}$$

where ζ_k indicates the interference, and can be expressed as:

$$\zeta_k = \sum_{m=1}^{M} s_{k,m} P_m h_{m,k} + \sum_{m=1}^{M} \sum_{i \neq k}^{K} s_{k,m} s_{i,m} P_k h_{k,i} \tag{4}$$

where h_k and $h_{k,i}$ indicates the channel gain and interference gain of the i^{th} V2V pair. The first term in (4) denotes the interference power of the V2I pair sharing the same resource block with the V2V pair. The second term in (4) is the sum interference power from all V2V pairs sharing the same resource block.

A. Problem Statement

In this work, we aim is to maximize the SE and EE of V2I and V2V pairs while guaranteeing the strict URLLC requirements of V2V pairs. The resource allocation problem is formulated as an optimization-based problem. The composite

This article has been accepted for publication in a future proceedings of this symposium, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/pimrc54779.2022.9978038.10270247, 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)

efficiency for the V2I network is the ratio of the sum-rate of all the V2I pairs, bandwidth and the total power consumed for the V2I communications. Hence, it can be formulated as

$$\wp_m = \frac{\sum_{m=1}^{M} R_m}{\sum_{m=1}^{M} B_m \times (\sum_{m=1}^{M} P_m + MP_c)} \tag{5}$$

where B_m and P_c indicates the allocated bandwidth to the m^{th} V2I pair and circuit power, respectively. Similarly, the composite efficiency of V2V pair can be expressed as:

$$\wp_k = \frac{\sum_{k=1}^{K} R_k}{\sum_{m=1}^{M} B_m \times (\sum_{k=1}^{K} P_k + KP_c)} \tag{6}$$

Let τ_{trx} and $\tau_{\max}$ be the V2V user current transmission and the maximum transmission delays, respectively. The strict latency requirements of V2V users can be met such that the probability $\Pr\{\tau_{trx} \geq \tau_{\max}\}$ is less than the threshold value $\nu_{\max}^{delay}$, and it can be expressed as [25]:

$$\nu_k^{delay} = \Pr\{\tau_{trx} \geq \tau_{max}\} \leq \nu_{max}^{delay} \tag{7}$$

Similarly, the reliability constraint can be guaranteed by reducing the outage probability such that:

$$\nu_k^{outage} = \Pr\{R_k \leq R_k^{'}\} \leq \nu_{max}^{outage} \tag{8}$$

where $R_k^{'}$ indicates the target rate and $\nu_{\max}^{outage}$ is the rate outage threshold. So, the joint optimization problem can be stated as:

$$\mathbf{P}: \max_{P_m^{\star}, P_k^{\star}} \{w_1 \wp_m + w_2 \wp_k\} \tag{9a}$$

$$\text{subject to} \sum_{m=1}^{M} s_{k,m} \leq 1, \ \forall k; \tag{9b}$$

$$\Pr\{\tau_{trx} \geq \tau_{max}\} \leq \nu_{max}^{delay}, \ \in \{0,1\}, \forall k; \tag{9c}$$

$$\Pr\{R_k \leq R_k^{'}\} \leq \nu_{max}^{outage}, \ \in \{0,1\}, \forall k; \tag{9d}$$

$$P_m \leq P_{max}, \tag{9e}$$

$$P_k \leq P_{max}, \tag{9f}$$

where (9a) represents the objective function that aims to maximize the weighted composite efficiency (SE and EE) of V2I and V2V pairs. Constraint (9b) indicates that the k^{th} V2V pair can only reuse a single resource block of the V2I pair. Constraints (9c) and (9d) ensure that the strict URLLC requirements of the V2V pair are satisfied. (9e) and (9f) are the transmission power constraints of V2I and V2V users, respectively.

III. Intelligent Resource Management based on DRL

In this section, DRL-based resource management approach is discussed along with the proposed solution which is based on PBL. In this framework, every V2V pair acts as an agent and interacts with the environment to observe the behaviour. So, the agent observes the environment state s_t at each transmission time intervals (TTI), and selects resource block and transmission power by taking an action a_t based on the optimal policy π. The optimal policy π can be achieved by estimating the Q-function $Q(s_t, a_t)$ through DL. The environment gets to a new state s_t+1 according to the decisions or actions initiated by agents, and it achieves a reward value r_t. In this work, we evaluate the reward value by the weighted composite efficiency of the V2I and V2V pairs subject to URLLC requirements of the V2V pair.

1) State-space: The state space is essential for learning the optimal parameter settings. It can be represented as $\mathcal{S}$, and it consists of useful information of all agents observed at each TTI. At each TTI, the state-space s_t of an agent consists of CSI, selected resource blocks, QoS requirements, and the received interference. The CSI at t^{th} TTI over the m^{th} resource block for the k^{th} agent (V2V pair) can be expressed as $H_k^t[m] = \{h_k^t[m], h_{m,b}^t[m], h_{m,k}^t[m]\}$, where $h_k^t[m]$ and $h_{m,b}^t[m]$ indicates the channel gain of V2V pair and channel gain from V2I pair to BS over m^{th} resource block, respectively. And $h_{m,k}^t[m]$ represents the interference channel gain on the m^{th} resource block. The chosen m^{th} resource blocks captured by the neighbor vehicular users in previous TTI can be represented by $N^{t-1}[m]$. QoS requirements can be indicated by v_q which includes strict URLLC requirements. We present the received interference as $I^{t-1}[m]$. The state-space can be represented as:

$$s(t) = [H_k^t[m], N^{t-1}[m], I^{t-1}[m], v_q] \tag{10}$$

2) Action space: After observing the environment state each V2V agent will take action, which includes selecting of resource block and transmit power level, and it needs to satisfy the constraints (9b)-(9f) while taking these actions. The action space can be described as:

$$a(t) = \{\{N_1, ..., N_k\}, \{P_k(1), ..., P_k(k)\}\} \tag{11}$$

where N_k and $P_k(k)$ represents the resource block and transmission power level, respectively, selected by the agent.

3) Reward: A V2V agent chooses the resource block and appropriate transmission power level that leads to less interference to neighbouring vehicular users while meeting the URLLC QoS requirements. Therefore, an efficient reward function is required to determine the optimal policy π. The key objective is to take intelligent decisions and maximize the reward by learning the optimal policy π based on the observed environment state. So to manage the resources intelligently, we present the reward function by considering the QoS requirements in the IoV network. It can be represented as:

$$r(t) = (w_1 \wp_m + w_2 \wp_k) - w_3 \left(\sum_{k \in K} (\nu_k^{delay} + \nu_k^{outage}) \right) \tag{12}$$

where w_1, w_2 and w_3 are the weight values. Our reward function has two parts, namely, the composite efficiency of V2I and V2V pairs, and the URLLC requirement of V2V pairs. Part 1 indicates the objective function (SE and EE) and part 2 is the cost function in terms of URLLC requirements for V2V users. Since the key objective is to determine the optimal policy, so, to achieve better performance, immediate

This article has been accepted for publication in a future proceedings of this symposium, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/pimrc54779.2022.9978038.10270247, 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)

reward and future rewards need to be reviewed. The expected cumulative discounted reward can be expressed as:

$$R_t = \mathbf{E}\big[\sum_{j=0}^{\infty} \beta_j r_{t+j}\big] \tag{13}$$

where $\beta_j \in [0, 1]$ indicates the discount factor.

A. PBL for Intelligent Resource Management

The optimal policy π can be determined by applying different DRL techniques such as policy gradient (PG), Q-learning and deep Q-learning (DQL). In this work, we train our model to find the optimal policy to allocate the resource block and transmission power level to maximize the reward value. The Q-learning approach can be useful when state-action space has low dimension, because with large dimension it will update the Q-value rarely which will result in slow convergence [14]. In the IoV environment, the policy gradient-based approach converges to local optima and also suffers from high variance values which result in poor performance in terms of convergence [24]. Whereas, in DQN, every action is chosen and evaluated by the target Q-network which leads to an overestimation of Q-value [25]. We propose an efficient PBL approach based on DDQN coupled with Thompson sampling in order to find the optimal policy π. In DDQN, the action a_t is selected by the DQN and a target Q-network is used to evaluate that action. We use Thompson sampling to allow the V2V agent to select an action a_t based on the highest probability value. The V2V agent will not choose the actions with low probability which will lead to better convergence performance. Therefore, for Q-value $Q^{\pi}(s_t, a_t)$ the action that maximizes it can be formulated as:

$$a_t = \underset{a}{argmax} Q^{\pi}(s_t, a; \theta) \tag{14}$$

where θ represents the weight. In this work, we formulate an efficient exploration-exploitation approach by using Thompson sampling. Bayesian linear regression approach is used to estimate the distribution over Q-value. For each action a_t taken by the agent, we form a dataset with values from the replay memory. Then we form a matrix $\psi_{a_t}^{\theta}$ and estimate the posterior distribution. It can be expressed as:

$$\overline{w}_{a_t} = \frac{1}{\alpha^2} Cov_{a_t} \psi_{a_t}^{\theta} Q'^{\pi}(s_t, a_t; \hat{\theta}) \tag{15}$$

$$w_{a_t} \sim \mathcal{N}(\overline{w}_{a_t}, Cov_{a_t}) \tag{16}$$

where $\psi_{a_t}^{\theta}$ and $\overline{w}_{a_t}$ refers to the variance and mean, respectively. Samples are selected for every action a_t around the mean value and co-variance Cov. Once the V2V agent takes action a_t based on the observed environment state s_t, the reward value r_t and current state s_{t+1} will get back to the V2V agent. Then all these values are stored in replay memory. The output can be expressed as:

$$y \leftarrow r_t + \gamma Q'^{\pi}(s_t, a_t; \hat{\theta}) \tag{17}$$

where $Q'^{\pi}(s_t, a_t; \hat{\theta})$ indicates the target network that approximates the Q-value with action a_t. Whereas, DQN chooses the action a with a maximum Q-value as mentioned in (14). The Q-value is updated based on the estimation from the $Q'^{\pi}(s_t, a_t; \hat{\theta})$. The loss function can be expressed as follows:

$$\Delta(y, Q'^{\pi}(s_t, a_t; \hat{\theta})) = \mathbf{E}[(y - Q'^{\pi}(s_t, a_t; \hat{\theta}))^2] \tag{18}$$

B. Transfer learning

In transfer learning, learned knowledge is transferred to new task to improve the learning performance. We assume that there are two tasks: old task $T_o = (\mathcal{S}, \mathcal{A}, \mathcal{P}_1, r_1, \beta_1)$ and new task $T_n = (\mathcal{S}, \mathcal{A}, \mathcal{P}_2, r_2, \beta_2)$, and their respective optimal Q-functions are Q_1 and Q_2, respectively; where $\mathcal{P}_1$ and $\mathcal{P}_2$ are transition probabilities. We also assume that both these tasks have similarities. The aim is to improve the performance of the system in dense environment by using the learned knowledge of T_o and Q_1. We propose a novel approach to improve the model performance and efficiency to determine the optimal policy. The vehicular users who are experiencing poor performance can learn from neighbouring vehicles in the IoV network. Newly joined vehicular users can also learn information from the master vehicular user instead of creating its own model. It needs to select the neighbouring vehicular user as a master to learn the information which includes resource block selection and transmission power level while exchanging the information such as QoS requirements, CSI and received interference. The vehicular user who has the highest probability with the learning user is chosen as the master. It can be defined as:

$$\triangle(T_o, T_n) = ||Q_1 - Q_2||_{\infty} \leq \hat{\triangle}(T_o, T_n) \tag{19}$$

PBL optimizes the state-value action function $Q^{\pi}(s_t, a_t, \theta)$ according to the transfer information. The weight parameter

Algorithm 1 PBL for intelligent resource allocation

```
Input: Environment simulator
Initialize: weight θ, target weight θ̂, policy π and model
Set memory= ∅
for each TTI do
    Observe the state space s_t
    Take action a_t for the selection of resource block and
    power level following the Thompson sampling
    Determine reward r(t)
    Save data into memory
    Sample and train with DDQN
    Update policy π
    Update parameter θ by minimizing a loss function
    if Vehicular user has poor performance then
        Find the master with highest probability
        Obtain learned strategy and create action space
        Choose action according to equation (20)
        Perform steps from (7) to (11)
        else
        Perform steps from (5) to (11)
    end if
end for
```

This article has been accepted for publication in a future proceedings of this symposium, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/pimrc54779.2022.9978038.10270247, 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)

θ will be optimized, and the transfer learning parameters will converge to the state-value action function and find an optimal policy π. After finding the master with the highest probability, it uses the learned action information from the master and its action space to create a new range of action space. Then the agent employs Thompson sampling to select the actual action space based on the observed environment. The chosen action can be expressed as:

$$\hat{a} = ca_l + (1-c)a_c \tag{20}$$

where c refers to the transfer rate $c \in [0,1]$, a_l and a_c indicate the learned action information and current action space, respectively. Details are mentioned in Algorithm 1. Mostly, V2V agent takes action based on its observed information and will not share the learned strategy with other V2V pairs if the V2V user is not experiencing any performance loss as it does not require to use the transfer learning framework from any master user. Our proposed decentralized transfer learning framework improves the performance of the system because the V2V agent will not continuously transfer the knowledge as it can learn based on its own observation.

IV. SIMULATION RESULTS

In this section, simulation results are provided to show the performance of the proposed algorithm in terms of maximizing the composite efficiency, which integrates both the EE and SE metrics subject to URLLC requirements. The carrier frequency is 2 GHz and we set the simulation setup according to the urban case of 3GPP TR 36.885 [26]. We generate the data samples from the environment where V2I and V2V users are scattered randomly in a cell. We obtain the CSI of all users in the network according to their present positions. Large-scale fading and small-scale fading is also considered, and we divide the channels into non-line-of-sight and line-of-sight. In order to meet the latency and reliability constraints, we set the $\tau_{max} = 5ms$, and reliability 99.999 %, respectively with the 1 bits/Hz target SE. All the parameters are shown in Table I. We have three hidden layers with a learning rate of 0.01 and a discount factor of 0.75. We compare our simulation results

TABLE I: Simulation Parameters

Parameter	Value
Carrier frequency	2 GHz
Resource block bandwidth	180 KHz
Cell radius	400 m
Bandwidth	5 MHz
No. of resource blocks	25
V2I power	23 dBm
V2V power level	[23, 15, 10] dBm
Circuit power	15 dBm
No. of V2I users	20
No. of V2V users	[20, 40, ..., 80]
Noise power	-107 dBm
Learning rate	0.01
Distance between vehicles	100 m
Path Loss	LOS & NLOS
Channel	Rayleigh fading
Shadow distribution	Log-normal

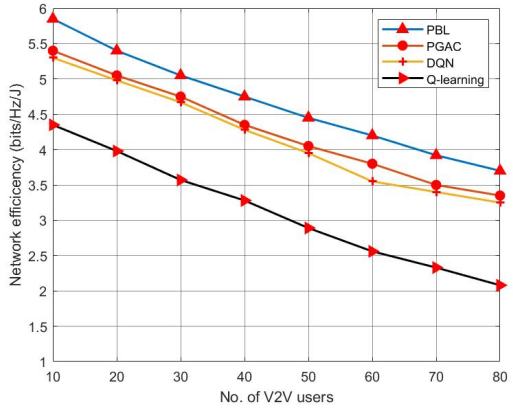

Fig. 2: Composite efficiency against number of V2V users.

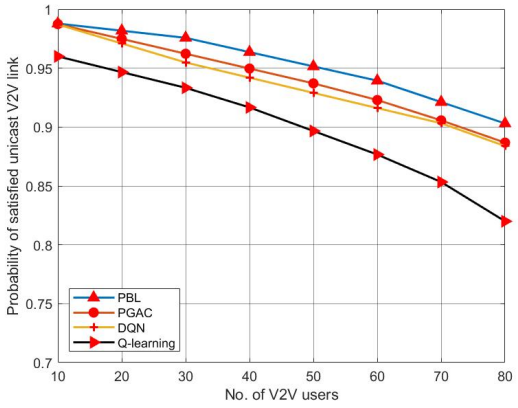

Fig. 3: Probability of satisfying V2V pairs.

with actor-critic learning based on the PG approach, DQN and Q-learning.

In Fig. 2, the result of composite efficiency is shown for the different schemes with an increasing number of V2V users, and when the vehicle speed is 50 km/h. It can be seen that the proposed scheme performs better than PGAC and DQN. This is because our proposed algorithm adapts the policy π in order to manage the resources intelligently and efficiently reducing the interference to V2I pairs. And as the number of users are increased, the composite efficiency of V2I and V2V pairs decreases due to the increasing interference and power consumption. The value is 5.12 bits/Hz/J for the PBL when the vehicular users are 30, whereas the values of PGAC and DQN are 4.75 bits/Hz/J and 4.67 bits/Hz/J, respectively. Q-learning technique performs poorly due to its slow convergence. When the number of vehicular users is increased to 80, the composite efficiency also decreases with it. The PBL approach performs better than the rest of the schemes.

The result showing the probability of satisfying V2V pairs is shown in Fig. 3., where the speed of the vehicular user is 50 km/h. It can be observed that PBL achieves better performance by finding the optimal resource management policy to meet

This article has been accepted for publication in a future proceedings of this symposium, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/pimrc54779.2022.9978038.10270247, 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)

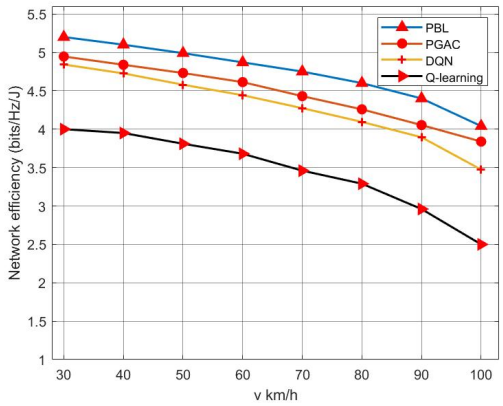

Fig. 4: Network efficiency with varying speed.

the URLLC requirements. As the number of V2V users is increased the network efficiency decreases with it. This is because as the user pairs are increased they are all required to establish the connection, and need to maximize the objective function while guaranteeing the URLLC requirements. Hence, increasing the number of pairs leads to toppling the satisfied pairs.

Fig. 4 shows the effect of the varying speed on the network efficiency when the number of vehicular users is 30. It can be seen that the network efficiency decreases with the increase in user speed. This is due to the influence of the larger observation variability from the dynamic environment. Our proposed approach outperforms the other schemes initially and then its performance drops as the increase of speed. This is because of the larger quantized action values which results in larger dimension.

V. Conclusion

In this work, we investigated the resource block and power allocation mechanism in IoV networks. We proposed a novel DRL-based approach to improve the network efficiency while meeting the URLLC requirements in a highly dynamic and resource limited environment. We optimized the composite efficiency, which incorporates the SE and EE metric while ensuring the QoS requirements of V2V pairs. Simulation results shows that the proposed PBL approach performs better while meeting the QoS requirements.

References

[1] O. Kaiwartya, A. H. Abdullah, Y. Cao, A. Altameem, M. Prasad, C. T. Lin, & X. Liu. "Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects". IEEE Access, 4, 2016.

[2] L. Liang, H. Peng, G. Y. Li, and X. Shen, "Vehicular communications: A physical layer perspective," IEEE Trans. Veh. Technol., vol. 66, no. 12, pp. 10647–10659, Dec. 2017.

[3] C. Wang, C. Chou, P. Lin, and M. Guizani, "Performance evaluation of IEEE 802.15.4 non beacon-enabled mode for internet of vehicles," IEEE Transaction on Intelligent Transportation System, vol. 16, no. 6, pp. 3150–3159, Dec. 2015.

[4] W. Saad, Z. Han, A. Hjorungnes, D. Niyato, and E. Hossain, "Coalition formation games for distributed cooperation among roadside units in vehicular networks," IEEE J. Sel. Areas Commun., vol. 29, no. 1, pp. 48–60, Jan. 2011.

[5] H. Seo, K. D. Lee, S. Yasukawa, Y. Peng, and P. Sartori, "LTE evolution for vehicle-to-everything services," IEEE Commun. Mag., vol. 54, no. 6, pp. 22–28, Jun. 2016.

[6] J. Cheng, J. Cheng, M. Zhou, F. Liu, S. Gao, & C. Liu, "Routing in internet of vehicles: a review," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 5, pp. 2339–2352, Oct. 2015.

[7] J. Mei, K. Zheng, L. Zhao, Y. Teng, & X.Wang, "A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems," IEEE Trans. Wireless Commun., vol. 17, Jun. 2018.

[8] Z. Liu, Y. Han, J. Fan, L. Zhang and Y. Lin, "Joint Optimization of Spectrum and Energy Efficiency Considering the C-V2X Security: A Deep Reinforcement Learning Approach," 2020 IEEE 18th International Conference on Industrial Informatics (INDIN), 2020, pp. 315-320, doi: 10.1109/INDIN45582.2020.9442103.

[9] Q. Wen, B. Hu and L. Zheng, "Outage-Constrained Device-to-Device Links Reuse Maximization and Its Application in Platooning," in IEEE Wireless Communications Letters, vol. 8, no. 6, pp. 1635-1638, Dec.2019.

[10] W. Sun, E. G. Str¨om, F. Br¨annstr¨om, K. C. Sou, and Y. Sui, "Radio resource management for D2D-based V2V communication," IEEE Trans. Veh. Technol., vol. 65, no. 8, pp. 6636–6650, Aug. 2016.

[11] M. I. Ashraf, M. Bennis, C. Perfecto, and W. Saad, "Dynamic proximityaware resource allocation in vehicle-to-vehicle (V2V) communications," in Proc. IEEE Globecom Workshops, Dec. 2016, pp. 1–6.

[12] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," IEEE Commun. Surv. Tut., vol. 21, no. 3, pp. 2224–2287, Mar. 2019.

[13] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," Proc. IEEE, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[14] H. Ye, G. Y. Li and B. F. Juang, "Deep Reinforcement Learning Based Resource Allocation for V2V Communications," in IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3163-3173, April 2019.

[15] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," IEEE Trans. Veh. Technol., vol. 67, no. 1, pp. 44–55, Jan. 2018.

[16] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Deep reinforcement learning based mode selection and resource allocation for cellular V2X communications," IEEE Internet Things J., vol. 7, pp. 6380–6391, Jun, 2020.

[17] W. Liu, G. Qin,Y. He, and F. Jiang, "Distributed cooperative reinforcement learning-based traffic signal control that integrates V2X networks dynamic clustering," IEEE Trans. Veh. Technol., vol. 66, no. 10, pp. 8667–8681, Oct. 2017.

[18] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, and M. Wu, "Machine learning for vehicular networks: Recent advances and application examples," IEEE Veh. Technol. Mag., vol. 13, no. 2, pp. 94–101, Jun. 2018.

[19] Z. Guan, Y. Wang, and M. He, "Deep Reinforcement Learning-Based Spectrum Allocation Algorithm in Internet of Vehicles Discriminating Services," Applied Sciences, vol. 12, no. 3, p. 1764, Feb. 2022, doi: 10.3390/app12031764

[20] H. Zhang, X. Liu, X. Bian, Y. Cheng, S. Xiang, "A Resource Allocation Scheme for Real-Time Energy-Aware Offloading in Vehicular Networks with MEC", Wireless Communications and Mobile Computing, vol. 2022, Article ID 8138079, 17 pages, 2022.

[21] Y. Zhang, D. Lan, C. Wang, P. Wang and F. Liu, "Deep Reinforcement Learning-aided Transmission Design for Multi-user V2V Networks," 2021 IEEE Wireless Communications and Networking Conference (WCNC), 2021, pp. 1-6, doi: 10.1109/WCNC49053.2021.9417249.

[22] F. Zhou, L. Feng, P. Yu, W. Li, X. Que and L. Meng, "DRL-based Low-Latency Content Delivery for 6G Massive Vehicular IoT," in IEEE Internet of Things Journal, 2021, doi: 10.1109/JIOT.2021.3064874.

[23] X. Hu et al., "A joint power and bandwidth allocation method based on deep reinforcement learning for V2V communications in 5G," in China Communications, vol. 18, no. 7, pp. 25-35, July 2021.

[24] H. Yang, X. Xie and M. Kadoch, "Intelligent Resource Management Based on Reinforcement Learning for Ultra-Reliable and Low-Latency IoV Communication Networks," in IEEE Transactions on Vehicular Technology, vol. 68, no. 5, pp. 4157-4169, May 2019, doi: 10.1109/TVT.2018.2890686.

[25] H. van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-Learning", AAAI, vol. 30, no. 1, Mar. 2016.

[26] 3rd Generation Partnership Project: Technical Specification Group Radio Access Network: Study LTE-Based V2X Services: (Release 14), Standard 3GPP TR, Jun. 2016.