Deep Reinforcement Learning for Time Allocation and Directional Transmission in Joint Radar-Communication

Joash Lee[1], Yanyu Cheng[2], Dusit Niyato[2], Yong Liang Guan[3], David González G.[4]
[1] Energy Research Institute @ NTU, Nanyang Technological University, Singapore
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[4] Advanced Connectivity Technologies Group, Continental AG, Germany

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771580

***Abstract*—Current strategies for joint radar-communication (JRC) rely on prior knowledge of the communication and radar systems within the vehicle network. In this paper, we propose a framework for intelligent vehicles to conduct JRC, with minimal prior knowledge, in an environment where surrounding vehicles execute radar detection periodically, which is typical in contemporary protocols. We introduce a metric on the usefulness of data to help the vehicle decide what, and to whom, data should be transmitted. The problem framework is cast as a Markov Decision Process (MDP). We show that deep reinforcement learning results in superior performance compared to non-learning algorithms. In addition, experimental results show that the trained deep reinforcement learning agents are robust to changes in the number of vehicles in the environment.**

***Index Terms*—Deep reinforcement learning, resource allocation, joint radar-communication**

I. Introduction

As the task of driving becomes increasingly autonomous, vehicles are being equipped with a growing number and variety of sensors. Vehicles are also being developed to drive cooperatively with each other [1], which will necessitate high data-rate transmission of sensory and perception information between vehicles in the order of gigabits per second [2].

The use of the millimeter-wave (mmWave) band is often regarded as a method to meet the demand for high data-rate vehicular communication [2], [3]. However, commercialization of mmWave technologies has raised concerns over interference with automotive radar sensing [3]. Methods to jointly conduct radar sensing and communication functions within a shared frequency range are referred to as joint radar-communication (JRC). Existing JRC strategies can be categorized into (i) methods that use dual-function signals, and (ii) coordinated transmission of dedicated radar and communication signals through division in the time, frequency or spatial dimensions [4]. A major advantage of the second category is that existing hardware for both radar and communication can be used.

A straightforward approach to coordinated transmission is to allocate portions of time or frequency in a fixed manner. More adaptive methods such as deep reinforcement learning (DRL) were proposed in [5]–[7]. Key advantages of the learning-based approach are that prior knowledge of system parameters is not required, and that the system can respond to instantaneous changes in the environment. Division of space, on the other hand, can be achieved through the use of beamforming. Beamforming has been separately proposed for both radar detection and communication methods [8], [9]. However, at the time of writing, the authors are not aware of a unified beamforming technique for JRC that maximizes the signal-to-noise ratio (SNR) of the network users by exploiting their changing spatial locations.

In a cooperative driving setting, the need to transmit high data-rates within the constraints of the allocated frequency spectrum, time and space raises further challenges. Cooperating vehicles must decide what, and to whom, information should be transmitted at any given point in time. These decision-making problems may be guided by the "age of information" ("AoI"), defined as the duration of time since the last received packet was generated [10], as a performance metric. In this paper, we are concerned with a further aspect of data packets which we term the *spatial signature*: the discretized spatial location represented by each packet of data.

The main challenge for intelligent vehicles in the near future is the competition for wireless resources imposed by radar sub-systems on surrounding vehicles that operate periodically and independently. Periodic operation of sensory systems, typical on contemporary vehicles, potentially causes noise to the mmWave communication of the ego vehicle. Consequently, in this paper, we develop a framework for an intelligent vehicle to conduct JRC in an environment where surrounding vehicles execute radar periodically.

We propose the framing of our JRC problem as a Markov Decision Process (MDP) where the ego vehicle simultaneously decides: (i) how to divide time for JRC, (ii) the content of data to be transmitted, and (iii) the direction in which the chosen data is transmitted. To aid decision-making, we propose a metric for the usefulness of a data packet to the receiving vehicle in terms of the data's AoI and spatial signature. The overall objective is for the ego vehicle to maximize the transmission of useful sensory information (based on our proposed metric), subject to radar noise generated by neighboring vehicles. We apply DRL to maximize performance based on the mentioned objective. Our approach requires minimal prior-knowledge of

This research is a collaboration between the Energy Research Institute @ NTU, the Continental-NTU Corporate Lab, and the Computer Networks and Communications Lab in Nanyang Technological University (NTU). This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s)."

the vehicular network, scales linearly with the number of vehicles, and is applicable to cooperative driving settings.

The paper is organized as follows. We review related work in Section II that motivates the system model we propose in Section III. In Section IV, the system is then formulated as an MDP. Methods to solve the MDP are proposed in V. Finally, the performance of these methods are evaluated in Section VI.

II. Related Work

In this section, we review literature in vehicle sensing and perception, and machine learning for wireless resource allocation. Based on this, we propose a unified framework for intelligent time allocation and data sharing for JRC.

A. Vehicle Sensing and Perception

The types of sensors fitted to autonomous vehicles include cameras, LIDAR and radar. Sensory data is combined through "sensor fusion", which results in more certainty [11]. Certainty in environmental perception can be gained through a number of mechanisms which we review below.

Complementary strengths: Different sensors have different strengths. For example, LIDAR and radar are better at measuring the distance of an object to the vehicle, whereas camera systems are better at object recognition [12]. Radar sensing also excels at measuring the velocities of surrounding objects, and is robust to deterioration in visibility due to adverse weather conditions such as fog or heavy precipitation [12].

Sampling rate: Methods to adjust the sampling rate of individual sensors have been proposed by [13]–[15] and reviewed in [16]. The general strategy is to increase the sampling rate of a sensor node when it observes an interesting event [13], [14], or when a proposed error criterion exceeds a predetermined bound [15]. Motivating factors for adaptive sampling rates include constraints on energy use, overall system bandwidth, and computational power. JRC in an automotive setting is similarly constrained by bandwidth availability.

Sharing data: Sensors on nearby vehicles can provide the vehicle of concern with higher sensor redundancy, or even reveal data on objects that are fully or partially obstructed from view. We illustrate an example in Figure 1.

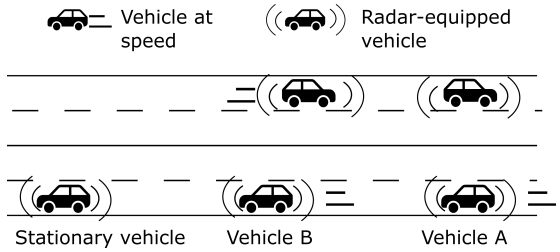


Fig. 1: A schematic of the dual carriageway environment

Unfortunately, camera and LIDAR sensors produce a high data rate of $100 - 700$ Mb/s and $10 - 100$ Mb/s respectively [2]. This means that current 4G systems are inadequate for the transmission of raw sensory data by $O(10)$. On the other hand, the use of mmWave systems will likely enable sharing of sensory data on the order of gigabits per second.

B. Machine Learning for Wireless Resource Allocation

The management of finite wireless resources is often formulated as an optimization problem to maximize throughput, transmission power, or more recently, AoI. These approaches require an accurate mathematical model of the system. Such information may be difficult to identify in practice, or result in non-convex or high-dimensional problems [17]. As such, recent studies use deep neural networks, which allow for the learning of high-dimensional problems with minimal prior knowledge of system parameters. Framing the problem under the reinforcement learning paradigm offers further advantages. While supervised learning requires a dataset of correct responses to environmental inputs, reinforcement learning relies only on a reward signal given by the system model.

III. System Model

We consider a vehicular system consisting of a set $\mathcal{N}$ of $N = |\mathcal{N}|$ vehicles traveling along a dual-carriageway. All vehicles within each lane travel at the same speed, with vehicles on the outer lane traveling at higher speeds. All vehicles have the same radar and sensory capabilities. As time advances, each vehicle collects data through its sensory perception systems, and stores the data in its memory.

At each time step, each vehicle chooses between transmitting sensory data and radar detection. If a vehicle decides to communicate, it chooses which data in its memory to transmit, and in which direction (through beamforming). Any received data deemed as useful is stored in the memory of the receiving vehicle, and may be re-transmitted in subsequent time steps. Consequently, the task of communication can be viewed as a message-passing problem.

In our proposed system, radar sub-systems on surrounding vehicles that operate periodically and independently, as is typical on contemporary vehicles. This causes potential interference to the mmWave communication of the ego vehicle. In contrast, our proposed intelligent vehicle learns to schedule its radar and communication operation based on its environment. The system objectives of our intelligent JRC system are:

1) To decide when to perform radar detection instead of communication, with the aim of obtaining more accurate measurements of the speeds of surrounding objects.
2) To predict which data is the most useful to transmit.
3) To predict the direction of vehicles that would find the transmitted data most useful.
4) To choose the best time step to maximize the SNR for communication.

The system, from the perspective of the ego vehicle, is modeled as an MDP, which we describe in Section IV. The mentioned system objectives are expressed mathematically in the form of the reward function of the MDP.

IV. Problem Formulation

Our proposed problem of real-time radar-communication scheduling data sharing is formulated as an MDP. Details on the observation space, environment transition model, action space and reward function are described further in this section.

Since each vehicle in the environment acts independently, we also refer to them as agents. The intelligent vehicle of concern is referred to as the "ego vehicle".

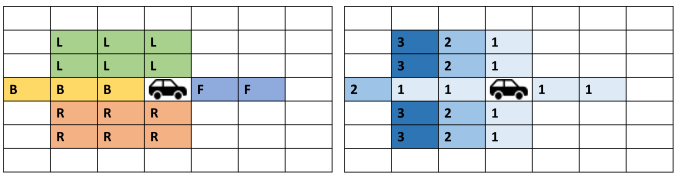
A. Observation

The ego vehicle's observation $\mathbf{o}$, as defined in (1), consists of several features describing the local environmental condition, the configuration of its surrounding vehicles, as well as the sensory data accumulated by the ego vehicle.

$$\mathbf{o} = [\alpha, \mathbf{e}, \mathbf{x}, \mathbf{v}, \mathbf{d}, \mathbf{d}_\theta]. \tag{1}$$

The scalar quantity α represents the age of radar information last gathered by the ego vehicle (i.e. the number of time steps elapsed since the vehicle last chose to perform radar detection). The environmental vector $\mathbf{e} = \{w, m, v\}$ contains three environmental features that are indicative of how important it is to perform radar detection at a given time step. The feature w indicates the weather condition, m indicates the presence of moving objects nearby, and v is the speed of the ego vehicle. Each environmental feature is ranked on an integer scale such that $\{w, m, v\} \in \{0, 1, \ldots, E\}$. A ranking of zero indicates the safest possible condition, while higher values represent increasing levels of risk. For example, $w = 0$ would indicate clear and sunny weather with excellent visibility, while higher values indicate poor visibility due to weather conditions such as heavy precipitation or fog. Higher values of m and v would indicate the presence of a moving object nearby, and high speed of the ego vehicle respectively.

The vectors $\mathbf{x} \in \mathbb{R}^{2(N-1)}$ and $\mathbf{v} \in \mathbb{R}^{N-1}$ represent the positions and velocities respectively of each of the other $N-1$ vehicles, relative to the ego vehicle. The position of each vehicle is two-dimensional across the plane of the road, while the velocity is one-dimensional since we consider only one-dimensional travel along a straight road.



(a) (b)

Fig. 2: Sample schematics of (a) $\mathbf{d}_\theta$ and (b) $\mathbf{d}$

The features $\mathbf{d}$ and $\mathbf{d}_\theta$ represent scalar fields discretized in the spatial domain that indicate scalar quantities associated with sensory data available in the memory of the ego vehicle. Represented as two-dimensional arrays, the origins of $\mathbf{d}$ and $\mathbf{d}_\theta$ correspond to the instantaneous position of the ego vehicle. The array $\mathbf{d} \in \mathbb{R}^{A \times B}$ indicates the AoI corresponding to locations for which the ego vehicle has sensory data. Data exceeding a maximum age of α_{max} is considered to be expired and is deleted. Thus, each element of the array $d_{ab} \in [0, \alpha_{max}]$. Feature $\mathbf{d}_\theta$ describes the azimuth, relative to the traveling direction of the ego vehicle, of each data point. The azimuth of the available data is discretized according to the cardinal directions with respect to the orientation of the ego vehicle, so the space for this feature is $d_{\theta,ab} \in \{F, R, B, L, null\} = D_\theta \cup null$, where F, R, B and L represent the directions front, right, back and left respectively, and $null$ represents the absence of data for the coordinate ab.

B. Action

The set of actions $\mathcal{A}$ comprises radar detection a_{radar} and combinations of decisions on which data to transmit a_d, and in which direction to transmit the data a_θ, along with a null action. The set of choices of data to transmit a_d matches the discrete set of data azimuth states $\mathbf{d}_\theta$. Points in the vehicle's memory with feature $\mathbf{d}_\theta$ that matches the action choice a_d are transmitted. Expressed mathematically, the features of the data transmitted by a vehicle are:

$$\mathbf{d}_\theta^{<send>} = \mathbf{1}_{a_d}(\mathbf{d}_\theta), \tag{2}$$

$$\mathbf{d}^{<send>} = \mathbf{d}_\theta^{<send>} \circ \mathbf{d}, \tag{3}$$

where $\mathbf{1}_{a_d}(\cdot)$ is an indicator function for values equal to a_d, and the operator $\circ$ is the Hadamard or element-wise product.

The direction of data transmission a_θ is selected from a discrete set, based on the hardware capabilities of the vehicle. The entire action set can thus be represented as $\mathcal{A} = \{(a_d, a_\theta) | a_d \in \mathcal{A}_d \text{ and } a_\theta \in \mathcal{A}_\theta\} \cup \{(a_{radar}, null)\}$.

When the data selected for transmission by vehicle i is directed towards neighboring vehicle j, the scalar fields associated with transmitted data are translated by the positional difference between vehicles i and j, such that the origin of the data received by vehicle j matches its instantaneous position. The features of the data received by vehicle j can be expressed mathematically as:

$$\mathbf{d}_{ji}^{<rec>}(x_1, x_2) = \mathbf{d}_{ij}^{<send>}(x_1 + x_{1,ji}, x_2 + x_{2,ji}), \tag{4}$$

$$\mathbf{d}_{\theta,ji}^{<rec>}(x_1, x_2) = \mathbf{d}_{\theta,ij}^{<send>}(x_1 + x_{1,ji}, x_2 + x_{2,ji}), \tag{5}$$

where x_1 and x_2 are the two-dimensional coordinates of the data field, and $x_{1,ji}$ is the position of vehicle j with respect to vehicle i. The total data received by vehicle j from all the vehicles in the environment is the sum:

$$\mathbf{d}_j^{<rec>} = \sum_{i \in \mathcal{N} \setminus j} d_{ij}^{<send>}. \tag{6}$$

C. Transition Model

Vehicle headway is the duration of time between successive vehicles passing a given point on a lane [18]. We assume free-flowing traffic that is identically and independently distributed over the length our road-segment of concern, and so model the arrival of vehicles using a Poisson distribution [18]. By assuming that each time step of the MDP is sufficiently small, the arrival of vehicles into the environment can be modeled by a Bernoulli distribution. The arriving vehicles are uniformly distributed across all lanes in the environment.

As a vehicle advances forward into the next time step, its data increases in age by one, and all data features are translated by the distance traversed (since their origins are relative to the vehicle's instantaneous position). New data may be acquired

by the vehicle's sensory and perception systems, or received from a neighboring vehicle. Expressed mathematically, the transition function of the data age feature is:

$$\begin{aligned}\mathbf{d}_{t+1}(x_1, x_2) =&(\mathbf{d}_t + \mathbf{1}_{D_\theta}(\mathbf{d}_{\theta,t}))((x_1 + v_t \times \Delta t), x_2) \\ &+ \mathbf{d}_t^{<new>} + \mathbf{d}_t^{<rec>}\end{aligned} \tag{7}$$

$$\begin{aligned}\mathbf{d}_{\theta,t+1}(x_1, x_2) =&\mathbf{d}_{\theta,t}((x_1 + v_t \times \Delta t), x_2) \\ &+ \mathbf{d}_{\theta,t}^{<new>} + \mathbf{d}_{\theta,t}^{<rec>}\end{aligned} \tag{8}$$

where $\mathbf{d}_{t+1}$ is the data age feature at time t, $((x_1 + v_t \times \Delta t), x_2)$ indicates a translation of the data age scalar field by the distance $v_t \times \Delta t$ traveled in the last time step, $\mathbf{d}_t^{<new>}$ is the scalar field of ages of newly perceived data on the environment, and $\mathbf{d}_t^{<rec>}$ is the scalar field of the age of data received from neighboring vehicles.

When a transmission arrives at a targeted vehicle j, the power received from the ego vehicle i is given by:

$$P_{ij} = \frac{P_t G^2 \lambda^2}{(4\pi)^2 x_j^2}, \tag{9}$$

where P_t is the transmission power at the ego vehicle (the source), G is the transmission and receiving antenna gain, λ is the transmission wavelength, and x_j is the position of vehicle j relative to the ego vehicle.

At the receiving device on the receiving vehicle, the SNR for the signal measured at the receiving vehicle j is:

$$SNR_{ij} = \frac{P_{ij}}{\sum_{k \in \mathcal{N} \setminus i} \sigma_k}, \tag{10}$$

where $\sum_{k \in \mathcal{N} \setminus i} \sigma_k$ is the sum of radar noise produced by other vehicles k. We model communication under a Binary PSK scheme, and the interfering radar noise as additive white Gaussian noise (AWGN). It follows that the bit error rate (BER) of the received signal of concern may be given by the following equation:

$$BER_{ij} = Q(\sqrt{2 \times SNR_{ij}}), \tag{11}$$

where $Q(\cdot)$ is the Q-function. For our problem, we can alternatively interpret the success rate of transmissions as:

$$\eta_{ij} = (1 - BER_{ij}). \tag{12}$$

D. Reward

The reward at each time step t is a weighted sum of two parts. The first part of the summation encourages the vehicle to maximize SNR, while also transmitting the most useful data with the lowest possible age. The last term, r_{radar}, encourages the vehicle to perform radar detection when necessary.

$$r(\mathbf{o}, (a_d, a_\theta)) = w_{comm}(r_{SNR} \times r_{data}) + w_{radar} r_{radar}, \tag{13}$$

where w_{comm} and w_{radar} are weights that we tune to balance the relative importance of communication and radar functions.

The term r_{SNR} encourages the ego vehicle i to maximize the SNR of its data transmission at the intended receivers.

$$r_{SNR} = \sum_{j \in \mathcal{N} \setminus i} \eta_{ij}. \tag{14}$$

The term r_{data} encourages the ego vehicle to send data that is more useful to other vehicles, and is defined as follows:

$$\begin{aligned}r_{data} = \sum_{j \in \mathcal{N} \setminus i} v_j \Big(&\sum_{a,b} (W_j)_{ab} \circ \\ &\big(\max(0, \alpha_{max} - (d_j^{<rec>})_{ab} \big) \Big),\end{aligned} \tag{15}$$

where $\mathbf{W}_j$ is weight array representing how much the receiving vehicle j values data at each positional coordinate (a, b). The level of risk v_j is linked to vehicle j's individual state, and describes how much it values help in perceiving its environment. In our model, we consider increased vehicle speed to result in an increased requirement for sensory data.

In our simulated environment, the importance of performing radar sensing is linked to the environmental condition as described by the local state vector $\mathbf{e}$, and the age of radar data gathered by the ego vehicle:

$$r_{radar} = -\mathbf{1}_{a \neq a_{radar}} \big(\exp(\mathbf{e}\boldsymbol{\beta}) \times f(\alpha) \big), \tag{16}$$

where $\mathbf{1}_{a \neq a_{radar}}$ is the indicator function for actions that are not radar detection a_{radar}, $\boldsymbol{\beta}$ is a coefficient vector, and $f(\alpha)$ is a non-decreasing function of α. The formulation for r_{radar} shows that when a communication action is chosen instead of radar detection, a higher penalty is received when the environmental features indicate more unfavorable conditions. Similarly, a higher penalty will be received as the time from the previous radar detection α increases, since this results in higher uncertainty on the perceived environment. The exact function $f(\alpha)$ can be selected depending on the system parameters. In our experiments, we use $f(\alpha) = \mathbf{1}_{\mathbb{Z}^+}(\alpha)$, a formulation known to encourage reduction in the average value of the quantity of concern (i.e. α) in reinforcement learning.

V. Method

We propose solving the problem introduced in Section IV with the DRL algorithms Advantage Actor Critic (A2C) and Proximal Policy Optimization (PPO). They are compared with non-learning algorithms that resemble contemporary protocols on sensing and communication. Both A2C and PPO are chosen because they do not rely on the Markov assumption and knowledge of the entire system state s.

A. Deep Reinforcement Learning

Advantage Actor Critic (A2C): The goal of reinforcement learning is to learn to make decisions within an MDP, so as to maximize the expected sum of discounted rewards:

$$J(\theta) = E_{s \sim Pr^{\pi_\theta}, a \sim \pi_\theta} \left[\sum_{t=i}^{T} \gamma^{t-i} r_t \right], \tag{17}$$

where Pr^π is the probability distribution of states under a policy π_θ, γ is the discount rate, and r_t is the reward at time t as defined in 13. The A2C algorithm directly minimizes this objective function. It is described in detail by [19].

Proximal Policy Optimization (PPO): Policy-based algorithms such as A2C are known to have high variance, resulting in fluctuations in the reward obtained. PPO [20] counters this

problem by limiting the size of each update through optimizing a surrogate objective function instead. Theoretical assurances on convergence are discussed in [20].

B. Benchmark Non-learning Algorithms

Round Robin: This non-learning algorithm acts independently of the agent's environmental state, imitating how an industrially available system might operate. At every k^{th} time step, the agent performs omni-directional radar detection. For the $(k-1)$ time steps in between, the agent cycles through the set of all possible actions. An example action sequence is $\{a_{t=1}, a_{t=2}, \ldots\} = (F,F),(F,R),F,B),(F,L),\ldots\}$. The other $N-1$ non-ego vehicles in the environment are programmed to act according to this algorithm.

Heuristic: In this algorithm, the agent chooses its transmission direction a_θ such that it transmits to the nearest possible vehicle. The agent selects the transmitted data to be $a_d = a_\theta$. This is based on the heuristic logic that the receiving vehicle, if positioned to the left of the agent, would prefer receiving data that was also gathered to the left of the transmitting agent.

VI. Experiments

Firstly, in Section VI-A, the performance of our proposed methods is evaluated in terms of the reward attained in each episode in an 8-vehicle environment. In both the PPO and A2C algorithms, the architecture of the neural networks are set to be identical. Subsequently, in Section VI-B, we examine the robustness of the trained DRL agents to changes in the number of vehicles in the environment.

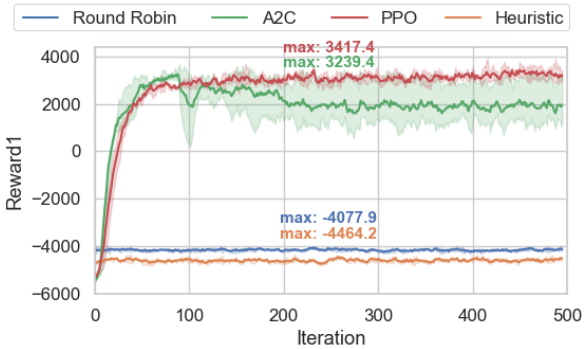


Fig. 3: Average total reward r per episode for each iteration of the training process.

A. Performance

The training progress in terms of total reward achieved is shown in Figures 3 to 5. The overall performance of the round robin agent, as measured by total reward shown in Figure 3 is representative of the reward achieved by the remaining 7 vehicles, since they are configured to operate with the same policy. Compared to the round robin agent, the agent following our heuristic policy achieves a slightly lower reward. As shown by the communication reward r_{comm} in Figure 5, the heuristic policy is able to make better decisions on what data should be transmitted to which vehicle. However, without a strategy

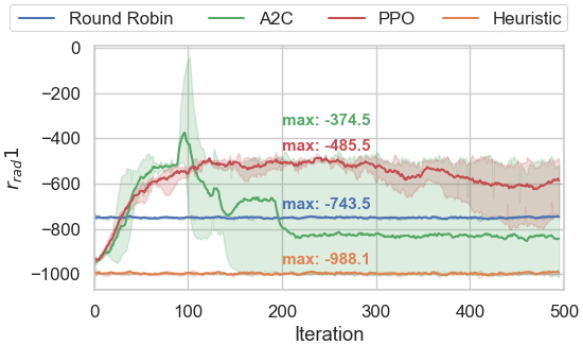


Fig. 4: Average total radar reward r_{rad} per episode for each iteration of the training process.

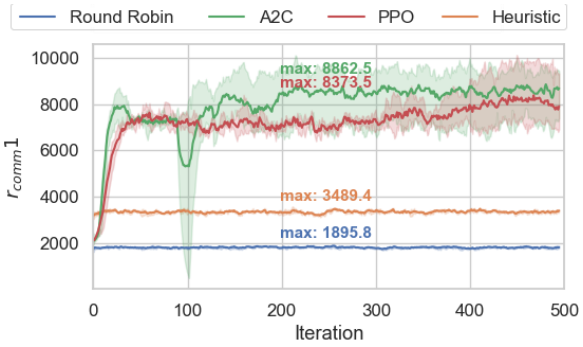


Fig. 5: Average total communication reward r_{comm} per episode for each iteration of the training process.

for radar detection, the heuristic agent performs less well in terms of radar reward r_{rad} (see Figure 4). When the ego agent is trained with either A2C or PPO, it achieves a steady increase in reward in the training process, and outperforms both the round robin and heuristic agents by a significant margin. A higher maximum reward is achieved by the PPO agent, which also demonstrates better stability than the A2C agent. The simultaneous increase in communication reward r_{comm} achieved by both reinforcement learning agents (Figure 5) indicates that they learned to transmit more information that is useful to the other agents.

B. Effect of varying the number of vehicles

We devise a set of experiments with the objective of determining how robust the learned solution of a reinforcement learning agent is to the number of vehicles in the environment. Firstly, we train PPO agents separately in environments where the number of vehicles N_{train} are set to 8, 12, 16, 20 and 24 respectively. Subsequently, we test the trained PPO agents in an environment with a different number of other vehicles N_{test}. For example, the PPO agent trained in an environment with 8 vehicles is tested in environments with 12, 16, 20 and 24 vehicles respectively. Since the policy network for PPO accepts an input size corresponding to N_{train}, we handle the inputs at test time as follows: when the number of inputs to the policy network is larger than the number of vehicles in the environment (i.e. $N_{train} > N_{test}$), the observation features

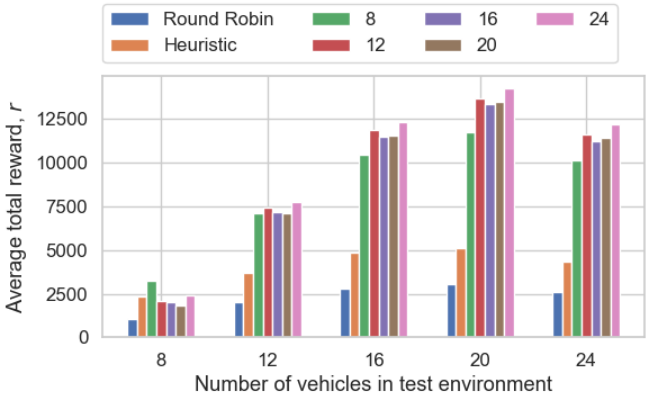


Fig. 6: Total reward per episode obtained by agents trained with different numbers of vehicles in the environment. The x-axis shows the number of vehicles in the test environment.

for non-existent vehicles are set to zero. Conversely, when $N_{train} < N_{test}$, the agent only observes only the N_{train} nearest agents by Euclidean distance.

The results are presented in Figure 6. We note that the trained PPO agents perform better when $N_{test} = N_{train}$. Furthermore, they perform similarly when N_{test} is close to N_{train}, even without prior training in the test environment, indicating that the learned solution has a degree of generality. This generality is attested by the superior performance of the PPO agents to the heuristic agent, despite operating in a different environment. Interestingly, the agent trained in the 24-vehicle environment performed the best when $N_{test} \geq 12$. This may be attributed to a wider distribution of observational data in the training phase due to the presence of more vehicles. The agent trained with $N_{train} = 12$ also performs well when $N_{test} > 12$. This may be attributed to the nearest vehicles providing more useful information for decision making. Future work may investigate these effects further.

VII. Conclusion

We proposed a problem framework for joint radar communication (JRC) by an autonomous vehicle that is subjected to an environment where neighboring vehicles operate radar detection periodically, as is typical in contemporary protocols. We defined this as a Markov Decision Process (MDP) where the ego vehicle simultaneously decides (i) how to divide time between radar and communication functions, (ii) the content of data to transmit, and (iii) the direction in which the chosen data is transmitted, which effectively divides the available frequency band in the spatial dimension. The overall objectives were to maximize the throughput of useful data, minimize the age of information (AoI) of data transmitted, and conduct radar frequently enough for a sufficiently accurate measurement of the positions and speeds of surrounding objects. Experiments showed that solutions computed using deep reinforcement learning achieved significantly better results than non-learning algorithms, with the advantage of requiring minimal a priori knowledge of the environment.

References

[1] R. Hult, F. E. Sancar, M. Jalalmaab, A. Vijayan, A. Severinson, M. D. Vaio, P. Falcone, B. Fidan, and S. Santini, "Design and experimental validation of a cooperative driving control architecture for the grand cooperative driving challenge 2016," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1290–1301, Apr. 2018.

[2] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[3] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3834–3862, Jun. 2020.

[4] N. C. Luong, X. Lu, D. T. Hoang, D. Niyato, and D. I. Kim, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3834–3862, Jun. 2020.

[5] Q. H. Nguyen, T. H. Dinh, C. L. Nguyen, and D. Niyato, "irdrc: An intelligent real-time dual-functional radar-communication system for automotive vehicles," *IEEE Wireless Commun. Lett.*, Dec. 2020.

[6] J. Lee, D. Niyato, Y. L. Guan, and D. I. Kim, "Learning to schedule joint radar-communication requests for optimal information freshness," in *Proc. IEEE IV*, 2021.

[7] J. Lee, T. D. Niyato, Y. L. Guan, and D. I. Kim, "Learning to schedule joint radar-communication with deep multi-agent reinforcement learning," *IEEE Trans. Veh. Tech.*, 2021.

[8] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5g networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581–1592, Mar. 2020.

[9] S. Alland, W. Stark, M. Ali, and M. Hegde, "Interference in automotive radar systems: Characteristics, mitigation techniques, and current and future research," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 45–59, Sept. 2019.

[10] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. IEEE SECON*, Jun. 2011, pp. 350–358.

[11] F. Kunz, D. Nuss, J. Wiest, H. Deusch, S. Reuter, F. Gritschneder, A. Scheel, M. Stübler, M. Bach, P. Hatzelmann, C. Wild, and K. Dietmayer, "Autonomous driving at ulm university: A modular, robust, and sensor-independent fusion approach," in *Proc. IEEE IV*, 2015, pp. 666–673.

[12] J. Steinbaeck, C. Steger, G. Holweg, and N. Druml, "Next generation radar sensors in automotive sensor fusion systems," in *Sensor Data Fusion*, 2017, pp. 1–6.

[13] C. Habib, A. Makhoul, R. Darazi, and R. Couturier, "Real-time sampling rate adaptation based on continuous risk level evaluation in wireless body sensor networks," in *Proc. IEEE WiMob*, 2017, pp. 1–8.

[14] A. Pal and K. Kant, "On the feasibility of distributed sampling rate adaptation in heterogeneous and collaborative wireless sensor networks," in *Proc. ICCCN*, 2016, pp. 1–9.

[15] B. Or, B. Z. Bobrovsky, and I. Klein, "Kalman filtering with adaptive step size using a covariance-based criterion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[16] D. Giouroukis, A. Dadiani, J. Traub, S. Zeuch, and V. Markl, "A survey of adaptive sampling and filtering algorithms for the internet of things," in *ACM DEBS*, Jul. 2020, p. 27–38.

[17] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for v2v communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[18] L. Li and X. M. Chen, "Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey," *Transp. Res. Part C: Emerg. Technol.*, vol. 76, pp. 170–188, Mar. 2017.

[19] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. ICML*, vol. 48, Jun. 2016, pp. 1928–1937.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv 1707.06347*, 2017.