Trends in Deep Reinforcement Learning for Distributed Coordination and Cooperation with Homogeneous Multi-Agents

Joongheon Kim
School of Electrical Engineering, Korea University, Seoul, Republic of Korea
E-mail: joongheon@korea.ac.kr

***Abstract*—Deep Reinforcement Learning (DRL) has gained traction as a potent method for improving the real-time sequential decision-making abilities of autonomous vehicles, thereby stimulating a wealth of research in this area. However, in complex environments where autonomous vehicles must interact with various agents like pedestrians and other vehicles, the necessity for a multi-agent approach becomes evident. While agent communication is fundamental for efficient decision-making, its integration into DRL can be challenging. To address this problem, this paper explores inter-agent information sharing by a Communication Network (CommNet) that allows agents to efficiently make collective decisions only based on observed information. In this paper, the benefits of employing CommNet in a variety of real-world applications are evidenced, particularly where agents of autonomous vehicles must engage in information exchange in dynamic environments. Overall, the importance and potential benefits of the proposed strategy for autonomous vehicles are underscored to bolster their decision-making prowess.**

I. Introduction

Deep reinforcement learning (DRL) offers a considerable benefit in that it facilitates sequential decision-making grounded in its learned policies, obviating the necessity to explore every conceivable situation to identify the most effective solution [1]–[6]. Thus, in circumstances where optimization reliant on dynamic programming is applied due to vast search spaces, or in settings punctuated by unexpected uncertainties that cannot be mathematically modeled, DRL is capable of delivering immediate and adaptable reactions to the given states [7]. However, it is common for conventional DRL algorithms to struggle to achieve optimal policy cooperatively when several agents are involved, which results in subpar performance or even reward convergence failure [8]–[10]. The particular reason is that multiple agents can have an impact on one another's performances in determining the best course of action toward a shared objective. In addition, all agents compete with each other in order to find only their own optimal policy without cooperation since traditional DRL algorithms consider a single-agent environment. These factors make multi-agent deep reinforcement learning (MADRL) environment non-stationary. Recent research has concentrated on developing meaningful MADRL algorithms that allow agents to communicate and coordinate with one another in order to address this difficulty.

One promising strategy to find an optimal policy in MADRL is using a communication network (CommNet) [11], which is a type of neural network enabling agents to communicate and share information throughout the learning process. By allowing agents to interact with each other, CommNet helps them learn to coordinate their actions better without a central controller. This can stimulate faster reward convergence and more outstanding performance than previous techniques. Multifaceted applications to autonomous vehicles, such as cooperative charging scheduling [12], [13], surveillance [14], and traffic signal control [15], have demonstrated the utility of CommNet.

CommNet-based agents engage in information sharing by transmitting encoded hidden variables rather than directly relaying observed information, similar to the approach in Federated Learning (FL) [16]–[18]. This method offers enhanced security during the communication process. Therefore, even in sensitive autonomous driving scenarios, like ambulances operating within hospital settings where secure information sharing is crucial, the application of CommNet allows for the secure exchange of sensitive information among agents without infringing on data privacy.

This paper aims to present a comprehensive explanation of CommNet in MADRL applications. Findings from several previous publications demonstrate how well CommNet helps multiple agents learn to communicate and coordinate their actions appropriately. These discoveries can aid in creating more sophisticated and efficient MADRL algorithms and make them easier to adopt in real world applications that call for collaboration among multiple agents. Overall, this work advances knowledge of the effect of CommNet in MADRL and offers insights into its use in versatile scenarios. In addition, this paper hopes to encourage additional research in this field by paving the way for developing more practical and progressive MADRL algorithms.

II. Cooperative MADRL using Inter-Agent Communications

CommNet allows multiple agents to communicate with each other through a piece of shared information as illustrated in Fig. 1. Each j-th agent has its own neural network that takes in the agent's observations o_j, and ground truth state s_j. This information is entered into the first hidden layer as an input. Hidden layers consist of a number of neurons, and they are fully connected with next hidden layers [19]. Every i-th

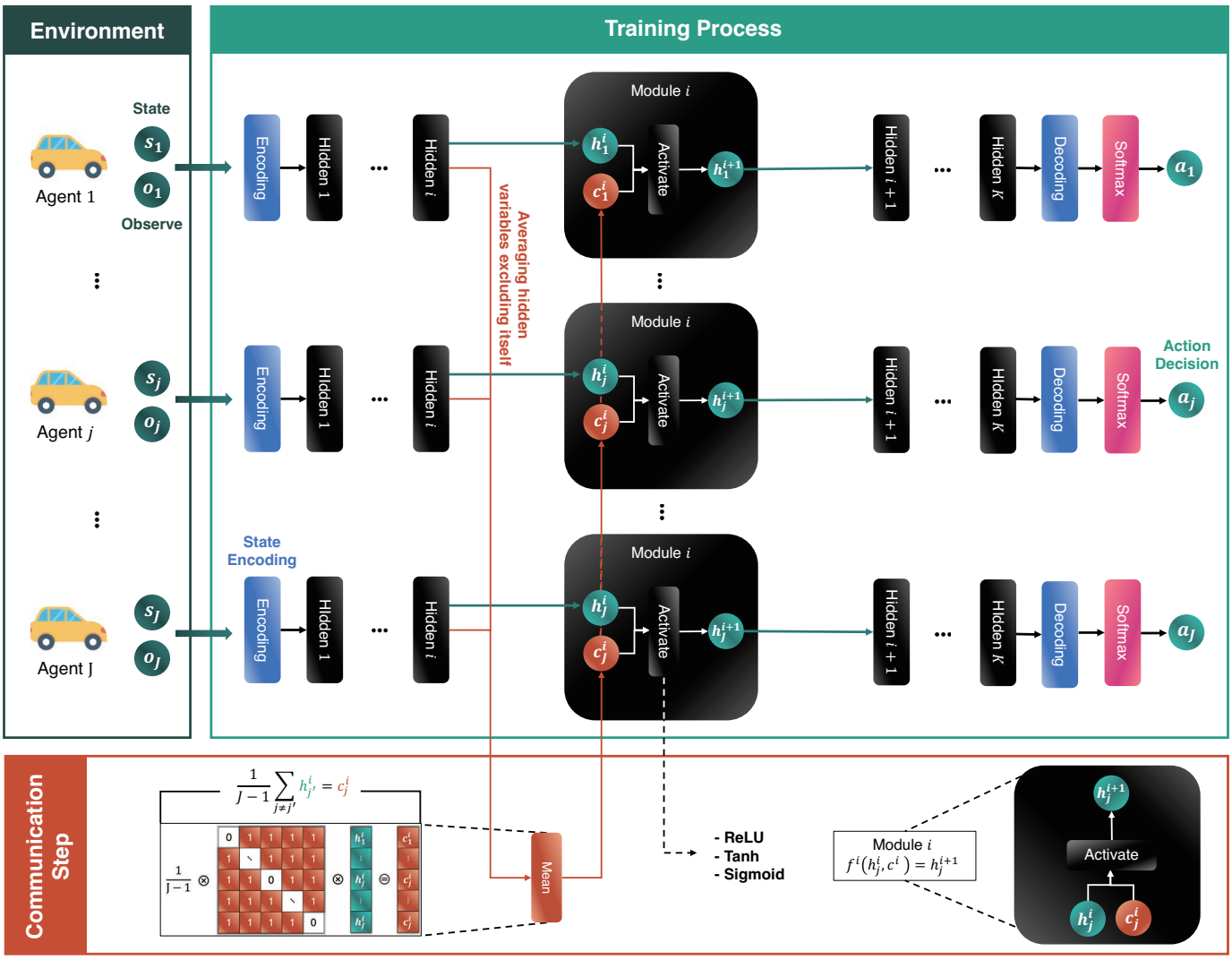

Fig. 1: Structure of inter-agent communication for collaborative multi-agent interaction leveraging CommNet.

hidden layer outputs a hidden state h_j^i which corresponds to the input of the i+1-th hidden layer. However, in CommNet, this hidden state is then combined with the hidden states of other agents through the communication step before being fed into the next hidden layer. In the communication step, a single j-th agent gets the communication variable c_j^i which is made by averaging the hidden states of other agents except for itself as depicted in the below part of Fig. 1. Using communication variable c_j^i, the output of the communication step, as an additional input allows the agent to update its policy based on the information received from other agents. As a result, agents can learn to effectively coordinate their policies through communication without centralized control or explicit coordination rules. Afterward, the i-th module $f^i(\cdot)$ transforms a concatenation of the hidden variable and communication variable $\left[h_j^i, c_j^i\right]$ into the next layer's hidden variable h_j^{i+1} using activation function such as ReLU, tangent-hyperbolic, or Sigmoid function. This communication process is repeatedly performed until it has reached the last hidden layer. Finally, agent can get the probability of all possible actions by conducting the softmax function to the decoded output of the last hidden layer. By doing so, each agent can make sequential decision-making considering not only its information but also other agents' information. In other words, agents take dependent actions cooperatively, albeit with getting independent experience in environments. In a nutshell, CommNet is a powerful and flexible approach to MADRL that enables agents to learn to communicate and coordinate their actions effectively to achieve a common goal.

III. Applications

This section presents the applications of CommNet, as summarized in Table I. These studies exemplify the application of real-time inter-agent information sharing in achieving objectives, specifically in the context of autonomous vehicle-type agents. To the best of the author's knowledge, they stand out for their unique employment of the CommNet algorithm in real-world autonomous driving settings, emphasizing the

TABLE I: Applications of Information Exchange among Different Agents in Autonomous Vehicles

	Shin *et al.* [12]	Jung *et al.* [13]	Yun *et al.* [14]	Park *et al.* [23]
Objective	Charging Scheduling	Charging Scheduling	Surveillance	Cellular Access
Vehicle Type	Electric Vehicles	UAVs	UAVs	UAVs
Agent	Charging Stations	Charging Towers	UAVs	UAVs
Action	Purchasing Energy	Purchasing Energy	2D Trajectory, Coverage	3D Trajectory
Reward	Payment Amount, Overcharging	Payment Amount, Overcharging	Support Rate, Resolution	Support Rate, QoS

novelty and uniqueness of such applications in these scenarios. In addition, they adopt single-agent DRL algorithms as comparators to evaluate the proposed method's performance. While there are other prominent MADRL algorithms, such as Value-Decomposition Networks (VDN) [20], QMIX [21], and Counterfactual Multi-Agent (COMA) [22], they typically base their state-value function $V(\mathcal{S})$ on a global state $\mathcal{S}$ for learning policies. This approach becomes impractical for real-time services due to the difficulty of aggregating extensively scaled global states. In contrast, the proposed approach permits agents to learn policies solely from their observed information, making it more suitable for realistic scenarios. Hence, this study compares with single-agent DRL algorithms rather than other MADRL algorithms to validate the power of policy learning via inter-agent communication, considering realistic scenarios.

A. Charging Station Scheduling for Electric Vehicles

Scenario Overview. In the Industry 4.0 Revolution era, the electric vehicle (EV) is considered as one of major players for autonomous vehicles, because it is easier to control their motors' rotation without delay. In light of these trends, optimizing the energy use and operation costs of EV charging stations (EVCSs) based on information regarding supplier-consumer patterns for cost-effectiveness is critical. Although previous researchers have studied EVCS operation optimization, they suggested a centralized approach. However, it is challenging to come up with a real-time centralized solution for processing massive dynamic time-varying data. To solve the given problem, learning-based and distributed approaches are widely utilized. Thus, Shin *et al.* propose a decentralized MADRL-based optimization to manage huge data while considering the use of energy storage system (ESS) and photovoltaic (PV) power production for EVCSs [12]. Here, a single private enterprise manages multiple EVCSs equipped with renewable energy resources. Each EVCS can provide energy to EVs with its own ESS and other EVCSs' ESS while charging energy with its mounted PV charger. At this time, EVCSs share only the remains of surplus energy after meeting their net demand. By doing so, every EVCS can meet net demand reducing overall operating costs by managing surplus energy. Shin *et al.* utilize CommNet for all EVCS agents' training policies to manage the charging/discharging of the energy stored in the ESSs cooperatively. Every EVCS agent jointly needs to minimize purchasing energy from the enterprise while observing the energy state and prices.

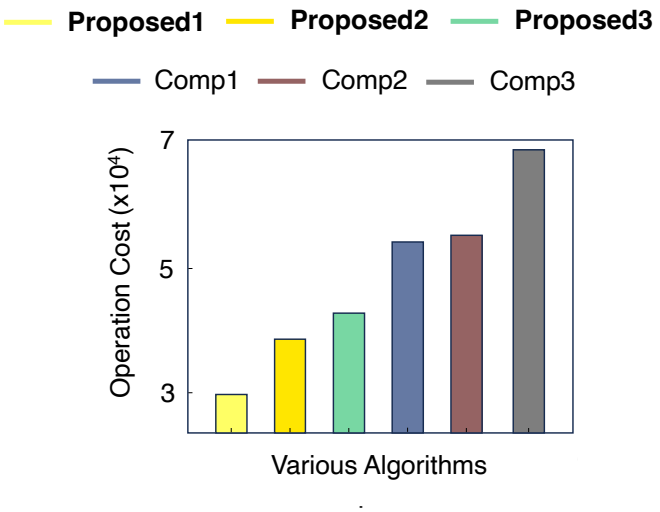

Fig. 2: Comparison of the operation cost in each training method.

Performance Evaluation. Shin *et al.* conducts a controlled experiment with independent variable σ that is the reward coefficient influencing the significance of the amount of energy charged into the ESS to fulfill the overall net demand. It means that a larger value of σ implies a greater necessity for residual energy, which in turn results in a higher operational cost for the EVCS. Here, three EVCS provide electric vehicle charging services. The proposed inter-agent information sharing algorithm is employed to learn policies by setting the value of σ to 1, 2, and 3, respectively. These strategies are explicitly labeled as Proposed 1, Proposed 2, and Proposed 3 as illustrated in Fig. 2. They compare the performance of the proposed mutual information sharing between multiple agents with the conventional DRL algorithms with $\sigma = 1$ including Deep Q-Network (DQN) [24] and Proximal Policy Optimization (PPO) [25], which corresponds to Comp1 and Comp2, respectively. In addition, there is a random algorithm where each EVCS agent takes action randomly without observing states, such as the cost of electricity and residual energy in ESS. As a result, the ESS operation cost of the EVCS agents trained by DQN or PPO is almost double to that of the proposed CommNet-based management system, when $\sigma = 1$. In addition, it is noteworthy that the proposed CommNet-based ESS management outperforms the policy learning in terms of operating costs, even though σ is two or three times bigger than other benchmarks. It can be understood that by sharing their current state among different EVCS agents, it

minimizes the amount of energy purchased by cooperatively sharing energy from EVCS with sufficient remaining energy to those with limited energy at present.

B. Charging Scheduling for UAV Networks

Scenario Overview. The energy management framework also can be adopted in unmanned aerial vehicle (UAV) networks. UAV is one of the sixth generation (6G) core technologies because it can provide flexible network service [26]. In particular, efficient energy management is essential since they have an insufficient battery capacity, where the operation time is limited to few hours. In addition, it is burdensome to control ESS management centrally while managing sensing data gathered by multiple UAVs. Thus, Jung *et al.* propose cloud-assisted charging scheduling via CommNet [13]. Here, multiple distributed charging towers serve plug-and-play charging during run-time operations. Every charging tower trains its policy for efficiently providing energy to UAVs with intelligent energy sharing collaboratively. As in [12], all charging tower agents have the common goal of minimizing payment amount and overcharging.

Performance Evaluation. In the considered scenario, there are four charging towers that provide charging services to UAVs, and each charging tower can share energy with one another. Jung *et al.* execute experiments involving the proposed inter-agent communication strategy, contrasting it with DQN and a MADRL-disable rule-based random action. These comparative elements are referred to as Comp1 and Comp2, respectively. The total purchased energy in each training progress is depicted in Fig. 3. It can be observed that the charging towers in the proposed algorithm purchase the least amount of energy. It implies that the proposed algorithm efficiently shares energies across charging towers to minimize operating expenses while maximizing shared energy. This outcome arises from the differences in information sharing among the charging towers. By exchanging information among the charging towers, they make decisions in an optimal manner, aiming to efficiently supply energy to the entire UAV network, rather than focusing on individually optimal choices.

C. Surveillance for UAV Networks

Scenario Overview. One of the many-sided UAV applications is flexible mobile surveillance [26], where UAVs provide on-demand surveillance by dynamically updating the locations. Furthermore, they can access extreme environments whereas physical limitations exist. However, there are various uncertainties that UAVs face, such as hardware damage regarding accidents with obstacles or insufficient battery. Furthermore, when multiple UAVs collide with each other, overlapping their surveillance coverage leads to service inefficiency. Thus, providing reliable autonomous surveillance services is essentially required. Yun *et al.* provide MADRL-based multi-UAV control using CommNet with one leader UAV and the others as non-leader UAV [14]. Here, the leader UAV decides its action by reflecting the information of all the others, while non-leader UAVs take an optimal action based only on their own

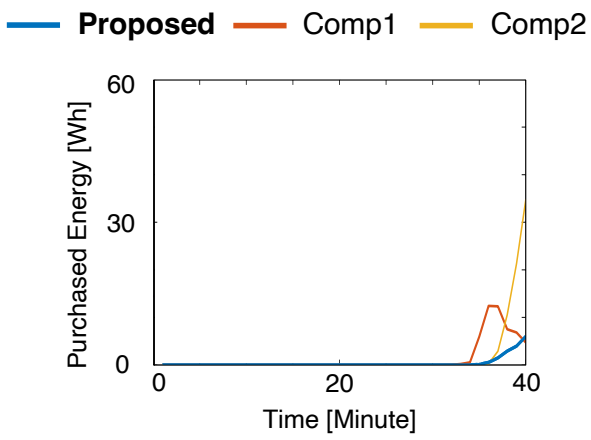

Fig. 3: Comparison of purchased energy in each training method.

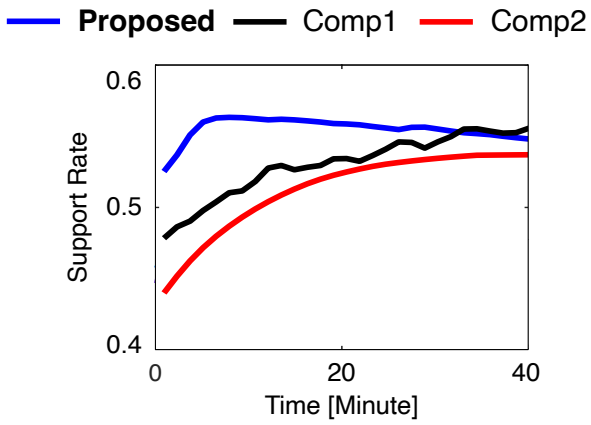

Fig. 4: Comparison of support rate in each training method.

information, like a single agent DRL. As noted in Table I, multiple UAVs in the system jointly move 2D trajectories or control video resolution (*i.e.*, coverage radius), aiming to monitor a large number of users with high video resolution.

Performance Evaluation. In the considered scenario, there are four UAVs, differentiated by the number of UAVs learning their CommNet policies and classical deep neural network (DNN). Yun *et al.* comprises only one CommNet-based leader UAV and three DNN-based non-leader UAVs. The remaining two benchmarks, Comp1 and Comp2, consist only of CommNet-based UAVs and DNN-based UAVs respectively. These four UAVs have the capability to move in a cardinal direction or perform quality control on surveillance image at each time step in $2,400\,m \times 2,400\,m$ 2D grid map. Additionally, there are three non-DRL UAVs to test adaptability to environmental uncertainties that cannot be mathematically modeled. These UAVs provide surveillance services from fixed locations and experience random failures with a certain probability. Fig. 4 shows the support rate of UAVs with trained policies. The suggested surveillance scheme exhibits the high-

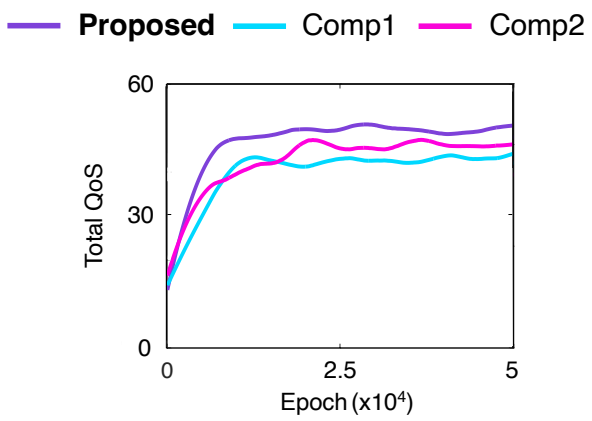

Fig. 5: Comparison of QoS in each training method at POMDP environment.

est level of support rate throughout nearly all episodes, and has the same support rate as Comp1 at the end of the episode. This result indicates that the proposed scheme consistently demonstrates the strongest surveillance performance. However, Comp2, which does not utilize CommNet (*i.e.,* no inter-agent communications) shows the lowest surveillance performance in all episodes. Therefore, when comparing the performance difference between the Proposed/Comp1, which incorporates inter-agent information sharing (*i.e.,* locations of users and other UAVs), and Comp2, which does not, it can be confirmed that the CommNet-based strategy enables a more cooperative and effective response to uncertainties such as UAV failures. This strategy allows for the successful achievement of shared objectives.

D. Mobile Cellular Access for UAV Networks

Scenario Overview. UAVs can also provide on-demand wireless communication services anywhere at a low cost since there is no need to construct an extra ground base station. Here, an UAV serving as the base station is referred to as UAV-BS. Park *et al.* propose CommNet-based multi-UAV-BS control for reliable mobile cellular access [23]. This paper demonstrates the effectiveness of inter-agent communications in partially observable Markov decision process (POMDP) and fully observable Markov decision process (FOMDP) [27]. Assuming the FOMDP means that the agent can observe entire environmental information, where this assumption is not realistic due to the fact that each agent has a limited observation because of physical limitation. Therefore, it is vital to adopt CommNet to multi-agent cooperation in real world, formulated by POMDP. UAV-BS cooperatively tries to maximize the support rate and quality of service (QoS) of ground user equipments (UEs), as in Table I.

Performance Evaluation. The configuration of multiple UAV-BSs is similar to work in [14], but in this case, they move in a three-dimensional (3D) trajectory in $6,000\,m \times 6,000\,m \times 2,500\,m$ grid map, and the coverage radius is determined by the altitude of them. In addition, every UE requests different data rates for services such as video streaming, online gaming, or web surfing. For the received QoS of ground UEs, theoretical data rates are calculated, taking into account factors such as the distance between the UAV-BS and UE, and interference from other UAV-BSs. These theoretical data rates are then matched with the Modulation and Coding Scheme (MCS) table of IEEE 802.11ad, and a quality function $f_v(\cdot)$ in [28] is applied to calculate the actual data rate. Fig. 5 shows the QoS received by terrestrial UEs over policy training epochs. Benchmarks are identical to the aforementioned autonomous surveillance system [14]. Benchmarks utilizing CommNet (Proposed and Comp1) show a faster reward convergence speed and more stable learning performance than Comp2 which only consists of the DNN-based UAV-BSs. In addition, the proposed mobile access network is superior among all benchmarks in terms of service quality. However, it is noteworthy that Comp1, which only has CommNet-based UAV-BSs, has the lowest QoS value. The improved performance can be attributed to the effective coordination between CommNet- and DNN-based policies, which leverages the abundant experience accumulated by DNN-based agents within a specific area in POMDP, particularly in large-scale maps. Nevertheless, the proposed inter-agent information sharing strategy indicates the ability to establish a more cooperative and effective mobile access network with the highest QoS value compared to Comp2, which does not engage in information sharing.

IV. Concluding Remarks

This paper presents the advantages of employing inter-agent information sharing in the context of MADRL for autonomous vehicles. The results derived from various experiments showcase that the integration of decentralized inter-agent communication can substantially improve the efficiency of multiple agents operating within complex and dynamic environments, yielding enhanced learning speeds and superior performance. Moreover, the preliminary findings from potential real-world applications of the proposed strategy provide promising prospects for future research and development. Notable examples of these applications include the collaboration of multiple robots in a smart factory setting and the cooperation of multiple agents for providing real-time services in caching scenarios for saving computing resources in communication environments. The authors of this paper believes that that the insights generated through this research can pave the way for more sophisticated autonomous vehicles, capable of operating safely and effectively in real-world conditions.

Acknowledgement

This research was funded by the National Research Foundation of Korea (NRF-Korea), Basic Research Laboratory (BRL) (2021R1A4A1030775). The author thanks to Mr. Chanyoung Park for his contribution on research initiation, during his graduate study under the guidance of Prof. Joongheon Kim.

REFERENCES

[1] M. Shin and J. Kim, "Randomized adversarial imitation learning for autonomous driving," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 4590–4596.

[2] D. Kwon, J. Jeon, S. Park, J. Kim, and S. Cho, "Multiagent DDPG-based deep learning for smart ocean federated learning IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9895–9903, October 2020.

[3] Y. J. Mo, J. Kim, J.-K. Kim, A. Mohaisen, and W. Lee, "Performance of deep learning computation with tensorflow software library in GPU-capable multi-core computing platforms," in *Proc. IEEE International Conference on Ubiquitous and Future Networks (ICUFN)*, Milan, Italy, July 2017, pp. 240–242.

[4] Y. Kwak, W. J. Yun, S. Jung, and J. Kim, "Quantum neural networks: Concepts, applications, and challenges," in *Proc. IEEE International Conference on Ubiquitous and Future Networks (ICUFN)*, Jeju, Korea, August 2021, pp. 413–416.

[5] Y. Kwak, W. J. Yun, S. Jung, J.-K. Kim, and J. Kim, "Introduction to quantum reinforcement learning: Theory and PennyLane-based implementation," in *Proc. IEEE International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea, October 2021, pp. 416–420.

[6] Y. Kwak, W. J. Yun, J. P. Kim, H. Cho, J. Park, M. Choi, S. Jung, and J. Kim, "Quantum distributed deep learning architectures: Models, discussions, and applications," *ICT Express*, vol. 9, no. 3, pp. 486–491, September 2023.

[7] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L.-C. Wang, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 44–52, June 2019.

[8] X. Tan, L. Zhou, H. Wang, Y. Sun, H. Zhao, B.-C. Seet, J. Wei, and V. C. Leung, "Cooperative multi-agent reinforcement-learning-based distributed dynamic spectrum access in cognitive radio networks," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19 477–19 488, October 2022.

[9] C. Park, W. J. Yun, J. P. Kim, T. K. Rodrigues, S. Park, S. Jung, and J. Kim, "Quantum multi-agent actor-critic networks for cooperative mobile access in multi-UAV systems," *IEEE Internet of Things Journal*, pp. 1–1, 2023 (Early Access).

[10] N.-N. Dao, D.-N. Vu, W. Na, J. Kim, and S. Cho, "SGCO: Stabilized green crosshaul orchestration for dense IoT offloading services," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2538–2548, September 2018.

[11] S. Sukhbaatar, R. Fergus *et al.*, "Learning multiagent communication with backpropagation," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, Barcelona, Spain, December 2016, pp. 2244–2252.

[12] M. Shin, D.-H. Choi, and J. Kim, "Cooperative management for PV/ESS-enabled electric vehicle charging stations: A multiagent deep reinforcement learning approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3493–3503, May 2020.

[13] S. Jung, W. J. Yun, M. Shin, J. Kim, and J.-H. Kim, "Orchestrated scheduling and multi-agent deep reinforcement learning for cloud-assisted multi-UAV charging systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 5362–5377, June 2021.

[14] W. J. Yun, S. Park, J. Kim, M. Shin, S. Jung, D. A. Mohaisen, and J.-H. Kim, "Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7086–7096, October 2022.

[15] J. Gao, X. Shi, and J. James, "Attn-CommNet: Coordinated traffic lights control on large-scale network level," in *Proc. of IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Washington, DC, USA, November 2021, pp. 289–293.

[16] H. Baek, W. J. Yun, Y. Kwak, S. Jung, M. Ji, M. Bennis, J. Park, and J. Kim, "Joint superposition coding and training for federated learning over multi-width neural networks," in *Proc. of IEEE Conference on Computer Communications (IEEE INFOCOM)*, London, United Kingdom, May 2022, pp. 1729–1738.

[17] M. Shin, C. Hwang, J. Kim, J. Park, M. Bennis, and S. Kim, "XOR mixup: Privacy-preserving data augmentation for one-shot federated learning," *CoRR*, vol. abs/2006.05148, 2020. [Online]. Available: https://arxiv.org/abs/2006.05148

[18] S. Park, S. Jung, H. Lee, J. Kim, and J.-H. Kim, "Large-scale water quality prediction using federated sensing and learning: A case study with real-world sensing big-data," *Sensors*, vol. 21, no. 4, February 2021.

[19] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, March 1996.

[20] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. of International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Stockholm, Sweden, July 2018, pp. 2085–2087.

[21] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7234–7284, January 2020.

[22] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. of the AAAI conference on artificial intelligence*, vol. 32, no. 1, New Orleans, LA, USA, February 2018, pp. 2974–2982.

[23] C. Park, H. Lee, W. J. Yun, S. Jung, C. Cordeiro, and J. Kim, "Cooperative multi-agent deep reinforcement learning for reliable and energy-efficient mobile access via multi-UAV control," *arXiv preprint arXiv:2210.00945*, 2022.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

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

[26] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, September 2019.

[27] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson, "Deep variational reinforcement learning for POMDPs," in *Proc. of International Conference on Machine Learning (ICML)*, Stockholm, Sweden, July 2018, pp. 2117–2126.

[28] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Non-convex optimization and rate control for multi-class services in the Internet," *IEEE/ACM Transactions on Networking*, vol. 13, no. 4, pp. 827–840, Aug. 2005.