Research Article

A Value Factorization Method for MARL Based on Correlation between Individuals

Liqin Xiong, Lei Cao, Xiliang Chen, Jun Lai, and Xijian Luo

Command and Control Engineering College, PLA Army Engineering University, Nanjing, Jiangsu 210007, China

Correspondence should be addressed to Lei Cao; caolei_nj2022@126.com

Received 9 May 2022; Accepted 13 July 2022; Published 9 August 2022

Copyright © 2022 Liqin Xiong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Value factorization is a popular method for cooperative multi-agent deep reinforcement learning, which effectively solves explosion of state-action spatial dimension and partial observability problems. However, most existing algorithms only consider the impact of individuals rather than correlation between individuals, which leads to poor coordination between agents in complex environments. In order to resolve this problem, this paper proposes a multi-agent deep reinforcement learning value factorization method based on correlation between individuals, CI-VF, which promotes coordination between agents effectively. Firstly, the individual value function vectors are obtained according to the output of individual networks in each round. Secondly, a Spearman correlation coefficient matrix can be calculated by the vectors to measure the correlation degree of agents, and the joint correlation coefficient can be obtained to optimize joint value function. Next, we use optimized joint value function to train individual networks. Experimental results show that our method outperforms QMIX and other baselines in various scenarios under the StarCraft Multi-Agent Challenge environment.

1. Introduction

Cooperative multi-agent tasks in real world usually involve cooperation of various number or types agents, such as resource allocation of wireless sensor network [1] and vehicle scheduling of intelligent transportation system [2], which poses great challenge to realize multi-agent collaboration that traditional intelligent algorithms [3–5] are no longer suitable for. One solution to the problem is value factorization (VF), a classic multi-agent deep reinforcement learning (MADRL) method, which factorizes the globally shared joint value function into individual value functions based on local observation to urge agents to learn an optimal joint action policy.

VF’s training mode is centralized training with decentralized execution (CTDE) [6, 7], which optimizes individual networks and greedily selects actions according to joint value function and individual value functions, respectively. Compared with the completely centralized training method [8], the VF method effectively alleviates explosion of state-action spatial dimension and partial observability problems. While compared with independent Q-learning [9], it can better deal with the problems of non-stationarity of environment and multi-agent collaboration. Therefore, the VF method is worthy of study extremely.

The VF method was first proposed in an additive way [6], where joint value function is represented by the simple sum of individual value functions. Later, Rashid et al. [7] proposed an improved algorithm, QMIX, which represented joint value function as a monotonic non-linear combination of individual value functions through monotonic mixing network. QMIX is a classic and efficient VF method and performs well in StarCraft Multi-Agent Challenge (SMAC) environment. Besides, there are many optimized VF algorithms. For example, weighted QMIX [10] is proposed to solve the problem that QMIX falls into suboptimal policy due to monotonic constraint. QTRAN [11] is proposed to solve the problem that QMIX has a strong structural constraint. Also, Qatten [12] is proposed to solve the problem of multi-agent credit assignment by attention mechanism.

However, the existing VF algorithms only consider individual impact on the whole system and ignore the
correlation between individuals. This correlation may be critical sometimes, especially in cooperative environment. For example, an attacking agent with low health needs the protection of defensive agents; otherwise, it is extremely vulnerable to be killed by enemies. Even though this correlation is already reflected in the constraint of joint value function and individual value function, it is fuzzy and insufficient.

This paper proposes a MADRL method, CI-VF, to study correlation between individuals, which utilizes Spearman correlation coefficient of individual value functions to quantify correlation between individuals and takes the advantage of the correlation to guide network training. CI-VF describes the interaction between individuals mainly through the relative variation trend of individual value function in the training process, which is based on the common sense that if there are two closely related agents in a cooperative environment, their individual value functions should be relevant to some extent. By enhancing this correlation, each agent will select actions that are more beneficial to itself and its partners to learn the optimal joint policy. We evaluate CI-VF on the challenging StarCraft II micromanagement mission and show that CI-VF outperforms QMIX and other baselines in various scenarios.

The rest of this paper is organized as follows. Section 2 briefly introduces the relevant work and Section 3 introduces basic theories of VF. Section 4 proposes CI-VF, a VF method based on correlation between individuals. Section 5 carries out comparative experimental analysis on CI-VF and baselines. Section 6 summarizes and prospects the future work of our research.

2. Related Work

In recent years, significant breakthroughs have been made in deep reinforcement learning (DRL), and MADRL has gradually become a research hotspot. MADRL combines multi-agent system with DRL, aiming to utilize the powerful decision-making and presentation capabilities of DRL to solve multi-agent collaboration problems, such as UAV (unmanned aerial vehicle) path optimization [13], network resource allocation [14], and so on. VF is one of the MADRL methods which can solve the multi-agent collaboration problem in cooperative environment. It was first proposed by representing joint value function as the sum of individual value functions, which achieved initial success in solving the problem of multi-agent reinforcement learning [6]. Since then, researches began to pay attention to the field of VF.

QMIX [7] is one of the most classical and effective VF algorithms, which uses mixing network with global information to decompose joint value function into a complex non-linear combination of individual value functions, thus improving the representational ability, application scope, and stability of the algorithm. Later on, realizing that QMIX was easy to be trapped in suboptimal policy due to monotonic constraints, Rashid et al. [10] proposed a weighted QMIX algorithm, WQMIX, which introduced weights to weight the square error of joint value function when updating networks, so as to obtain the optimal joint policy. Xu et al. [15] proposed a MMD (maximum mean discrepancy) mixing network with global status information, which combined distributed reinforcement learning and VF to adapt to the environment with high randomness.

However, although VDN (value decomposition network), QMIX, and WQMIX performed well in the experiment, they lacked theoretical basis. Therefore, Yang et al. [12] proposed a cooperative multi-agent reinforcement learning framework, Qatten, which derived the factorization formula of joint value function theoretically and adopted the mixing network based on multi-head attention to approximate it, enhancing the theoretical and rationality of the VF method. Zhao et al. [16] also proposed DQMIX, a novel value-based multi-agent reinforcement learning (MARL) method, from a distributional perspective, which employed a distribution mixing network to integrate individual value distributions into the global value distribution and further proved that DQMIX can meet the distributional-individual-global-max (DIGM) principle with respect to the expectation of distribution. The experimental results showed DQMIX coped well with the randomness of long-term returns.

VDN, QMIX, and WQMIX have the strong structural constraints due to their joint value functions obtained from additive or monotonic networks. In order to relax this constraint, Son et al. [11] proposed a general VF method, QTRAN, which transformed joint value function into decomposable function with same optimal actions, thereby improving the scalability of the algorithm. However, although QTRAN has good theoretical guarantee, its empirical performance is not good. In order to bridge the gap, Son et al. [17] proposed an improved version of QTRAN, QTRAN++, which used joint action value obtained from multi-head monotonic mixing network to imitate the real joint action value obtained from semi-monotonic mixing network to optimized the loss function, so as to stabilize the training process. Later, Wang et al. [18] proposed a MADRL algorithm based on duplex dueling network architecture, QPLEX. It used duplex dueling structure to transform individual-global-max (IGM) principle into consistency constraint of advantage functions and introduced an extensible multi-head attention mechanism, which improved the stability and extendibility without limiting the representational ability. Zhou et al. [19] proposed a novel framework, LSF-SAC, which took a variational inference-based information-sharing mechanism as extra state information to assist individuals in value function factorization. Making full use of shared latent state information, LSF-SAC significantly expanded the ability of value function factorization and performed well in SMAC.

In addition, some knowledge in other fields has also been introduced into the VF method. Zhang et al. [20] proposed an attention-based method, AVD-Net (attention value decomposition network), which introduced attentional mechanism into VDN and QMIX to learn the correlation between individuals, so as to effectively factorize joint value function and urge agents to learn the cooperative relationship adaptively. Wu et al. [21] combined attention mechanism and communication to learn a more general VF
formula, so as to complete multi-agent collaboration in complex environments. Liu et al. [22] proposed an attention relational encoder (ARE) for the representation of attentional relational state in decentralized MARL. It used attention mechanism to aggregate the information of neighbor agents to expand the observation space of agents, which improved computational efficiency and flexibility and performed well in the micromanagement tasks of StarCraft II. Iqbal et al. [23] proposed a randomized entity-wise factorization method for imagined learning, REFIIL, which randomly selected a subset of observed entities and predicted the utility of each agent in these subsets. Making full use of the independence of agents, the learning efficiency and extendibility of REFIIL were greatly improved.

Our method also draws on the idea of correlation coefficient in mathematics, which has been applied to multi-agent systems. Wan et al. [24] proposed an energy-efficient sleep scheduling mechanism with similarity measure for wireless sensor networks, which divided sensor nodes into different categories by measuring the similarity degree of member nodes through correlation coefficient and effectively reduced energy consumption. Zhang et al. [25] proposed an agent-level coordination-based MARL method, which screens the communication messages between agents by analyzing the correlation between individual agents based on the Pearson, Spearman, and Kendall correlation coefficients, effectively solving the problem of the high dimensionality of input space of the state-action value network.

3. Background

This section introduces the basic theories related to VF, mainly including the modeling method, correlation measurement approach, and classical VF algorithms.

3.1. Decentralized Partially Observable Markov Decision Process (Dec-POMDP).

Dec-POMDP is a classical modeling method of MADRL. A fully cooperative MARL task can be described as a Dec-POMDP denoted by a tuple $G = \langle N, S, U, P, r, Z, O, \gamma \rangle$ [26], where $N$ denotes the number of agents and $s \in S$ denotes the global state. At each time step, each agent $i$ selects an action $u^i \in U$ which forms a joint action $u \in U^N$. The state transition probability function $P(s'|s, u): S \times U \times S \rightarrow [0, 1]$ describes a transition from state $s$ to state $s'$. $r(s, u): S \times U \rightarrow \mathbb{R}$ is the reward function shared by all agents. Also, individual observations $z \in Z$ are drawn by observation function $O(s, u): S \times A \rightarrow Z$. $\gamma \in (0, 1)$ is the discount factor.

Each agent has an action-observation history $h^i \in H \equiv (Z \times U)^\ast$, forming a joint action-observation history $h = (h^1, \ldots, h^n)$. Also, each agent selects actions according to a stochastic policy $\pi^i(u|h^i): H \times U \rightarrow [0, 1]$, which forms a joint policy $\pi = (\pi^1, \ldots, \pi^n)$. The joint action-value function is denoted by $Q^i(s, u) = E_{\pi \in \Pi} \max_{h^i \in H^i} \sum_{t=0}^{\infty} \gamma^t r_t$, where $R_t = \sum_{t=0}^{\infty} \gamma^t r_t$, is the discount return.

3.2. Spearman (Rank) Correlation Coefficient. In statistics, Spearman correlation coefficient [27, 28] is a non-parametric measure of rank correlation, which calculates the statistical dependence between the ranks of two variables. It measures the extent to which a monotone function is used to describe the correlation between two variables.

**Definition 1.** Suppose there are two groups of sample data $X$ and $Y$ with size $N$, and we assign the ranks $r_X$ and $r_Y$ to the elements $X_i$ and $Y_j$, respectively, according to their average descending position in the overall data. Spearman correlation coefficient is defined as follows:

$$r_s = \rho_{X,Y} = \frac{\text{cov}(r_X, r_Y)}{\sigma_{r_X} \sigma_{r_Y}},$$

where $\text{cov}(r_X, r_Y)$ is the covariance of the rank and $\sigma_{r_X}, \sigma_{r_Y}$ are the standard deviations of the rank.

When there are no elements with the same value in the sample, Spearman correlation coefficient can also be calculated by the following formula.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where $d_i = r_{Xi} - r_{Yi}$ is the difference between the corresponding ranks of the $i$-th pair elements. Table 1 shows the value and corresponding ranks of elements in both cases.

Since there is no constraint on the distribution and continuity of the variable, Spearman correlation coefficient can measure the monotonic correlation between two sets of ordered data. It also indicates the direction of the correlation between variables $X$ and $Y$. If $Y$ increases as $X$ increases, then Spearman correlation coefficient is positive; otherwise, Spearman correlation coefficient is negative or 0. The closer to complete monotonic correlation, the greater the absolute value of Spearman correlation coefficient. In the case of complete monotonic correlation, the absolute value is 1.

3.3. QMIX. QMIX is a classical VF method, which represents the joint action value function $Q_{\text{tot}}(h, u)$ as a complex non-linear combination of individual action value functions $Q^i$ based only on local observation. It makes the result of $\arg\max Q_{\text{tot}}(h, u)$ consistent with on each $Q^i$ by imposing monotonicity constraints on $Q_{\text{tot}}(h, u)$ and $Q^i$. That is, when (3) holds, (4) can be derived.

$$\frac{\partial Q_{\text{tot}}(h, u)}{\partial Q^i} \geq 0 \quad \forall i \in \{1, 2, \ldots, n\},$$

$$\arg\max_u Q_{\text{tot}}(h, u) = \begin{pmatrix} \arg\max_u Q^1 \\ \vdots \\ \arg\max_u Q^n \end{pmatrix}.$$  

QMIX adopts three neural networks. (1) Individual network: the individual network is represented by DRQN [29], which calculates individual value function $Q^i(h^i, u^i)$ according to local observation $o^i$ and the last action $u^i_{t-1}$ at each time step. (2) Hypernetwork: the hypernetwork
4. Correlation between Individuals in Value Factorization

As we all know, the VF method for MADRL is an effective approach to solve the multi-agent collaboration problem, which mainly uses IGM principle and attention mechanism to evaluate the contribution of agents, so as to realize multi-agent collaboration. However, the existing VF algorithms only evaluate individual impact on the whole system and ignore the interaction between individuals. Such interaction is sometimes critical, especially in a cooperative environment [25]. For example, an attacking agent with “weak defense” but “high damage and long range” attributes usually relies on a defensive agent with “strong defense and high health” attributes. Without the protection of defensive agents, the attacking agent is easy to become enemies’ prime target and be killed effortlessly, which will result in a significant reduction in combat effectiveness. Therefore, the interaction between individuals should be considered appropriately to urge agents to select actions that are beneficial not only to themselves but also their partners, so as to learn the optimal joint policy.

The interaction between individuals reflects the correlation between individuals, which will be embodied in the variation trend of agent’s individual value function in cooperative environment. In other words, if two agents interact with each other, their individual value functions corresponding to the selected actions in the training process should be relevant to some extent. That is, \( Q_i \) will increase or decrease as \( Q_j \) increases. For example, the individual value function corresponding to “attack” will increase if an agent receives a positive reward after hitting an enemy. At the moment, if another relevant agent selects “charge,” it is more likely to kill the enemy, and the individual value function corresponding to “charge” also increases.

In order to fully consider the interaction between individuals, we propose a novel VF method, CI-VF, to study the correlation between individuals. Our method parameterizes the interaction between individuals according to Spearman correlation coefficient of individual value functions and takes the parameterized results as an auxiliary term to train network, whose aim is to improve the multi-agent collaborative ability by stimulating this correlation. The individual value function vector in each round is a set of ordered data, and the correlation between individuals can be measured by the change trend of the set of ordered data. Therefore, it is reasonable and feasible to use Spearman correlation coefficient to describe the correlation degree of individual value functions. The agent will select the best action for itself and its partners by using this correlation, so as to learn the optimal joint policy.

The overall framework of CI-VF is shown in Figure 2. Weighted network module represents the weight calculation process of individual value functions, including hypernetwork and attention mechanism. Mixing network module represents the integration process of individual value functions to joint value functions in the VF method. Spearman correlation module is the main work of this paper, which represents the quantification process of the correlation between individual agents. The module is designed as follows.

In each round, agents iterate \( T \) steps. Also, at each time step, agent \( i \) will select an action and obtain the individual action-value function corresponding to the action, denoted as \( Q_i^t \). The individual value function in each round is denoted

| Case 1: with the same value | Case 2: without the same value |
|-----------------------------|-------------------------------|
| \( X \) | \( rX \) | \( X \) | \( rX \) |
| 3   | 2   | 1.2 | 2  |
| 4   | 3.5 | 5   | 4  |
| 2   | 1   | 4   | 3  |
| 4   | 3.5 | 7   | 5  |
| 7   | 5   | 0   | 1  |

## Table 1: The corresponding ranks of elements in the sample.

4 Mathematical Problems in Engineering
Figure 1: Overall architecture of QMIX.

[Diagram of QMIX architecture]

Figure 2: The overall framework of CI-VF.

[Diagram of CI-VF framework]

Then, we can get a new joint value function $Q_{tot}^i = Q_{tot} + r_{s}^i$ and the loss function $L_r(\theta)$. $L_r(\theta)$ is defined as

$$L_r(\theta) = \sum_{i=1}^{n} \left( Q_{tot}^i - Q_{tot}(h, u) \right)^2.$$  \hspace{1cm} (10)

The pseudocode of CI-VF is as follows (Algorithm 1).

5. Experiments

5.1. Settings. In this section, we evaluate the performance of CI-VF in the StarCraft II micromanagement mission and compare it with the classic VF algorithms such as QMIX, VDN, IQL, and so on. During training, our units are controlled by CI-VF, while the enemies are controlled by the built-in AI algorithm. Also, all training parameters are the same as QMIX in SMAC.

CI-VF mainly includes CIVDN and CIQMIX, which are proposed for VDN and QMIX, respectively. We evaluate CIVDN and CIQMIX on four different types of maps, including 8m, 2s3z, 3s_vs_5z, and MMM2, and compare them to baselines. The features of these maps are shown in Table 2.

5.2. Validation. First, we evaluate CIQMIX and CIVDN in easy scenarios as shown in Figures 3(a) and 3(b), and they complete multi-agent collaboration tasks well both in homogeneous and heterogeneous scenarios. In addition, compared with previous algorithms, their learning efficiency and stability are slightly improved on the whole, but the improvement is not obvious. It is mainly because previous algorithms also work well in easy scenarios. IQL has significantly bad performance mainly because it is a poor multi-agent credit assignment scheme that does not take into account the interaction between individuals.

Then, we test CIQMIX and CIVDN in the hard scenario, 3s_vs_5z, and results are presented in Figure 3(c). 3s_vs_5z is a homogeneous and asymmetric scenario, in which our agents must learn to keep distance with enemies to disperse enemies’ forces and then surround and kill them one by one.
Table 2: The features of 8m, 2s3z, 3s_vs_5z, and MMM2.

| Maps     | Scenario complexity | Combat group   | Combat pattern |
|----------|---------------------|----------------|----------------|
| 8m       | Easy                | Homogeneous    | Symmetric      |
| 2s3z     | Easy                | Heterogeneous  | Symmetric      |
| 3s_vs_5z | Hard                | Homogeneous    | Asymmetrical   |
| MMM2     | Super hard          | Heterogeneous  | Asymmetrical   |

Algorithm 1: Value factorization method based on correlation between individuals (CI-VF).

For episode = 1, ..., M do
  For i = 1, ..., N do
    Agent i greedily selects action $u_i^t$ based on $Q_{t-1}^i(s_{t-1}, u_{t-1}, \theta)$
    Obtain the joint action $u_t$
    End for
  Transfer to next state $s_{t+1}$ and get total reward $r_t$ after executing $u_t$
  Store the sample $(s_t, o_t, u_t, r_t, s_{t+1})$ to the experience pool $D$
  For i = 1, ..., N do
    Small batch samples $(s_i, o_i, u_i, R_i, s_{i+1})$ are randomly taken from $D$
    Calculate individual correlation coefficient $r_{i, j}^s$ and joint correlation coefficient $r_{j, tot}^s$ according to samples, so as to get new joint value function $Q_{tot}^s$
    Set $y_t = \begin{cases} r_j & \text{for terminal } s \\ r_j + y \max_{a, \theta} Q_{tot}^s(s, a, \theta) & \text{for non-terms} \end{cases}$
    Update the network parameters $\theta$ by the gradient descent method
  End for
  Update parameters $\theta^- = \theta$ every $K$ steps
End for

Figure 3: Continued.
Hisputsahighdemandonalgorithmsexperiments. Experimental results show that CI-VF still works well in asymmetric scenario. Compared with previous algorithms, its performance is improved significantly, especially CIQMIX. Obviously, this improvement owes to the consideration of the correlation between individuals.

Finally, we test CIQMIX and CIVDN in super-hard level scenario, MMM2, and the results are shown in Figure 3(d). MMM2 is a heterogeneous and asymmetric scenario, consisting of Medivac, Marauder, and Marine, in which we had to rely on three types of units working together to win. In other words, Medivac draws enemies’ fire and heals the wounded, and Marauder breaks through enemies’ defense and harvests wounded enemies, while Marine provides protection and attacks enemies’ Medivac. Therefore, it is necessary to take account of interaction between individuals. Experimental results also indicate that CI-VF stands out in performance than previous algorithms. This just further proves the necessity and validity of considering correlation between individuals.

In a word, whether the scenario is homogeneous or heterogeneous, easy or hard, CI-VF has the better performance than baselines.

In order to test the performance of all algorithms clearly and comprehensively, we conduct 5000 rounds of tests on the trained models to obtain the statistical win rate, which is the ratio of the rounds of winning to the total test rounds, with a maximum of 1. The results are shown in Table 3, and the highest values are given in bold.

| Test win rate | CIQMIX | CIVDN | QMIX [7] | VDN [6] | IQL [9] |
|---------------|--------|-------|----------|---------|---------|
| 8m            | 0.9822 | 0.9946| 0.9890   | 0.9898  | 0.6184  |
| 2s3z          | 0.9880 | 0.9846| 0.9878   | 0.9808  | 0.9526  |
| 3s_vs_5z      | 0.9936 | 0.9962| 0.9876   | 0.9902  | 0.1964  |
| MMM2          | 0.6834 | 0.7622| 0.0002   | 0.3596  | 0.3730  |

From the table, we can find that CI-VF has almost the best performance, and the final win rates are higher than those of baselines in all scenarios, especially in MMM2.

6. Conclusions and Future Work

In this paper, we propose a MADRL VF method based on correlation between individuals, which provides a new perspective for solving multi-agent credit assignment problem. We first analyzed the necessity of considering interaction between individuals, then introduced Spearman correlation coefficient to measure interaction between individuals and analyzed its rationality and feasibility, and finally optimized joint value function according to joint correlation coefficient obtained by individual value function. Experimental results in SMAC environment show that our method has a better performance than baselines in various scenarios.

For future work, it will be important to consider whether there are better ways to measure interaction between individuals more than Spearman correlation. For example, designing a correlation network that calculates interaction between individuals according to observations and rewards may be worthy of further study.

**Data Availability**

No data were used to support this study.

**Conflicts of Interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper.

**Acknowledgments**

This study was supported by the National Natural Science Foundation of China (61806221).
References

[1] W. Guo, N. Xiong, H.-C. Chao, S. Hussain, and G. Chen, “Design and analysis of self-adapted task scheduling strategies in wireless sensor networks,” Sensors, vol. 11, no. 7, pp. 6535–6554, 2011.

[2] K. Gao, F. Han, P. Dong, N. Xiong, and R. Du, “Connected vehicle as a mobile sensor for real time queue length at signalized intersections,” Sensors, vol. 19, no. 9, p. 2059, 2019.

[3] X. Wang, Q. Li, N. Xiong, and Y. Pan, “Ant colony optimization-based location-aware routing for wireless sensor networks,” in Proceedings of the International conference on wireless algorithms, systems, and applications, pp. 109–120, Springer, Nanjing, China, June 2008.

[4] Y. Jiang, G. Tong, H. Yin, and N. Xiong, “A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters,” IEEE Access, vol. 7, pp. 118310–118321, 2019.

[5] Y. Lu, S. Wu, Z. Fang, N. Xiong, S. Yoon, and D. S. Park, “Exploring finger vein based personal authentication for secure IoT,” Future Generation Computer Systems, vol. 77, pp. 149–160, 2017.

[6] P. Sunehag et al., “Value-decomposition networks for cooperative multi-agent learning based on team reward,” in Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 2085–2087, Stockholm, Sweden, July 2018.

[7] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, “Qmix: monotonic value function factorisation for deep multi-agent reinforcement learning,” in Proceedings of the International Conference on Machine Learning, pp. 4295–4304, PMLR, Ukraine, July 2018.

[8] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, p. 1, Louisiana USA, February 2018.

[9] A. Tampuu, T. Matisen, D. Kodelja et al., “Multiagent cooperation and competition with deep reinforcement learning,” PLoS One, vol. 12, no. 4, p. e0172395, 2017.

[10] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, “Weighted QMIX: expanding monotonic value function factorisation for deep multi-agent reinforcement learning,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[11] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, “Qtran: learning to factorize with transformation for cooperative multi-agent reinforcement learning,” in Proceedings of the International Conference on Machine Learning, pp. 5887–5896, PMLR, California, USA, June 2019.

[12] Y. Yang et al., “Qatten: a general framework for cooperative multi-agent reinforcement learning,” 2020, https://arxiv.org/abs/2002.03939.

[13] M. N. Alpdmir, “Tactical UAV path optimization under radar threat using deep reinforcement learning,” Neural Computing & Applications, vol. 34, no. 7, pp. 5649–5664, 2022/04/01 2022.

[14] J. Jia, R. Yu, Z. Du, J. Chen, Q. Wang, and X. Wang, “Distributed localization for IoT with multi-agent reinforcement learning,” Neural Computing & Applications, vol. 34, no. 9, pp. 7227–7240, 2022/05/01 2022.

[15] Z. Xu, D. Li, Y. Bai, and G. Fan, “MMD-MIX: value function factorisation with maximum mean discrepancy for cooperative multi-agent reinforcement learning,” in Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–7, Shenzhen, China, June 2021.

[16] J. Zhao, M. Yang, X. Hu, W.-g. Zhou, and H. Li, “DQMIX: a distributional perspective on multi-agent reinforcement learning,” vol. abs/2202, 2022 https://arxiv.org/abs/2202.10134, Article ID 10134.

[17] K. Son, S. Ahn, R. D. Reyes, J. Shin, and Y. Yi, “QTRAN++: improved value transformation for cooperative multi-agent reinforcement learning,” 2020, https://arxiv.org/abs/2006.12010.

[18] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, “QPLEX: duplex dualing multi-agent Q-learning,” in Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, April 2020.

[19] H. Zhou, T. Lan, and V. Aggarwal, “Value functions factorization with latent state information sharing in decentralized multi-agent policy gradients,” 2022, https://arxiv.org/abs/2201.01247.

[20] Y. Zhang, H. Ma, and Y. Wang, “AVD-net: attention value decomposition network for deep multi-agent reinforcement learning,” in Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7810–7816, IEEE, Milano, Italy, January 2021.

[21] B. Wu, X. Yang, C. Sun, R. Wang, X. Hu, and Y. Hu, “Learning effective value function factorization via attentional communication,” in Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 629–634, IEEE, Toronto, ON, Canada, October 2020.

[22] X. Liu and Y. Tan, “Attentional relational state representation in decentralized multiagent reinforcement learning,” IEEE Transactions on Cybernetics, vol. 52, no. 1, pp. 252–264, 2022.

[23] S. Iqbal, C. A. S. De Witt, B. Peng, W. Böhmer, S. Whiteson, and F. Sha, “Randomized entity-wise factorization for multi-agent reinforcement learning,” in Proceedings of the International Conference on Machine Learning, pp. 4596–4606, PMLR, July 2021.

[24] R. Wan, N. Xiong, and N. T. Loc, “An energy-efficient sleep scheduling mechanism with similarity measure for wireless sensor networks,” Human-centric Computing and Information Sciences, vol. 8, no. 1, p. 18, 2018.

[25] Y. Zhang, Q. Yang, D. An, and C. Zhang, “Coordination between individual agents in multi-agent reinforcement learning,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 13, pp. 11387–11394, 2021.

[26] F. A. Oliehoek, M. T. J. Spaan, and N. Vlassis, “Optimal and approximate Q-value functions for decentralized POMDPs,” Journal of Artificial Intelligence Research, vol. 32, pp. 289–353, 2008.

[27] J. H. Zar, “Spearman rank correlation,” Encyclopedia of biostatistics, vol. 7, 2005.

[28] C. Croux and C. Dehon, “Influence functions of the Spearman and Kendall correlation measures,” Statistical Methods and Applications, vol. 19, no. 4, pp. 497–515, 2010.

[29] M. Hausknecht and P. Stone, “Deep recurrent q-learning for partially observable mdps,” in Proceedings of the 2015 aaai fall symposium series, Arlington, Virginia, November 2015.