

Diversifying Message Aggregation in Multi-Agent Communication via Normalized Tensor Nuclear Norm Regularization

Yuanzhao Zhai¹, Kele Xu¹, Bo Ding^{1*}, Dawei Feng¹, Zijian Gao¹, Huaimin Wang¹

¹ National University of Defense Technology, Changsha 410005, China
 yuanzhaozhai@nudt.edu.cn, kelele.xu@gmail.com, dingbo@nudt.edu.cn,
 davyfeng.c@gmail.com, gaozijian19@nudt.edu.cn, whm_w@163.com

Abstract

Aggregating messages is a key component for the communication of multi-agent reinforcement learning (Comm-MARL). Recently, it has witnessed the prevalence of graph attention networks (GAT) in Comm-MARL, where agents can be represented as nodes and messages can be aggregated via the weighted passing. While successful, GAT can lead to homogeneity in the strategies of message aggregation, and the “core” agent may excessively influence other agents’ behaviors, which can severely limit the multi-agent coordination. To address this challenge, we first study the adjacency tensor of the communication graph and demonstrate that the homogeneity of message aggregation could be measured by the normalized tensor rank. Since the rank optimization problem is known to be NP-hard, we define a new nuclear norm, which is a convex surrogate of normalized tensor rank, to replace the rank. Leveraging the norm, we further propose a plug-and-play regularizer on the adjacency tensor, named *Normalized Tensor Nuclear Norm Regularization* (NTNNR), to actively enrich the diversity of message aggregation during the training stage. We extensively evaluate GAT with the proposed regularizer in both cooperative and mixed cooperative-competitive scenarios. The results demonstrate that aggregating messages using NTNNR-enhanced GAT can improve the efficiency of the training and achieve higher asymptotic performance than existing message aggregation methods. When NTNNR is applied to existing graph-attention Comm-MARL methods, we also observe significant performance improvements on the StarCraft II micromanagement benchmarks.

Introduction

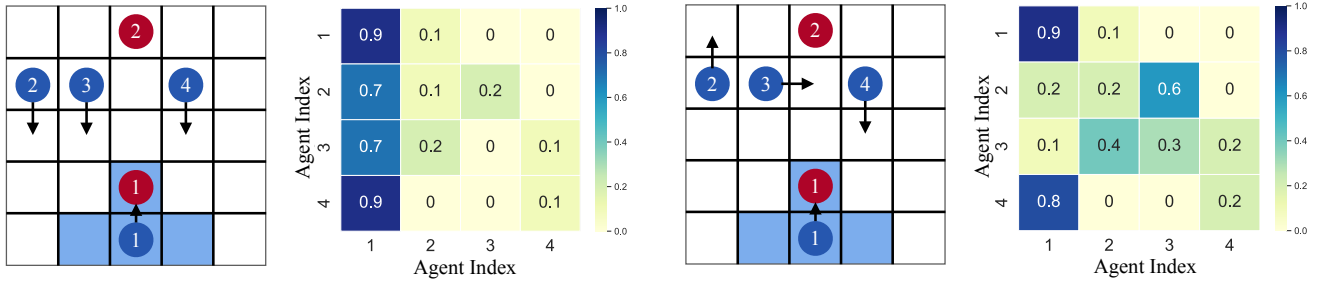
Multi-Agent Reinforcement Learning (MARL) has achieved remarkable success in a range of challenging sequential decision-making tasks, such as traffic control (Zhou et al. 2020), swarm robotics (Zhai et al. 2021) and multi-player strategy games (Yuan et al. 2022). As an under-explored issue in MARL, communication is a key component for multi-agent coordination where agents can exchange their local observations via communication messages. These messages are aggregated by decentralized agents and further utilized to augment individual local observations for learning policies and selecting actions, allowing the agents to jointly optimize the objectives.

Although sustainable efforts have been made, efficient communication between agents is still far from being solved. How to aggregate messages is a key factor that determines communication efficiency. To model the interactions between agents, MARL has widely utilized graph neural networks (GNNs) (Scarselli et al. 2008) to allow for a graph-based representation. The multi-agent system is usually modeled as a complete graph, and each agent corresponds to a node. As one of the most popular GNNs variants, GAT has shown great potential in Comm-MARL (Zhu, Dastani, and Wang 2022). Message aggregation can be achieved via attention-weighted message passing in the communication graph.

Despite the success of the GAT in Comm-MARL, we show that a lack of diversity still persists in the obtained message aggregation strategy. In essence, many nodes in the graph may pay undue attention to a few “key” nodes and are often excessively influenced. This issue is identified in various tasks modeled by GAT (Brody, Alon, and Yahav 2022), and multiple agents exacerbate the problem severely. For the multi-agent scenarios where the importance of messages is conditioned on agents’ state, homogeneous message aggregation strategies mean most agents may pay excessive attention to some emergent message, resulting in inefficient communication. Moreover, since many Comm-MARL methods adopt the parameter-sharing scheme, agents with homogeneous message aggregation strategies tend to obtain similar behaviors, severely limiting the diversity of behaviors for better coordination (Chenghao et al. 2021). As shown in Figure 1, the behavior obtained by methods with homogeneous message aggregation strategies can be suboptimal, highlighting the urgent need for diverse message aggregation strategies.

In this paper, we aim to enable agents to explore diverse message aggregation strategies. Firstly, we study the adjacency tensor of the multi-agent communication graph, which consists of adjacency matrices generated by the multi-head attention mechanism of GAT. We present that the homogeneity of message aggregation could be measured by the normalized tensor rank and normalized tensor nuclear norm. Accordingly, we propose a novel *Normalized Tensor Nuclear Norm* (NTNN) regularizer, which regularize adjacency tensors to actively enrich the diversity of the message aggregation strategies in Comm-MARL. In this way, agents

*Corresponding authors: Kele Xu and Bo Ding



(a) With homogeneous message aggregation strategies obtained by GAT, predators 2, 3, and 4 are excessively influenced by the message of predator 1. All predators tend to pursue prey 1 while ignoring prey 2.

(b) With diverse message aggregation strategies, predators 1 and 2 intend to capture prey 1, while predators 3 and 4 explore the environment. In this way, the predators could obtain better coordination.

Figure 1: A toy experiment in the predator-prey scenario. Predators are marked in blue, and preys are in red. Four predators with a limited vision of one grid are pursuing two static preys through communication. Exact two predators are required to be present in the grid cell of a prey for a successful capture. Element at row i and column j of the adjacency matrix represent the attention score of agent j 's communication message to agent i .

could discover diverse behaviors and tend to find better coordination. In brief, our main contribution is threefold:

- We firstly propose to measure the diversity (or the homogeneity) of the message aggregation via the normalized tensor rank of the adjacency tensor.
- We define a novel normalized tensor nuclear norm to replace the rank. The norm can be further utilized as the regularizer to discover diverse message aggregation strategies for multi-agent communication.
- Experiments show that aggregating messages using GAT with NTNNR can improve training efficiency and asymptotic performance. Our regularizer also brings significant performance improvements for existing graph-attention Comm-MARL methods, using the plug-and-play manner.

Related Work

Attention in Graphs

For graph-structured data, attention mechanisms have been widely used to model the pairwise interactions between nodes. For example, many previous attempts employed GNNs with attention mechanisms (Lee et al. 2019; Brody, Alon, and Yahav 2022), which generalizes the standard node representation update pattern, e.g., averaging or max-pooling of neighbors (DKipf and Welling 2017; Hamilton, Ying, and Leskovec 2017). During the message passing, the attention mechanism allow nodes to compute a weighted average of their neighbors, and softly select their most relevant neighbors. GAT (Veličković et al. 2018) is one of the most popular GNNs variants. GAT generalizes the multi-head self-attention mechanism (Vaswani et al. 2017) from sequences to graphs, which allows the model to attend to information from different representation subspaces jointly.

Despite the effectiveness, GATv2 (Brody, Alon, and Yahav 2022) finds by a theoretical analysis that the ranking of the attention scores generated by GAT may be unconditioned on the query node, which is called the static attention

problem. To address this problem, GATv2 proposes to modify the order of weight calculation operation in GAT, outperforming GAT in many public datasets of GNNs.

Message Aggregation Methods in Comm-MARL

In Comm-MARL, message aggregation strategies for agents determine how to aggregate received messages and partial observation to select the next actions. Some works aggregate messages with no preference, such as concatenation (Foerster et al. 2016; Kim et al. 2019; Kim, Park, and Sung 2020), averaging (Sukhbaatar, Fergus et al. 2016; Singh, Jain, and Sukhbaatar 2019), summing up (Du et al. 2021), recurrent neural networks (Peng et al. 2017), and so on. Since messages encode the senders' personal understanding of their observations, some may be more important than others.

To aggregate messages unequally, the attention mechanism is often utilized to calculate received messages' weights and then aggregate them together (Das et al. 2019; Agarwal, Kumar, and Sycara 2020). Considering the graph topology, GAT has been proved an effective tool to aggregate messages (Liu et al. 2020; Li et al. 2021; Niu, Paleja, and Gombolay 2021). GA-Comm (Liu et al. 2020) propose a two-stage graph-attention mechanism. The hard attention determines whether communication between agents is necessary, while the soft attention calculates the attention weight. DICG (Li et al. 2021) introduces the deep implicit coordination graph with the self-attention mechanism for message aggregation.

Inherited from the static attention problem, most methods mentioned above lack diversity in terms of message aggregation. Although replacing GAT with GATv2 in Comm-MARL can prove the diversity of message aggregation theoretically, agents with similar observation still tend to obtain homogeneous message aggregation strategies in practice. Complementary with GATv2, our method regularize the adjacency tensor of GAT, encouraging diverse message aggregation actively.

Diversity in MARL

As an emerging topic in MARL, maintaining diversity for policies is meaningful for various scenarios, such as emergent behavior (Tang et al. 2020), exploration (Mahajan et al. 2019) or learning to adapt (Balduzzi et al. 2019). Diverse policies can be discovered by evolution methods (Cully et al. 2015; Pugh, Soros, and Stanley 2016), specially designed reward function (Lowe et al. 2019; Baker et al. 2019; Tang et al. 2020), role-based learning (Wang et al. 2020, 2021), population-based training (Vinyals et al. 2019; Parker-Holder et al. 2020; Lupu et al. 2021) or iterative policy optimization (Zhou et al. 2021; Zahavy et al. 2021).

Based on the value decomposition framework, some attempts aim to maintain diversity through non-shared individual Q-functions for each agent. EOI (Jiang and Lu 2021) combines the gradient from the intrinsic value function (IVF) and the total Q-function to train each agent’s local Q-function. CDS (Chenghao et al. 2021) maximize the mutual information between agents’ identities and their trajectories to diversify individual Q-functions. However, a recent work (Fu et al. 2022) theoretically shows that policy gradient with individual policy or communication can be comparable to popular value-based learning methods for maintaining diverse policies. They propose to obtain diverse policies with an auto-regressive policy gradient. Each agent selects actions according to different other agents’ actions, which can be seen as a Comm-MARL method.

Most of the aforementioned methods did not study diverse message aggregation strategies, which can be a potential way to enrich diversity in Comm-MARL. We adopt the policy gradient method with parameter sharing and utilize GAT with NTNNR to aggregate messages. To our best knowledge, we are the first to discover diverse policies through diversifying message aggregation.

Background and Notations

We model the multi-agent tasks as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) augmented with communication, which can be described as a tuple $\langle N, \mathcal{S}, \mathcal{U}, \mathbf{P}, \mathbf{R}, \mathcal{O}, \mathcal{M}, \mathcal{G}, \gamma \rangle$. N is the number of agents. \mathcal{S} represents the space of global states. \mathcal{O} denotes the space of observations of robots, and each agent receives a private observation $o_i \in \mathcal{O}$ according to the observation function $\sigma(s_i) : \mathcal{S} \rightarrow \mathcal{O}$. \mathcal{M} represents the space of messages. Agents generate messages $m_i \in \mathcal{M}$ encoded by its observations and others’ messages at the last timestep, which could be modeled by the multi-agent communication graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$. Node $v_i \in \mathcal{V}$ represent agents, and edges $e_{ij} \in \mathcal{E}$ represent communication links. We denote h_i as the feature of v_i . Combining with local observations, each agent aggregates communication messages and generates its own action $u_i = \pi_{\theta}(o_i, m_{j \neq i})$, where π_{θ} is the policy with parameter θ shared across all agents. For states $s, s' \in \mathcal{S}$ and a joint action $\mathbf{u} \in \mathcal{U}^N$, the transition probability of reaching state s' from state s by executing action \mathbf{a} is $\mathbf{P}(s'|s, \mathbf{u})$. \mathbf{R} is the joint reward function. $\gamma \in [0, 1]$ denotes the discount factor. Agent i aims to maximize its discounted reward $\mathbb{E}_{s \sim \rho_{\pi}, \mathbf{u} \sim \pi} [r_i^t] = \sum_{t=0}^{\infty} \gamma^t r_i^t(s^t, \mathbf{u}^t)$, where ρ_{π} is the

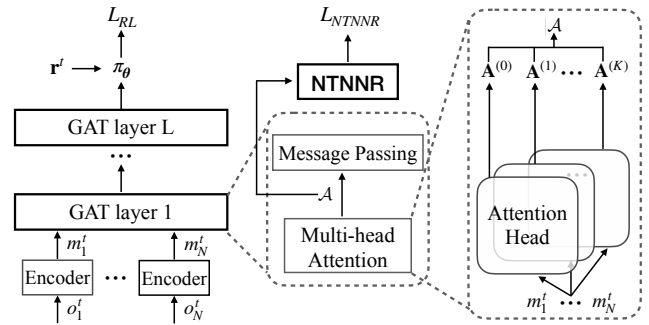


Figure 2: Schematics of NTNNR. We regularize the NTNN of the adjacency tensor \mathcal{A} which consists of adjacency matrices generated by the multi-head attention mechanism.

discounted state distribution induced by the policy.

In this paper, we denote adjacency tensors by boldface Euler script letters \mathcal{A} . Adjacency matrices are denoted by boldface capital letters \mathbf{A} ; vectors are denoted by boldface lowercase letters \mathbf{a} , and scalars are denoted by lowercase letters a . For the communication graph of GAT, adjacency matrices $\mathbf{A} \in \mathbb{R}_+^{N \times N}$ generated by the multi-head attention mechanism can be regarded as a three-way adjacency tensor $\mathcal{A} \in \mathbb{R}_+^{N \times N \times K}$, where the dimension of the third way is the number of attention heads K . We denote the (i, j, k) -th entry of \mathcal{A} as \mathcal{A}_{ijk} . The frontal slice $\mathcal{A}(:, :, k)$ is denoted compactly as $\mathbf{A}^{(k)}$.

Methodology

In this section, we describe the details of NTNNR (Figure 2), which actively enrich the diversity of message combination and can be integrated into graph-attention Comm-MARL methods.

Measuring Message Aggregation’s Diversity with the Normalized Tensor Rank

For each agent i , GAT computes a learnable weighted average of the representations of all neighbors $j \in \mathcal{N}_i$.

$$e(\mathbf{h}_i, \mathbf{h}_j) = \text{LeakyReLU}(\mathbf{W}'[\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j]), \quad (1)$$

where \mathbf{W} and \mathbf{W}' are learnable, and \parallel denotes vector concatenation.

We first consider the case that a single attention head is used. Then the attention scores, as the elements of the adjacency matrix \mathbf{A} , are normalized across all neighbors using the softmax function:

$$a_{ij} = \text{Softmax}_j(e(\mathbf{h}_i, \mathbf{h}_j)) = \frac{\exp(e(\mathbf{h}_i, \mathbf{h}_j))}{\sum_{j' \in \mathcal{N}_i} \exp(e(\mathbf{h}_i, \mathbf{h}_{j'}))}. \quad (2)$$

The adjacency matrix $\mathbf{A} \in \mathbb{R}_+^{N \times N}$ satisfies the following properties:

$$\begin{cases} \sum_{j=1}^N a_{ij} = 1 & \forall i \in 1, \dots, N, \\ a_{ij} \geq 0 & \forall i \in 1, \dots, N, j \in 1, \dots, N. \end{cases} \quad (3)$$

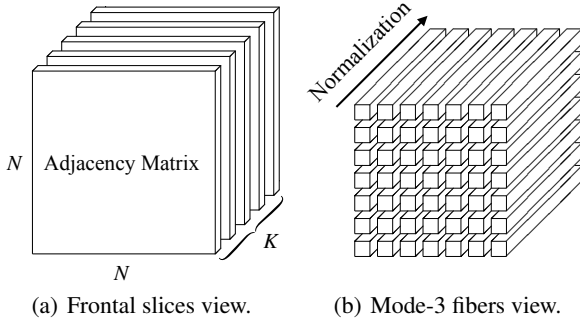


Figure 3: Multiple views of the adjacency tensor \mathcal{A} .

We denote the vectors selected from the i -th and j -th rows of the matrix as \mathbf{a}_i and \mathbf{a}_j , which represent the attention scores of agent i and j respectively for aggregating messages. If agent i and j have homogeneous message aggregation strategies, the difference between \mathbf{a}_i and \mathbf{a}_j is minor. In this case, \mathbf{a}_i and \mathbf{a}_j could be approximately regarded as linearly dependent. On the contrary, diverse message aggregation strategies mean linearly independent vectors. Therefore, we could measure the diversity (or the homogeneity) of the message aggregation with the matrix rank of the adjacency matrix \mathbf{A} .

With the multi-head attention mechanism in GAT, independent K attention mechanisms execute the attention function in parallel. Then we obtain a three-way adjacency tensor $\mathcal{A} \in \mathbb{R}_+^{N \times N \times K}$. As shown in Figure 3(a), K frontal slices $\{\mathbf{A}^{(k)}\}_{k=1, \dots, K}$ represent independent adjacency matrices. From another view, as shown in Figure 3(b), the mode-3 fiber $\mathcal{A}(i, j, :)$ represents the attention scores of agent j to agent i using different attention heads. Multiple heads are considered to attend to information from different representation subspaces. Thus we aim to maintain the diversity in both the frontal slices view and the mode-3 fibers view of the adjacency tensor.

Extended from the matrix rank, tensor rank could be defined in various ways. CP rank (Kolda and Bader 2009) denotes the smallest number of rank one tensor decomposition. But both CP rank and its convex relaxation is hard to obtain. To avoid this issue, the tractable Tucker rank (Kolda and Bader 2009) and its convex relaxation are more widely used. However, most existing tensor ranks can not directly measure the linear correlation from both frontal slices and mode-3 fibers views. This motivates us to define a new tensor rank to measure the homogeneity of message aggregation with multi-head attention GAT.

We denote $\hat{\mathcal{A}}$ as a result of applying normalization to \mathcal{A} along the 3-rd way. Specifically, we apply Softmax on every tube fibers $\mathcal{A}(i, j, :)$, i.e.,

$$\hat{\mathcal{A}}_{ijk} = \frac{\exp(\mathcal{A}_{ijk})}{\sum_{l \in [0, K-1]} \exp(\mathcal{A}_{ijl})}, \quad K \geq 2. \quad (4)$$

Then we can define the normalized tensor rank as:

$$\text{rank}_n(\mathcal{A}) = \sum_k \text{rank}(\hat{\mathbf{A}}^{(k)}). \quad (5)$$

Normalized Tensor Nuclear Norm Regularization

The rank optimization problem is known to be NP-hard. An alternative is to utilize the nuclear norm, and the matrix nuclear norm is defined as:

$$\|\mathbf{A}\|_* = \sum_i \sigma_i(\mathbf{A}), \quad (6)$$

where $\sigma_i(\mathbf{A})$ are singular values of \mathbf{A} .

For normalized tensor $\hat{\mathcal{A}}$, we denote $\hat{\mathbf{A}} \in \mathbb{R}_+^{NK \times NK}$ as the block diagonal matrix with its i -th block on the diagonal as the i -th frontal slice, i.e.,

$$\hat{\mathbf{A}} = \text{bdiag}(\hat{\mathcal{A}}) = \begin{bmatrix} \hat{\mathbf{A}}^{(0)} & & & \\ & \hat{\mathbf{A}}^{(1)} & & \\ & & \ddots & \\ & & & \hat{\mathbf{A}}^{(K-1)} \end{bmatrix}. \quad (7)$$

Based on the matrix nuclear norm, we define a novel tensor nuclear norm for the normalized tensor rank, which is called *Normalized Tensor Nuclear Norm* (NTNN):

$$\|\mathcal{A}\|_* = \frac{1}{K} \|\hat{\mathbf{A}}\|_*. \quad (8)$$

As a special case, if \mathcal{A} reduces to a matrix ($K = 1$), it is not necessary to normalize the third dimension. In this case, NTNN reduces to the matrix nuclear norm. Considering the nuclear norm is the convex relaxation of the matrix rank (Candès and Recht 2009), $\|\hat{\mathbf{A}}\|_*$ is a tight convex surrogate of $\text{rank}(\hat{\mathbf{A}})$. Combining Equation 5 and 7, $\|\mathcal{A}\|_*$ is a tight convex surrogate of $\text{rank}_n(\mathcal{A})$.

Regularizing NTNN of the adjacency tensor \mathcal{A} could maintain the diversity of message aggregation. With the increase of NTNN, the diversity of \mathcal{A} is enriched not only in the frontal slices view but also in the mode-3 fibers view, which makes agents' message aggregation strategies more diverse.

Overall Optimization Objective

In this part, we describe how to use NTNNR to diversify message aggregation strategies in graph-attention Comm-MARL algorithms.

Following most Comm-MARL methods, we implement our framework with the policy decentralization with shared parameters (PDSP) paradigm. Then the gradient of Comm-MARL's original loss function can be formulated as:

$$\nabla_{\theta} L_{RL}(\theta) = \mathbb{E}_{i,t} [\nabla_{\theta} \log \pi_{\theta}(u_i^t | o_i^t, m_{j \neq i}^t) \Psi_i^t], \quad (9)$$

where Ψ_i^t is related to the discounted reward r_i^t and has various forms depending on different algorithms (Schulman et al. 2015), and θ denotes all parameters of the policy network.

To discover diverse message aggregation strategies, we apply NTNNR to the adjacency tensor \mathcal{A} of GAT layers. The corresponding loss function of NTNNR in the l -th layer can be formulated as:

$$L_{NTNNR}(\theta_l) = -\|\mathcal{A}\|_*, \quad (10)$$

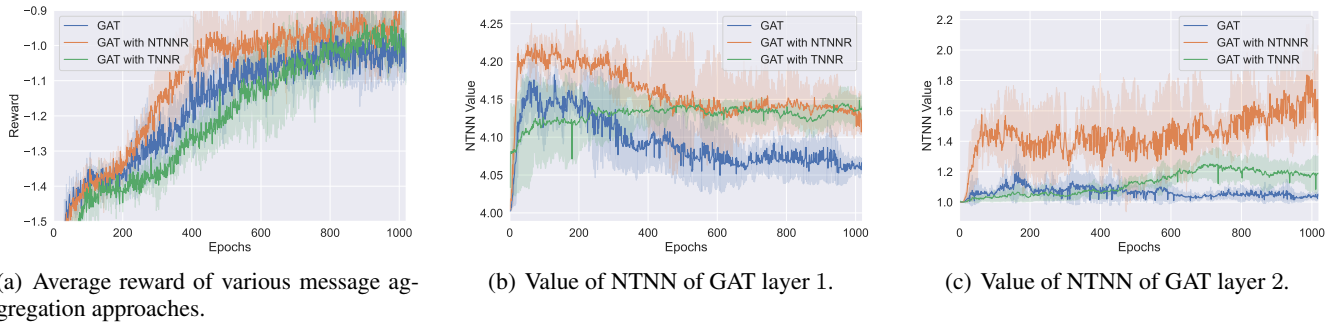


Figure 4: A case study in the predator-prey scenario indicates that NTNNR encourages diverse policies, thus achieving higher training efficiency and asymptotic performance.

Algorithm 1: Comm-MARL with NTNNR

Initialization: the number of agents N , the number of the communication graph layers L , parameters of the policy network θ

```

1: while Training do
2:    $L(\theta) \leftarrow 0$ 
3:   for each agent  $n \in \text{range}(N)$  do
4:     Calculate the Comm-MARL loss  $L_{RL}(\theta)$ 
5:      $L(\theta) \leftarrow L(\theta) + L_{RL}(\theta)$ 
6:   end for
7:   for each communication graph layer  $l \in \text{range}(L)$  do
8:     if Number of attention heads is greater than 1 then
9:       Normalize the adjacency tensor as Equation 4.
10:    end if
11:     $L_{NTNNR}(\theta_l) = -\|\mathcal{A}\|_*$ 
12:    Calculate the  $\lambda_l$  as Equation 12.
13:     $L(\theta) \leftarrow L(\theta) + \lambda_l L_{NTNNR}(\theta_l)$ 
14:  end for
15:   $\theta \leftarrow \text{optimize}(L(\theta))$ 
16: end while

```

where θ_l is part of parameters θ to obtain the adjacency tensor \mathcal{A} of the l -th GAT layer.

Overall, we update the model parameter θ by minimizing the following loss function:

$$L(\theta) = L_{RL}(\theta) + \sum_l \lambda_l L_{NTNNR}(\theta_l), \quad (11)$$

where λ_l is the regularization weights of NTNNR for layer l . To anneal λ_l during the training process, we introduce new scaling hyper-parameters β_l and obtain adaptive weight as follows:

$$\lambda_l = \frac{|L_{RL}(\theta)|}{\beta_l \times |L_{NTNNR}(\theta_l)|}. \quad (12)$$

Algorithm 1 details how NTNNR is integrated with generic Comm-MARL algorithms.

Experimental Results

In this part, we evaluate the performance of NTNNR in three widely-used scenarios: Predator-Prey, Traffic Junction, and

StarCraft II Multi-Agent Challenge.

In the mixed cooperative-competitive predator-prey scenario, we conduct ablation studies to show the effectiveness of NTNNR. We compare GAT with NTNNR with two baselines: vanilla GAT and applying our defined tensor nuclear norm without normalization (TNNR) to GAT. In the cooperative traffic junction scenario, we compare our proposed message aggregation method, GAT with NTNNR, against a variety of widely used message aggregation methods, including averaging used in CommNet (Sukhbaatar, Fergus et al. 2016), signature-based attention mechanism used in TarMAC (Das et al. 2019), GAT used in GA-Comm (Liu et al. 2020) and MAGIC (Niu, Paleja, and Gombolay 2021), and GATv2 (Brody, Alon, and Yahav 2022). Following the experimental setup in MAGIC, we utilize the two-layer GAT. The first layer contains two attention heads in the predator-prey scenario and four in the traffic junction scenario, while the second layer always contains one. For all methods, we uniformly adopt the REINFORCE (Williams 1992) with baseline as the training algorithm.

StarCraft II Multi-Agent Challenge (SMAC) (Whiteson et al. 2019) is a benchmark to evaluate various reinforcement learning works in recent years. Among them, we choose two state-of-the-art Comm-MARL methods, GA-Comm (Liu et al. 2020) and DICG-CE-LSTM¹ (Li et al. 2021), and then integrate NTNNR with them. All results are obtained by averaging over three runs.

Predator-Prey

Predator-Prey is one of the Multi-Agent Particle Environments (Lowe et al. 2017). In this scenario, we set 8 predators pursuing four fixed preys. To make the scenario mixed cooperative-competitive, two predators are required to be present in the grid cell of a prey for a successful capture. A predator obtains a reward of 0.3 if it captures a prey successfully. We set the maximum time steps to 30 and impose a step cost of 0.1.

We set scaling hyper-parameters to $\beta_1 = 0.2$, $\beta_2 = 0.005$ for the two GAT layers respectively. Figure 4(a) shows the average reward as the training epoch increases. Integrat-

¹DICG-CE-LSTM is the communication-augmented version of DICG.



(a) The frontal slices with TNNR.



(b) The frontal slices with NTNNR.

Figure 5: Visualization of the adjacency tensors generated by two-attention GAT with TNNR and NTNNR.

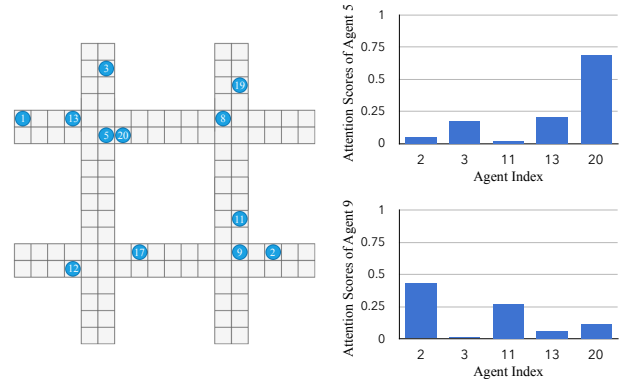
ing NTNNR into GAT when aggregating messages could boost the training efficiency, and obtain the highest asymptotic performance. Two baselines need more time to explore emergent strategies, demonstrating that NTNNR incentivizes more efficient exploration and, finally, achieves better coordination. We also record the corresponding NTNN values in Figure 4(b) and 4(c) respectively. We can observe that vanilla GAT keeps small NTNN values in both layers during the training stage, which suggests that all agents have homogeneous message aggregation strategies. In the early stage of training, GAT with NTNNR exhibits large NTNN values, encouraging agents to obtain more diverse message aggregation strategies and explore environments better.

Compared to TNNR, NTNNR can maintain larger NTNN values in the second layer with the same scaling hyperparameters. This is due to additional diversity among different attention heads. Note that even though GAT with NTNNR and with TNNR converges to similar NTNN values in the first layer, they have different message aggregation strategies. We visualize the adjacency tensors of two methods in similar states in Figure 5. Compared to TNNR, NTNNR can maintain diversity in inter and intra-frontal slices, indicating that normalization is critical for the regularizer when using the multi-head attention mechanism.

Traffic Junction

The second scenario we employ is cooperative. The hard-mode traffic junction scenario (Sukhbaatar, Fergus et al. 2016) consists of two-way intersecting routes on an 18×18 grids with four arrival points, and cars (agents) with one-grid limited vision, requiring communication to avoid collisions. We set the maximum number of cars in the environment to 20 and the maximum time steps to 50. New cars get added to the environment with a probability of 0.05. Success indicates that there are no collisions within an episode. The action space for each car is gas and break, and the reward consists of a step cost of 0.01 and a collision penalty of -10 .

We set $\beta_1 = 0.01, \beta_2 = 0.005$ for the two GAT layers respectively. For the agents around different arrival points, NTNNR can encourage them to obtain diverse message aggregation strategies. The message aggregation strategies of agents are constantly changing at different time steps in an



(a) A test frame in Traffic Junction. (b) Message aggregation strategies of agent 5 and 9

Figure 6: The visualization of a time step in the Traffic Junction scenario (hard mode).

episode. In order to analyze the impact of NTNNR on strategies, we visualize the message aggregation strategies of two agents at one time step evaluated with the well-trained policy in Figure 6. It is observed that agent 5 is at the upper left arrival point, while agent 9 is at the downright arrival point. Intuitively, even though they can communicate, the messages are useless to each other. The distributions of representative attention scores for message aggregation are shown in Figure 6(b). With NTNNR, agents 5 and 9 obtained diverse message aggregation strategies. This makes communication more efficient in the multi-agent system, avoiding unnecessary interference between unrelated agents.

Figure 7 shows the success rate per epoch attained by various message aggregation methods. GAT with NTNNR is competitive when compared to other methods. Our method not only provides a higher success rate but also can be more sample efficient. We suppose the phenomenon attributes to efficient communication brought by NTNNR, where agents find optimal coordination faster.

The effect of the scaling hyper-parameters β_l : To further analyze the effect of the regularization weights of

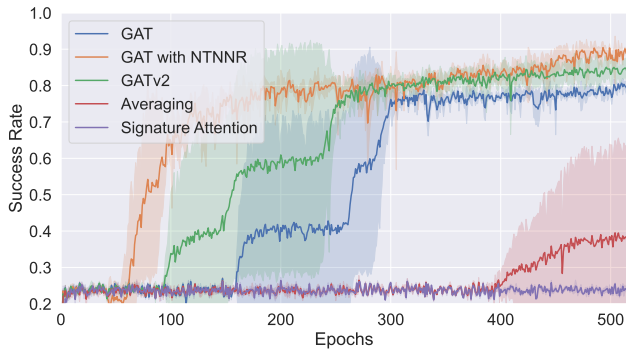


Figure 7: Success rates of various message aggregation approach in the traffic junction scenario.

NTNNR, we evaluate the performance with different β_1 and β_2 . We record the corresponding success rates in Table 1. From the first column and row, we can observe that utilizing NTNNR can significantly improve performance.

$\beta_1 \backslash \beta_2$	0	0.001	0.005	0.01	0.02
0	0.77	0.78	0.86	0.83	0.84
0.005	0.87	0.84	0.82	0.81	0.89
0.01	0.88	0.81	0.91	0.85	0.83
0.02	0.76	0.85	0.80	0.85	0.87

Table 1: Success rates with different scaling hyper-parameters of NTNNR.

We observe that the best performance achieved when $\beta_1 = 0.01, \beta_2 = 0.005$. Considering the first GAT layer contains four attention heads while the second layer only contains one, we recommend a larger regularization weight when the third dimension of the adjacency tensor is larger. Experiments in the predator-prey scenario also support this conclusion. Besides, it is observed that setting β_1 between 0.005 and 0.01 or setting β_2 between 0.001 and 0.02 can guarantee the performance improvement, showing acceptable robustness to the scaling hyper-parameters.

StarCraft II

In this section, we evaluate our method on SMAC, a more complex benchmark. We want to show that NTNNR is general and easily integrated with existing graph-attention Comm-MARL methods, using the plug-and-play manner. We choose two state-of-the-art methods, GA-Comm and DICG-CE-LSTM, and apply NTNNR to them. The scaling hyper-parameter is set to 0.05 and 0.005, respectively.

The average evaluation win rates are shown in Figure 8. Methods augmented by NTNNR achieve outstanding performance compared with their vanilla counterparts. We suppose the improvement is due to the emergent behaviors brought by the diverse message aggregation strategies. To better explain why our regularizer performs well, we further visualize the final trained strategies in Figure 9. In this 3s5z map, three

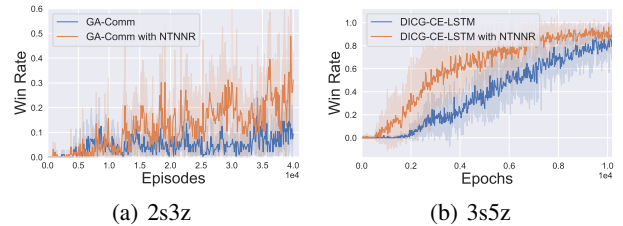


Figure 8: Performance comparison of methods with NTNNR over their vanilla counterparts in SMAC maps.



Figure 9: Visualization of the final strategies trained by DICG-CE-LSTM with NTNNR in the 3s5z map. We control the red stalkers and zealots.

parameter-sharing zealots with similar observations can select diverse actions and finally surround the enemy stalkers to attack. The sophisticated coordination reflects the effectiveness of diverse message aggregation in Comm-MARL.

Conclusion

In this paper, we present that the diversity of message aggregation in graph-attention Comm-MARL methods could be measured by the normalized tensor rank, and further define the corresponding nuclear norm to quantify the diversity. Then we propose a plug-and-play regularizer named NTNNR, to actively enrich the diversity of message aggregation. Experiments show that GAT with NTNNR can provide superior performance and better training efficiency compared to existing message aggregation methods. Furthermore, NTNNR can be easily applied to existing graph-attention Comm-MARL methods and improve their performance.

Assuredly, our method has some limitations. In some multi-agent coordination tasks with core agents, overly diverse message aggregation may be unreasonable. Therefore, NTNNR may not achieve significant performance improvements in these cases. In future work, we plan to quantify the diversity upper bound for multi-agent systems.

References

- Agarwal, A.; Kumar, S.; and Sycara, K. 2020. Learning transferable cooperative behavior in multi-agent teams. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*.
- Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; and Mordatch, I. 2019. Emergent Tool Use From Multi-Agent Autocurricula. In *International Conference on Learning Representations*.
- Balduzzi, D.; Garnelo, M.; Bachrach, Y.; Czarnecki, W.; Perolat, J.; Jaderberg, M.; and Graepel, T. 2019. Open-ended learning in symmetric zero-sum games. In *International Conference on Machine Learning*, 434–443. PMLR.
- Brody, S.; Alon, U.; and Yahav, E. 2022. How attentive are graph attention networks? *International Conference on Learning Representations*.
- Candès, E. J.; and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6): 717–772.
- Chenghao, L.; Wang, T.; Wu, C.; Zhao, Q.; Yang, J.; and Zhang, C. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34.
- Cully, A.; Clune, J.; Tarapore, D.; and Mouret, J.-B. 2015. Robots that can adapt like animals. *Nature*, 521(7553): 503–507.
- Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabbat, M.; and Pineau, J. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*, 1538–1546. PMLR.
- DKipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.
- Du, Y.; Liu, B.; Moens, V.; Liu, Z.; Ren, Z.; Wang, J.; Chen, X.; and Zhang, H. 2021. Learning correlated communication topology in multi-agent reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 456–464.
- Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29.
- Fu, W.; Yu, C.; Xu, Z.; Yang, J.; and Wu, Y. 2022. Revisiting Some Common Practices in Cooperative Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*.
- Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.
- Jiang, J.; and Lu, Z. 2021. The emergence of individuality. In *International Conference on Machine Learning*, 4992–5001. PMLR.
- Kim, D.; Moon, S.; Hostallero, D.; Kang, W. J.; Lee, T.; Son, K.; and Yi, Y. 2019. Learning to schedule communication in multi-agent reinforcement learning. *International Conference on Learning Representations*.
- Kim, W.; Park, J.; and Sung, Y. 2020. Communication in multi-agent reinforcement learning: Intention sharing. In *International Conference on Learning Representations*.
- Kolda, T. G.; and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500.
- Lee, J. B.; Rossi, R. A.; Kim, S.; Ahmed, N. K.; and Koh, E. 2019. Attention models in graphs: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(6): 1–25.
- Li, S.; Gupta, J. K.; Morales, P.; Allen, R.; and Kochenderfer, M. J. 2021. Deep Implicit Coordination Graphs for Multi-agent Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 764–772.
- Liu, Y.; Wang, W.; Hu, Y.; Hao, J.; Chen, X.; and Gao, Y. 2020. Multi-agent game abstraction via graph attention neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7211–7218.
- Lowe, R.; Foerster, J.; Boureau, Y.-L.; Pineau, J.; and Dauphin, Y. 2019. On the Pitfalls of Measuring Emergent Communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 693–701.
- Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.
- Lupu, A.; Cui, B.; Hu, H.; and Foerster, J. 2021. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, 7204–7213. PMLR.
- Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven: Multi-agent variational exploration. *Advances in Neural Information Processing Systems*, 32.
- Niu, Y.; Paleja, R.; and Gombolay, M. 2021. Multi-agent graph-attention communication and teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 964–973.
- Parker-Holder, J.; Pacchiano, A.; Choromanski, K. M.; and Roberts, S. J. 2020. Effective diversity in population based reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 18050–18062.
- Peng, P.; Wen, Y.; Yang, Y.; Yuan, Q.; Tang, Z.; Long, H.; and Wang, J. 2017. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*.
- Pugh, J. K.; Soros, L. B.; and Stanley, K. O. 2016. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 40.
- Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.
- Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Singh, A.; Jain, T.; and Sukhbaatar, S. 2019. Individualized controlled continuous communication model for multiagent cooperative and competitive tasks. In *International conference on learning representations*.
- Sukhbaatar, S.; Fergus, R.; et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29.
- Tang, Z.; Yu, C.; Chen, B.; Xu, H.; Wang, X.; Fang, F.; Du, S. S.; Wang, Y.; and Wu, Y. 2020. Discovering Diverse Multi-Agent Strategic Behavior via Reward Randomization. In *International Conference on Learning Representations*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2018. Graph attention networks. *International Conference on Learning Representations*.
- Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.
- Wang, T.; Gupta, T.; Peng, B.; Mahajan, A.; Whiteson, S.; and Zhang, C. 2021. RODE: learning roles to decompose multi-agent tasks. In *Proceedings of the International Conference on Learning Representations*. OpenReview.
- Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2020. Influence-based multi-agent exploration. *International Conference on Learning Representations*.
- Whiteson, S.; Samvelyan, M.; Rashid, T.; De Witt, C.; Farquhar, G.; Nardelli, N.; Rudner, T.; Hung, C.; Torr, P.; and Foerster, J. 2019. The StarCraft multi-agent challenge. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2186–2188.
- Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3): 229–256.
- Yuan, L.; Wang, J.; Zhang, F.; Wang, C.; Zhang, Z.; Yu, Y.; and Zhang, C. 2022. Multi-Agent Incentive Communication via Decentralized Teammate Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Zahavy, T.; O’Donoghue, B.; Barreto, A.; Flennerhag, S.; Mnih, V.; and Singh, S. 2021. Discovering Diverse Nearly Optimal Policies with Successor Features. In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*.
- Zhai, Y.; Ding, B.; Liu, X.; Jia, H.; Zhao, Y.; and Luo, J. 2021. Decentralized multi-robot collision avoidance in complex scenarios with selective communication. *IEEE Robotics and Automation Letters*, 6(4): 8379–8386.
- Zhou, M.; Luo, J.; Vilella, J.; Yang, Y.; Rusu, D.; Miao, J.; Zhang, W.; Alban, M.; Fadakar, I.; Chen, Z.; et al. 2020. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. *Conference on Robot Learning*.
- Zhou, Z.; Fu, W.; Zhang, B.; and Wu, Y. 2021. Continuously Discovering Novel Strategies via Reward-Switching Policy Optimization. In *International Conference on Learning Representations*.
- Zhu, C.; Dastani, M.; and Wang, S. 2022. A Survey of Multi-Agent Reinforcement Learning with Communication. *arXiv preprint arXiv:2203.08975*.

Analysis of Diversity Maintained by NTNNR

In this paper, we define the Normalized Tensor Rank and the Normalized Tensor Nuclear Norm (NTNN). Enlarging NTNN can enrich the diversity of tensor \mathcal{A} from both the frontal slice view and the mode-3 fiber view. For better comprehension of the effect, we will take a toy example. Suppose there are two agents in the multi-agent system, and the multi-head attention mechanism contains two heads, i.e., $N = 2, K = 2$. In this case, $\mathcal{A} \in \mathbb{R}_+^{2 \times 2 \times 2}$ could be expressed as:

$$\mathbf{A}^{(0)} = \begin{bmatrix} x_0 & 1 - x_0 \\ y_0 & 1 - y_0 \end{bmatrix}, \quad \mathbf{A}^{(1)} = \begin{bmatrix} x_1 & 1 - x_1 \\ y_1 & 1 - y_1 \end{bmatrix}. \quad (13)$$

where x_0, y_0, x_1 and y_1 are variables.

Applying normalization to \mathcal{A} along the 3-rd way and transforming the tensor to the block diagonal matrix form, we have:

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{x}_0 & 1 - \hat{x}_0 & & & & \\ \hat{y}_0 & 1 - \hat{y}_0 & & & & \\ & & \hat{x}_1 & 1 - \hat{x}_1 & & \\ & & \hat{y}_1 & 1 - \hat{y}_1 & & \end{bmatrix}. \quad (14)$$

To obtain the singular values, we calculate the eigenvalues of $\hat{\mathbf{A}}\hat{\mathbf{A}}^T$ as follows:

$$|\hat{\mathbf{A}}\hat{\mathbf{A}}^T - \lambda\hat{\mathbf{I}}| = 0. \quad (15)$$

We denote the four singular values as $\sigma_0, \sigma_1, \sigma_2$, and σ_3 respectively. With the properties of block diagonal matrix, we solve Equation 15 and have:

$$\begin{cases} \sigma_0^2 + \sigma_1^2 = 2(\hat{x}_0^2 - \hat{x}_0 + \hat{y}_0^2 - \hat{y}_0 + 1), \\ \sigma_0^2 \times \sigma_1^2 = (\hat{y}_0 - \hat{x}_0)^2, \\ \sigma_2^2 + \sigma_3^2 = 2(\hat{x}_1^2 - \hat{x}_1 + \hat{y}_1^2 - \hat{y}_1 + 1), \\ \sigma_2^2 \times \sigma_3^2 = (\hat{y}_1 - \hat{x}_1)^2, \\ \hat{x}_0 + \hat{x}_1 = 1, \\ \hat{y}_0 + \hat{y}_1 = 1. \end{cases} \quad (16)$$

The normalized tensor nuclear norm is the sum of the singular values of $\hat{\mathbf{A}}$, which is calculated as follows:

$$\begin{aligned} \|\mathcal{A}\|_* &= \frac{1}{K} \|\hat{\mathbf{A}}\|_* = \frac{1}{K} (\sigma_0 + \sigma_1 + \sigma_2 + \sigma_3) \\ &= \frac{1}{K} (\sqrt{(\sigma_0 + \sigma_1)^2} + \sqrt{(\sigma_2 + \sigma_3)^2}) \\ &= \sqrt{\hat{x}_0 + (1 - \hat{x}_0)^2 + \hat{y}_0 + (1 - \hat{y}_0)^2 + 2|\hat{y}_0 - \hat{x}_0|} \\ \text{s.t. } &\hat{x}_0 + \hat{x}_1 = 1, \quad \hat{y}_0 + \hat{y}_1 = 1. \end{aligned} \quad (17)$$

From Equation 17, $\|\mathcal{A}\|_*$ would reach the maximum solution when:

$$\begin{aligned} \mathbf{A}^{(0)} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \mathbf{A}^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{or} \\ \mathbf{A}^{(0)} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mathbf{A}^{(1)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \end{aligned} \quad (18)$$

Therefore, we demonstrate that NTNNR tries to maintain the diversity of the adjacency tensor in both the frontal slice view and the mode-3 fiber view.

Code

Our code will be released publicly to enhance the reproducibility. We use the following open-source repositories for baselines:

- MAGIC code: <https://github.com/CORE-Robotics-Lab/MAGIC>
- GA-Comm code: <https://github.com/starry-sky6688/MARL-Algorithms>
- DICG-CE-LSTM code: <https://github.com/sisl/DICG>

Implementation Details

Our implementation is on a desktop machine with one Intel i9-12900K CPU and one NVIDIA RTX3080 GPU. All the methods in the same scenario are run for the same number of total environment steps (or episodes) and the same number of iterations.

Table 2: Hyper-parameters in the the Predator-Prey and Traffic Junction Scenarios.

Parameter	Predator-Prey	Traffic Junction
Number of processes	16	16
Epoch size	10	10
Hidden units for LSTM encoder	128	128
Learning rate	0.001	0.001
Number of attention heads (the first GAT layer)	2	4
Number of attention heads (the second GAT layer)	1	1
Hidden units of each attention head	32	32
Scaling hyper-parameter β_1	0.2	0.01
Scaling hyper-parameter β_2	0.005	0.005

In the the Predator-Prey and Traffic Junction Scenarios, we distribute the training over 16 threads and each thread runs batch learning with a batch size of 500. The threads share the parameters θ of the policy network and update synchronously. We use RMSProp as the optimizer. Table 2 shows the details of our neural network and other hyper-parameters.

In the StarCraft II scenario, we follow the original implementations of the selected Comm-MARL methods. We adopt their network structures and hyper-parameter settings, except changing the number of attention heads to 4 and the hidden units of each attention head to 32 for GA-Comm.

Training Algorithms

Most of existing Comm-MARL methods utilize policy gradient methods with parameter sharing. Then the joint policy can be factorized as:

$$\pi_{\theta}(\mathbf{u}^t | \mathbf{o}^t, \mathbf{m}^t) = \prod_{i=0}^N \pi_{\theta}(u_i^t | o_i^t, m_{j \neq i}^t) \quad (19)$$

Augmented with communication, the environment and other agents' policies can be seen as stable for a agent.

Agents simultaneously select actions according to local observations and communication messages. So various single-agent policy-gradient methods can be utilized as the training algorithms for Comm-MARL.

All methods used in the predator-prey and traffic junction scenarios adopt the REINFORCE (Williams 1992) with baseline as training algorithms. In this case, Equation 9 of the manuscript can be written as:

$$\nabla_{\theta} L_{RL}(\theta) = \mathbb{E}_{i,t}[\nabla_{\theta} \log \pi_{\theta}(u_i^t | o_i^t, m_{j \neq i}^t) \psi_i^t], \quad (20)$$

where $\psi_i^t = r_i^t - V(o_i^t, m_{j \neq i}^t)$ is the advantage function. $V(o_i^t, m_{j \neq i}^t)$ is the value function.

GA-Comm adopts the REINFORCE algorithm. The Equation 9 of the manuscript can be written as:

$$\nabla_{\theta} L_{RL}(\theta) = \mathbb{E}_{i,t}[\nabla_{\theta} \log \pi_{\theta}(u_i^t | o_i^t, m_{j \neq i}^t) \sum_{t'=t}^T \gamma^{t'-t} r_i^{t'}], \quad (21)$$

where T is the maximum time steps for agents to interact with the environment.

DICG-CE-LSTM adopts the clipped PPO (Schulman et al. 2017) algorithm. Then the Equation 9 of the manuscript has the following form:

$$\begin{aligned} \nabla_{\theta} L_{RL}(\theta) = \mathbb{E}_{i,t}[\nabla_{\theta} \min & \left(\frac{\pi_{\theta}(u_i^t | o_i^t, m_{j \neq i}^t)}{\pi_{old}(u_i^t | o_i^t, m_{j \neq i}^t)} \psi_i^t, \right. \\ & \left. clip\left(\frac{\pi_{\theta}(u_i^t | o_i^t, m_{j \neq i}^t)}{\pi_{old}(u_i^t | o_i^t, m_{j \neq i}^t)}, 1 - \epsilon, 1 + \epsilon\right) \psi_i^t \right)], \end{aligned} \quad (22)$$

where ϵ is the hyper-parameter for clip, and π_{old} is the policy of the last update iteration.