计算机科学
COMPUTER SCIENCE

基于图注意力的分组多智能体强化学习方法

朱士昊, 彭可兴, 马廷淮

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700107

基于图注意力的分组多智能体强化学习方法

朱士昊[1] 彭可兴[2] 马廷淮[1,3]
1 南京信息工程大学软件学院 南京 210044
2 南京信息工程大学计算机学院 南京 210044
3 江苏海洋大学计算机工程学院 江苏 连云港 222005
(zhushihaosz@126.com)

摘 要 目前,多智能体强化学习在各类合作任务中被广泛应用。但在真实环境中,智能体通常只能获取部分观测值,导致合作策略的探索效率低下。此外,智能体共享奖励值,导致其难以准确衡量个体贡献。针对这些问题,提出一种基于图注意力的分组多智能体强化学习框架,其有效提高了合作效率并改善了个体贡献的衡量。首先,构建图结构的多智能体系统,通过图注意力网络学习个体与邻居的关系以进行信息共享,扩大智能体个体的感受野,从而缓解部分可观测的限制并有效衡量个体贡献。其次,设计了动作参考模块,为个体动作选择提供联合动作参考信息,使智能体在探索时更高效、多样。在两个不同规模的多智能体控制场景下,所提方法相比基线方法展现出显著的优势;同时,消融实验证明了图注意力分组方法和通信设置的有效性。

关键词:多智能体强化学习;图注意力网络;集中训练分散执行;多智能体协作;多智能体通信
中图分类号 TP391

Graph Attention-based Grouped Multi-agent Reinforcement Learning Method

ZHU Shihao[1],PENG Kexing[2] and MA Tinghuai[1,3]
1 School of Software,Nanjing University of Information Science and Technology,Nanjing 210044,China
2 School of Computer Science,Nanjing University of Information Science and Technology,Nanjing 210044,China
3 School of Computer Engineering,Jiangsu Ocean University,Lianyungang,Jiangsu 222005,China

Abstract Currently,multi-agent reinforcement learning is widely applied in various cooperative tasks. In real environments,agents always have access to only partial observations,leading to inefficient exploration of cooperative strategies. Moreover,sharing reward values among agents makes it challenging to accurately assess individual contributions. To address these issues,a novel graph attention-based grouped multi-agent reinforcement learning framework is proposed,which improves cooperation efficiency and enhances the evaluation of individual contributions. Firstly,a multi-agent system with graph structure is constructed,which learning relationships among the individual agents and their neighbors for sharing information. This approach expands individual agents' perceptual fields to mitigate constraints from partial observability and assess individual contributions. Secondly,an action reference module is designed to provide joint action reference information for individual action selection,enabling agents to explore more efficiently and diversely. Experimental results in two different scales of multi-agent control scenarios demonstrate significant advantages over baseline methods. Detailed ablation studies further verify the effectiveness of the graph attention grouping approach and communication settings.

Keywords Multi-agent reinforcement learning,Graph attention network,Centralized training decentralized execution,Multi-agent cooperation,Multi-agent communication

1 引言

神经网络与深度学习的飞速发展使多智能体系统能够自主协调机器高效完成合作任务,在交通、能源、通讯、军事领域得到广泛应用[1-3]。强化学习(Reinforcement Learning,RL)通过智能体与环境交互的方式,引导智能体探索最大化奖励的策略。现实环境中通常由多个智能体共同协作完成目标,因此多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)应运而生。

在真实环境中,诸多应用要求智能体根据部分环境信息

到稿日期:2024-07-16 返修日期:2024-11-04
基金项目:国家自然科学基金(62372243,62102187)
This work was supported by the National Natural Science Foundation of China(62372243,62102187).
通信作者:马廷淮(thma@nuist.edu.cn)

和其他智能体的部分信息进行决策[4]，如纸牌游戏中，玩家无法查看其他玩家的牌面信息，这给多智能体强化学习带来了挑战。此类问题被称为部分可观测问题。在部分可观测的场景下，智能体仅能直接获取自身信息及部分智能体的历史行为，难以推测其他智能体的潜在策略，加剧了多智能体系统的非平稳性[5]。一些研究通过共享环境奖励来补充全局信息，虽然这类方法有助于智能体理解其他智能体策略，但共享奖励值会使得智能体动作探索不足以及策略趋同，导致训练陷入次优[6]。为平衡个体与联合策略之间的关系，集中训练分散执行框架被广泛应用。该框架能够提供联合动作指导智能体个体的动作选择，其中联合动作如何分解为个体动作是核心问题之一。多智能体在合作任务时，特别是在接收联合动作值的设定下，无法准确衡量个体智能体的贡献[7]。

为此，本文提出了一种基于图注意力的分组强化学习方法（Graph Attention-Based Grouped Multi-Agent Reinforcement Learning Method，GAG），以图结构的方式建模多智能体系统，通过图注意力机制促使智能体自发分组通信，对联合和个体动作分别进行约束，从而准确衡量个体贡献。通过设计动作参考网络，引入动作参考，使个体智能体进行多样动作选择。此外，设计了值评估网络，通过保证个体与联合之间的一致性来减少动作参考网络带来的次优干扰。

本文的主要贡献如下：

1）提出一种基于图注意力的分组多智能体强化学习框架，采用图结构建模智能体与邻居的关系，以扩大智能体感受野，补充局限观测值；集合图注意力机制更新节点权重，衡量个体贡献。

2）提出动作参考网络，以捕捉智能体间的潜在关系，提供多个联合动作选择，给每个智能体提供多样化的探索空间。

3）基于 SUMO 交通信号控制、星际争霸 II 环境进行实验，证明了本文方法较主流基线算法有显著优势。通过充分的消融实验，证明了本文方法的有效性。

2 相关工作

2.1 基于通信的强化学习

智能体之间能够通过信息交换的方式来提高共享效率。在早期的显式通信研究中，为了获得其他智能体的状态，所有智能体共享一个额外的网络，在这个网络中有条件地传输消息，如 COMA（Counterfactual Multi-Agent）算法[8]。这些方法需要不断更新智能体之间的动态关系，导致训练不收敛以及共享信息中模式提取困难的问题。

学者们在通信关系建模上进行了大量研究。IS（Intention Sharing）算法[9]利用想象的轨迹生成意图信息与智能体进行通信。该方法基于个体策略进行了详细的分析，假定所有智能体的信息都是保密的，这与现实中大规模环境的要求不一致。最近，在图通信方面，Liu 等[10]提出 G2Anet（Game Abstraction Mechanism Based on Two-stage Attention Network）算法，该网络基于完全图捕获智能体之间的关系。MAGIC（Multi-Agent Graph-attention Communication）算法[11]是一个图注意力通信网络，使用动态图结构来确定何时通信以及向谁发送消息。

受 IS 算法的启发，本文方法通过引入图注意力网络来关注智能体的邻域信息，从而做出更好的决策。

2.2 基于值分解的强化学习

基于值分解的 MARL 算法的核心是学习将团队价值函数分解为单个智能体价值函数。

现有研究大多采用集中式训练和分散式执行（Centralized Training with Decentralized Execution，CTDE）框架。QMIX 算法[12]通过端到端集中式训练获得去中心化策略，设计了一个混合网络，在满足单调性约束的前提下，近似于联合动作值和单个动作值之间的关系。Son 等设计的 QTRAN 算法[13]通过一种新的价值函数来代替具有相同最优动作的联合动作-价值函数。

许多研究遵循上述方法的约束，对框架进行改进和重新设计，以解决复杂环境中的各种问题。GraphMIX 算法[14]对智能体之间的图关系建模，并使用图神经网络分解联合状态-动作值函数。ROMA（Role-Oriented Multi-Agent Reinforcement Learning）算法[15]允许学习角色划分，具有相似角色的智能体共享它们在子任务上的合作学习。DOP（Multi-agent Decomposed Policy Gradient Method）算法[16]将价值函数分解的思想引入到多智能体行为者批评框架中。该方法可以实现稳定、高效的多智能体离线策略学习。LIIR（Learning Individual Intrinsic Reward）算法[17]在每个时间步对个体智能体进行建模，并分配特定的奖励函数。

基于上述工作，本文设计一种方法，使得在集中式训练中，智能体可以访问其他智能体的信息，并平滑地估计联合动作值；在分散式执行中，智能体根据个体的动作、观测值独立地做出决策。

3 基于图注意力的分组通信方法

本文针对多智能体协作任务中的个体贡献分配问题，利用图注意力机制构建智能体间的显式关系图，并根据关系权重进行分组通信。同时，利用智能体间的隐式关系来提供更多的联合动作，丰富动作空间探索，提高模型探索和利用的效率。

如图 1 所示，本文方法由提供参考联合动作的动作参考网络和构建智能体关系的图注意力网络构成，且遵循 CTDE 框架，先集中训练后分散执行。下文中以各模块的先后顺序依次进行阐述。

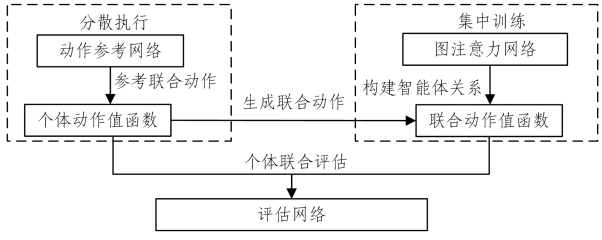


图 1 基于图注意力的分组多智能体强化学习框架

Fig. 1 Framework of graph attention-based grouped multi-agent reinforcement learning

在分散执行阶段，各智能体能够获得当前部分观测值以及其他智能体的历史动作，并进行个体动作选择。其中，动作

参考网络能够实时根据当前状态给出联合动作参考。在集中训练过程中，根据个体动作选择生成图结构的智能体表征；通过图注意力机制更新智能体之间的关系，准确衡量个体贡献。此外，引入值评估网络，以减少联合动作函数和个体动作函数的不一致性。

3.1 基于动作参考的分散执行

在部分可观测的设定下，各智能体通过有限通信来理解其他智能体的动作趋势。在每个时间步 t，第 i 个智能体根据当前自身的观测值 o_i^t 和前一步其他智能体 $-i$ 的动作 a_{-i}^{t-1}，计算个体动作价值 $q_i(o_i^t, a_i^t, a_{-i}^{t-1})$，实现分散执行。其中，个体动作由策略 $\pi_i(a|\tau_n^t)$ 选择。每个智能体的个体动作选择网络和内部结构如图 2 所示。

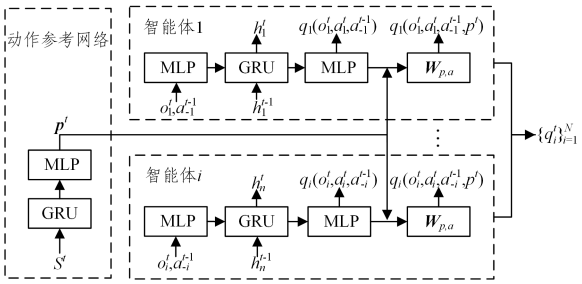


图 2 分散执行结构

Fig. 2 Architecture of decentralized execution

为了避免单调性约束导致训练陷入局部次优，设计了一个用强化方法训练的动作参考网络 P(S)，该网络能根据当前状态值学习联合动作选择。该模块基于 MAVEN(Multi-Agent Variational Exploration)方法[18]的 Z 空间，可以看作一个探索联合作用的空间，能通过任意策略学习方法进行训练。如图 2 左侧“动作参考网络”所示，动作参考网络输入当前状态值 S^t，通过神经网络学习参数 φ 来探索分层策略 $\pi_p(\cdot|S;\varphi)$，实现从 p 空间到不同动作模式的映射 $z\sim f_\varphi(S)$，得到隐向量 $\boldsymbol{p}^t$，即联合动作参考。动作参考网络的分层策略目标如式(1)所示：

$$J_P(\varphi)=\int R(\tau_a|p)\,p_\varphi(p|S)\rho(S)\mathrm{d}z\mathrm{d}s \tag{1}$$

其中，可学习神经网络由多层感知机和时序网络 GRU 构成，隐藏状态仅用于学习动作选择。本文采用 ε-贪心策略，以 ε 概率选择随机动作进行探索，以 1−ε 概率选择最优动作来进行探索，以此实现探索与利用的平衡。而完全贪心策略始终选择当前最优动作，长期来看存在探索不足的问题。至此，得到了联合动作参考值 $\boldsymbol{p}^t$，对应不同的动作组成动作参考矩阵 $\boldsymbol{W}_{p,a}$。每个智能体个体动作 $q_i(o_i^t, a_i^t, a_{-i}^{t-1})$ 均考虑动作参考信息，得到最终个体动作 $q_i(o_i^t, a_i^t, a_{-i}^{t-1}, p^t)$。在大规模网络 MARL 任务中，可以在离散执行中屏蔽 $\boldsymbol{W}_{p,a}$，仅在集中训练中使用，以提高部分训练效率。

3.2 基于图注意力的集中训练

在个体动作值函数的基础上，以图注意力网络构建智能体间的显式关系。联合状态-动作函数训练模块如图 3 所示。以拓扑图构建多智能体系统，表示为 $G=(V,E)$，其中节点由各智能体的个体动作值函数组成，表示为 $V=\{q_i^t\}_{i=1}^N$，边是智能体间的关系权重，属于 N^2 图的子集，满足 $E\subseteq V\times V$。至此，构建了一个完全无向图，并且根据不同的环境可以设定为有向图。

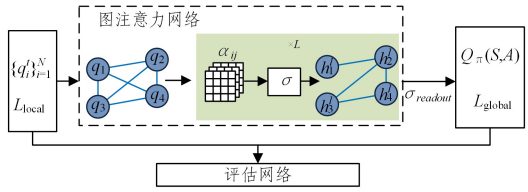


图 3 集中训练结构

Fig. 3 Architecture of centralized training

为了探索联合策略与个体策略之间的关系，将联合动作值函数 $Q_\pi(S^t,A^t)$ 分解为个体动作值函数 $\{q_i(\tau_i^t,a_i^t)\}_{i=1}^N$。基于价值的 CTDE 的重要前提是确保个体和联合贪婪动作选择的一致性。本文框架需要保证每个智能体尽可能达到个体动作函数最优值，从而使得联合动作值函数最优拟合，这被称为个体-全局最大化原则(Individual-Global Maximum, IGM)[12,19]，具体分解定义如式(2)所示：

$$\arg\max_A Q_\pi(S,A)=\begin{pmatrix}\arg\max_{a_1} q_1(\tau_1,a_1)\\ \vdots\\ \arg\max_{a_N} q_N(\tau_1,a_N)\end{pmatrix} \tag{2}$$

为了实现高效联合探索，在训练学习过程中加入全局状态信息进行辅助。本部分将图注意力机制与联合动作值函数相结合，如图 3 所示。

图注意力网络以图结构信息为输入，其中节点 $i\in V$ 表示个体智能体，节点特征表示对应智能体的个体动作值函数。节点间的边决定哪些智能体可以相互通信，边特征表示与节点智能体相关的特征。为了有效训练，节点特征用先前计算的个体动作值函数进行初始化，如式(3)所示：

$$h_i^0=q_i(o_i,a_i,a_{-i},p) \tag{3}$$

在复杂任务中，智能体 i 与相邻智能体 $\{j\in N_i\}$ 之间的关系存在差异，使得智能体 i 难以学习到一个稳定的策略。因此，需要智能体对其邻居分配不同的权重来计算相邻节点的特征。具体地，利用节点的隐藏状态计算智能体 i、智能体 j 之间的注意力系数 α_{ij}，具体过程如式(4)、式(5)所示：

$$e_{ij}=\boldsymbol{W}^k([f(h_i),f(h_j)]) \tag{4}$$

$$\alpha_{ij}=\frac{\exp(LeakyReLU(e_{ij}))}{\sum_{j\in N_i}\exp(LeakyReLU(e_{ij}))} \tag{5}$$

其中，W^k 动作是单层前馈神经网络，$f:R^N\to R^{N'}$ 是一个升维的线性变化，LeakyReLU 是非线性激活函数，N_i 是智能体 i 的邻居节点集。

在得到节点的注意力系数后，将一阶邻居节点特征聚合到中心节点，得到 1 跳节点特征。经过 L 层隐层计算后，L 跳节点特征被聚合到 i 节点，计算过程如式(6)所示：

$$h_i^l=\sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k\boldsymbol{W}^k h_j^{l-1}\right) \tag{6}$$

其中，σ 是单调递增的可学习函数，$\boldsymbol{W}^k$ 是对应输入的线性变化矩阵，α_{ij}^k 是第 k 个注意力头的系数。取 K 个注意力头组成多头注意力机制，遵循原始图注意力网络。将 K 设置为 3，得到的平均值能避免单头注意力机制的不稳定性。

经过图注意力网络更新节点特征后，智能体完成自发分组的通信，获取到了各自分组中的通信信息。最后，根据加权

后的个体动作表征计算联合动作值函数，取各节点最后一层的输出 h_i^L。联合动作值函数如式(7)所示：

$$Q_\pi(S,A)=\sigma_{\text{readout}}(\{h_i^L\}_{i\in V}) \tag{7}$$

其中，$\sigma_{\text{readout}}:\mathbb{R}^{F_L}\times\cdots\times\mathbb{R}^{F_L}\rightarrow\mathbb{R}^{F_L}$ 是从整个图的关系、节点异构特征到联合动作值函数体间的映射关系。

基于上述流程，整体框架的动作值函数更新过程如式(8)、式(9)所示：

$$Q_\pi(R,S';\theta^-)=r+\gamma Q_\pi'(S',\bar{A}';\theta^-) \tag{8}$$

$$\bar{A}'=[\arg\max_{a_i} q_i(o_i',a_i,a_{-i},p,\theta^-)]_{i=1}^N \tag{9}$$

为了更好地衡量个体在整体中的贡献程度，本文参考 GraphMIX[14] 算法，从联合和个体角度分别对所提框架进行训练，损失函数如式(10)、式(11)所示：

$$L_{\text{global}}=\sum_{i\in N}[(Q_\pi(R,S';\theta^-)-Q_\pi(S;\theta))^2]_i \tag{10}$$

$$L_{\text{local}}=\sum_{i\in N}[(R+\gamma\max_{a_i'}\{q_n'\}_{n=1}^N-\{q_n'\}_{n=1}^N)^2]_i \tag{11}$$

其中，各智能体的动作值函数用于最小化局部损失 L_{local}，联合状态动作值用于最小化全局损失 L_{global}，二者有效减小了个体贡献偏差。

为了确保算法的收敛性，本文方法满足 IGM 原则的充分必要条件，如式(12)、式(13)所示：

$$\sum_{i=1}^N q_i(\tau_i,a_i)-Q_\pi(\tau,A)+B_{\pi_b}=\begin{cases}0, & a=\bar{a}\\ \geqslant 0, & a\neq\bar{a}\end{cases} \tag{12}$$

$$B_{\pi_b}=\max_a Q_\pi(\tau,A)-\sum_{i=1}^N q_i(\tau_i,\bar{a}_i) \tag{13}$$

其中，$\bar{a}_i$ 为最优个体动作值。如图 3 中评估网络所示，B_{π_b} 是得到的最优联合状态值函数，旨在减少所有个体动作值函数与联合动作值函数之间的不一致性，同时减少动作参考网络可能导致的次优收敛。

由于部分可观测的设定，满足所有状态下的 IGM 原则存在困难。在本文方法中，给定任意 τ，总能找到一个对应的 a_i 满足式(13)，使得任务可分解。该评估网络通过更新参数 θ 来学习最优联合状态值函数，损失函数如式(14)所示：

$$L_B=(\min(\{q_i\}_{i=1}^N-\hat{Q}_\pi(R,S;\theta)+B(\tau),0))^2 \tag{14}$$

其中，$\hat{Q}_\pi$ 被固定为 Q_π。

为了减小动作参考网络陷入次优的可能性，设计损失函数，如式(15)所示：

$$L_P=-\frac{1}{M}\sum_\tau r\log p(\tau) \tag{15}$$

其中，τ 是智能体的历史动作观察，M 是采样大小。

本文方法的总损失函数如式(16)所示：

$$L=L_{\text{global}}+\lambda_{\text{local}}L_{\text{local}}+L_B+L_P \tag{16}$$

算法 1 展示了上述框架的伪代码流程。

算法 1 图注意力强化学习算法

输入：神经网络超参数、智能体局部观测值

输出：动作参考网络参数 φ、策略网络参数 θ

1. 初始化：经验回放池，网络参数
2. for 回合开始 do
3. 智能体观察初始状态和观测值
4. for t<T do
5. ε 概率，随机选择个体动作，否则，选择最优个体动作 $a_i^t=\arg\max_{a_i^t} q_i(o_i^t,a_i^t,a_{-i}^{t-1})$
6. 存储经验$(\tau^t,A^t,R^t,\tau^{t+1})$
7. 随机批次采样经验(τ,A,R,τ')
8. 根据动作参考网络计算动作参考 p^t
9. 计算个体动作值函数 $q_i(o_i^t,a_i^t,a_{-i}^{t-1},p^t)$
10. 按式(7)，通过图注意力网络计算联合动作值函数 $Q_\pi(S,A)$
11. 按式(16)，计算损失函数，更新参数 θ
12. 每 I 次，目标网络参数 θ^- 与 θ 同步
13. End for
14. End for

4 实验与结果分析

4.1 基线方法

为了证明本文方法的有效性，将其与 4 类基线模型进行对比，包括：分层强化学习方法 RODE[20] 和 MAVEN；基于值的强化学习方法 QMIX，QTRAN，GraphMIX，ROMA，DOP，LIIR；平均通信的强化学习方法 G2ANet 和 Central-V；多网络的强化学习方法 COMA 和 ATOC[21]。

实验中超参数的设置如表 1 所列。

表 1 超参数设置

Table 1 Hyperparameters setting

参数名	参数值
优化器	Adam
学习率	0.0001
折扣因子	0.95
采样大小	64
ε 初始值	1
ε 衰减率	0.99
软更新因子	0.001
记忆长度	50000

4.2 交通信号控制实验

模拟城市交通(Simulation of Urban Mobility，SUMO)是一个开源的、微观的、多模式的交通仿真环境[22]，能够模拟指定路网中大规模车辆的交通流变化。本文实验主要考虑交通信号灯对车流量的影响。

如图 4、图 5 所示，本文以南京长江大桥周边作为真实交通环境进行实验。该区域内有 41 个路口，包含两车道、三车道、四车道路口。为了确保各算法能够收敛，环境设定遵循先前工作[23]，实验设置为 3000 回合，每回合包含 360 个时间步(6 min)，路网中平均出现 830 辆车。本文模拟的交通流为高峰和非高峰混合交通流。

图 4 交通信号模拟南京长江大桥俯视图

Fig. 4 Traffic signal simulation overhead view of Nanjing Yangtze River bridge

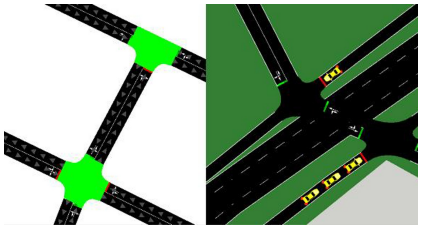


图 5 交通信号模拟路口示意

Fig. 5 Traffic signal simulated intersection idication

1)状态空间

本文设计的状态空间对各路口的车辆拥堵情况进行编码，拥堵情况由等待车辆的数量和累计时间组成，表示为 $S^t=\{S^{\text{num}},S^{\text{wt}}\}$。其中，$S^{\text{num}}$ 为所有路口排队车辆的数量，S^{wt} 为所有车辆在路口的等待时间。

2)动作空间

每个路口的信号灯直接影响当前的拥堵程度，智能体 i 要从动作空间中选择动作 a_i^t，$a_i^t\in A^t$。本文设计的动作空间如式(17)所示：

$$A^t=\{NSS_{\Delta t},NSL_{\Delta t},WES_{\Delta t},WEL_{\Delta t}\} \tag{17}$$

动作由信号灯开启方向和持续时间组成。其中，信号灯开启方向包括南北方向直行(NSS)、南北方向左转(NSL)、东西方向直行(WES)、东西方向左转(WEL)；Δt 代表信号灯的持续时间。本实验中，信号灯可以在不同动作选择之间自由切换，促使基线模型的学习能力最大化。

3)奖励设计

每经过一个时间步，环境中所有信号灯会获得由环境定义的奖励值。该奖励值通过所有车道阻塞车辆的数量计算得到，如式(18)所示：

$$r_n^t=-\sum_{j\in N_i}(queue_{i,j}^{t+d_i^a}+\beta\times wait_{i,j}^{t+d_i^a}) \tag{18}$$

其中，d_i^a 是信号灯动作 a 的持续时间；$queue_{i,j}^{t+d_i^a}$ 是路口 i 到路口 j 车辆的等待长度；$wait_{i,j}^{t+d_i^a}$ 是路口 i 中第一辆车的累计延误时间；β 是平衡系数，遵循先前工作[23]设定为 0.75。

图 6 展示了本文方法和基线方法在模拟交通网络中的实验效果。实线和阴影分别代表平均奖励和标准差。图 6 中各方法的平均收益率曲线总体呈上升趋势，其中 GAG 算法优于其他比较算法，当曲线收敛时，平均奖励值达到 −0.7，收敛速度显著快于基线算法。观察图 6(a)—图 6(c)可以看出，很多方法在上升过程中都实现了稳定收敛；但除了 GAG 和 QTRAN 之外，其他方法在学习过程中都陷入了一定程度的次优策略，不能获得更高的平均奖励值。尽管 QTRAN 和 GAG 表现相似，但 GAG 在图注意力网络训练时有额外的局部和全局约束，QTRAN 在整个训练过程中振荡程度大，在不同的训练回合中表现不同。

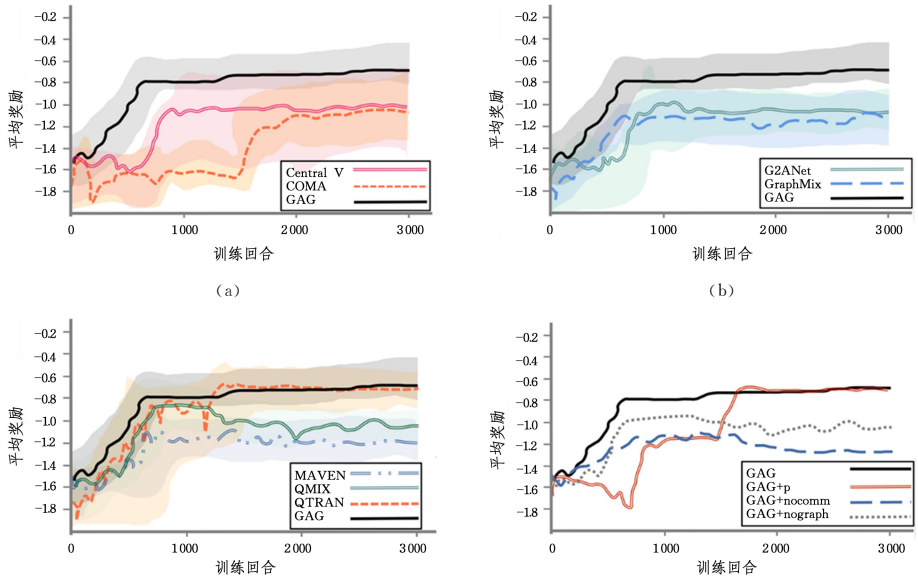


图 6 GAG 与基线方法在 SUMO 环境中的实验结果

Fig. 6 Experiment results of GAG and baseline methods in SUMO environment

图 6(a)比较了隐式通信算法。隐式通信中的智能体在信息共享时处理大量的信息，在大规模的合作任务中，收敛速度会受到严重影响。其中，COMA 在训练早期的平均奖励值没有增加，说明基于个体动作值学习的 COMA 不能有效地初始化参数。

图 6(b)比较了基于图神经网络的算法。可以看出，G2ANet 与 GraphMIX 表现相近。虽然两者都考虑了相邻智能体的信息，但由于图神经网络优化的限制，它们对动作空间的探索有限。

图 6(c)比较了价值分解的方法。同样地，具有评估网络的 QTRAN 表现良好，可以证明评估网络对于大规模智能体训练的有效性。与前两类方法相比，基于值分解的方法表现出了快速、鲁棒的学习能力和收敛速度，说明任务分解方法在大规模的网络环境中具有良好的泛化性。从图中还可以看出，QTRAN 和 QMIX 都优于 MAVEN，原因是在大规模仿真环境下，MAVEN 在 Z 空间中存在过多的噪声，且不具备如图神经网络般的更新关系的能力，因此收敛到次优。此外，QMIX 的表现不及 QTRAN 和 GAG，是因为 QMIX 虽然具有

良好的大规模任务分解的能力，但缺乏反网络或通信功能来提高其性能。

为了验证 GAG 各模块的有效性，进行了消融实验，结果如图 6(d)所示。GAG＋p 是仅保留 P 网络的变体算法，在分散执行过程中存在动作选择噪声，导致收敛速度明显减慢，但最终仍能获得较好的平均奖励值。GAG＋nocomm 是去除了通信功能的变体算法，个体之间缺乏通信信息来考虑全局状态，导致算法收敛结果差。GAG＋nograph 是去除图注意力网络的变体算法，尽管其训练效果稳定，但与完整算法仍存在显著差距，这表明聚合邻域信息确实有助于生成更好的节点表征。上述定量结果证明了 GAG 中各模块的有效性。

为了探索 GAG 中神经网络优化器的最优学习率参数，在不同学习率下进行了对比实验，结果如图 7 所示。0.0001 的学习率在较短的回合数内收敛到最优平均奖励。0.00005 的学习率虽然保持稳定上升的训练，但收敛速度慢，且未能收敛到最优结果。0.0005 和 0.0002 的学习率在收敛速度上有一定提升，但牺牲了最终收敛结果。以上实验结果证明，0.0001 的学习率是本文方法在收敛速度和收敛结果上的一个良好的平衡点。

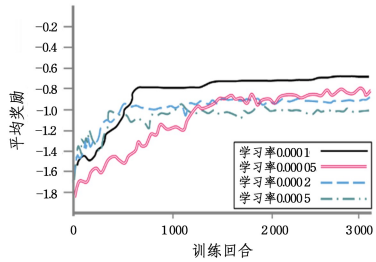


图 7 GAG 在不同学习率下的平均奖励

Fig. 7 Average reward of GAG with different learning rate

此外，对各算法在 SUMO 环境中的平均排放量进行了实验测量，结果如图 8 所示。将各算法训练至收敛，在 SUMO 环境中随机抽取 45 辆车进行测试，以每辆车在整个行驶过程中的平均排放量为指标。同时，在上述比较算法的基础上，增加了随机法作为对照组，该设定下采取随机行动。具体的排放指标包括：一氧化碳(CO/mg)、二氧化碳(CO2/mg)、氢化物(HC/mg)、空气质量指标(PMx/mg)、氮氧化物(NOx/mg)和燃料消耗(fuel/ml)。表 2 列出了各算法的平均排放量结果，图 8 给出了各算法与随机法的减排效果的百分比对比。可以看出，本文方法具有最好的减排效果。

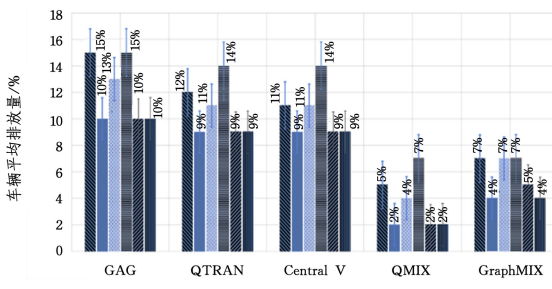


图 8 GAG 与基线方法对比随机法的平均排放量

Fig. 8 GAG and baseline methods compared with average emissions from random method

表 2 GAG 与基线方法在 SUMO 环境中的平均排放量

Table 2 Average emissions of GAG and baseline methods in SUNMO environment

指标	Random	GAG	QTRAN	Central V	QMIX	GraphMIX
CO/mg	10796	9283	9678	9785	10438	10183
CO2/mg	687389	620739	628738	627852	674582	661301
HC/mg	64	57	58	59	63	63
PMx/mg	14	12	14	13	14	14
NOx/mg	276	253	257	261	278	273
fuel/ml	297	269	270	269	293	282

4.3 星际争霸 II 实验

为了验证本文方法在复杂合作任务上的有效性，在星际争霸 II(StarCraft Multi-Agent Challenge，SMAC)[24] 环境中进行了实验。SMAC 是一个多智能体游戏环境，在星际争霸 II 的微场景下两支军队进行对战。根据不同场景设置的任务具有不同的学习难度，这里选择 2s3z 场景。在 SMAC 环境中，我方单位由 MARL 策略控制，敌方单位由游戏内置算法控制。每 5000 步，MARL 算法在 32 个评估集上进行一次评估。

图 9 给出了 GAG 以及基线的平均胜率。大部分算法在 200 万步左右获得稳定的收敛结果。与其他 MARL 算法相比，GAG 获得了最好的指标，这表明 GAG 在微动态环境下具有很强的鲁棒性。其中，基于图的方法 G2ANet 和 GraphMIX 在基线方法中表现最好，说明通信可以显著提高智能体的合作效率。QMIX 在训练时表现不稳定，这可能是由于约束过于简单，导致智能体在动态环境中学习时不能稳定收敛。

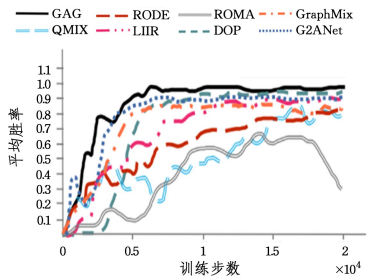


图 9 GAG 与基线方法在 2s3z 地图上的平均胜率

Fig. 9 Average win rate of GAG and baselines in 2s3z map

图 10 展示了本文方法及基线的平均奖励。所有算法均达到了稳定的收敛，本文方法和 G2Anet 取得了最佳表现，说明基于图的方法具有优秀的收敛速率。ROMA 算法在训练过程中出现明显的震荡，原因是基于角色的算法依赖多角色分工实现稳定收敛，在更多角色场景下训练会偏向稳定。

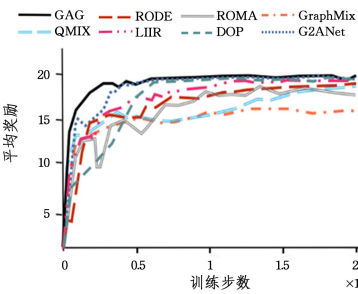


图 10 GAG 与基线方法在 2s3z 地图上的平均奖励

Fig. 10 Average reward of GAG and baselines in 2s3z map

结束语 本文提出了一种基于图注意力的分组强化学习算法。以图结构的方式建模多智能体通信，通过注意力机制促使智能体自发分组，解决信用分配问题。通过设计动作参考网络，引入动作参考使个体智能体进行多样动作选择。SUMO 和 SMAC 的对比实验表明，本文提出的方法有效提高了 MARL 算法的性能。排放量对比实验表明，本文方法可以有效减少交通污染物的排放。未来的工作将集中在通信内容和方式上[25]。考虑到通信带宽的限制和硬件的适配性，计划在该方法中加入层次控制，以适应更智能的现实环境。

参 考 文 献

[1] LI L, ZHAO W, WANG C, et al. Nash double Q-based multi-agent deep reinforcement learning for interactive merging strategy in mixed traffic[J]. Expert Systems with Applications, 2024, 237: 121458.

[2] OROOJLOOY A, HAJINEZHAD D. A review of cooperative multi-agent deep reinforcement learning[J]. Applied Intelligence, 2023, 53(11): 13677-13722.

[3] LI T, ZHU K, LUONG N C, et al. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2022, 24(2): 1240-1279.

[4] LIU Q, SZEPESVÁRI C, JIN C. Sample-efficient reinforcement learning of partially observable markov games[C]// Advances in Neural Information Processing Systems. 2022: 18296-18308.

[5] ZHANG K, YANG Z, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[M]// Handbook of Reinforcement Learning and Control. 2021: 321-384.

[6] YARAHMADI H, SHIRI M E, NAVIDI H, et al. Bankruptcy-evolutionary games based solution for the multi-agent credit assignment problem[J]. Swarm and Evolutionary Computation, 2023, 77: 101229.

[7] JIANG K, LIU W, WANG Y, et al. Credit assignment in heterogeneous multi-agent reinforcement learning for fully cooperative tasks[J]. Applied Intelligence, 2023, 53(23): 29205-29222.

[8] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 2974-2982.

[9] KIM W, PARK J, SUNG Y. Communication in multi-agent reinforcement learning: Intention sharing[C]// International Conference on Learning Representations. 2020: 1-15.

[10] LIU Y, WANG W, HU Y, et al. Multi-agent game abstraction via graph attention neural network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 7211-7218.

[11] NIU Y, PALEJA R R, GOMBOLAY M C. Multi-Agent Graph-Attention Communication and Teaming[C]// Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. 2021: 964-973.

[12] RASHID T, SAMVELYAN M, DE WITT C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(178): 1-51.

[13] SON K, KIM D, KANG W J, et al. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning[C]// International Conference on Machine Learning. 2019: 5887-5896.

[14] NADERIALIZADEH N, HUNG F H, SOLEYMAN S, et al. Graph convolutional value decomposition in multi-agent reinforcement learning[J]. arXiv: 2010.04740, 2020.

[15] WANG T, DONG H, LESSER V, et al. ROMA: multi-agent reinforcement learning with emergent roles[C]// Proceedings of the 37th International Conference on Machine Learning. 2020: 9876-9886.

[16] WANG Y, HAN B, WANG T, et al. Dop: Off-policy multi-agent decomposed policy gradients[C]// International Conference on Learning Representations. 2020: 1-24.

[17] DU Y, HAN L, FANG M, et al. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning[C]// Advances in Neural Information Processing Systems. 2019, 32: 1-12.

[18] MAHAJAN A, RASHID T, SAMVELYAN M, et al. Maven: Multi-agent variational exploration[C]// Advances in Neural Information Processing Systems. 2019: 1-12.

[19] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward[C]// Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. 2018: 2085-2087.

[20] WANG T, GUPTA T, MAHAJAN A, et al. Rode: Learning roles to decompose multi-agent tasks[J]. arXiv: 2010.01523, 2020.

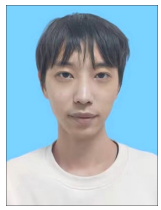
[21] JIANG J, LU Z. Learning attentional communication for multi-agent cooperation[C]// Advances in Neural Information Processing Systems. 2018: 1-11.

[22] WANG X, KE L, QIAO Z, et al. Large-scale traffic signal control using a novel multiagent reinforcement learning[J]. IEEE Transactions on Cybernetics, 2020, 51(1): 174-187.

[23] YANG S, YANG B, ZENG Z, et al. Causal inference multi-agent reinforcement learning for traffic signal control[J]. Information Fusion, 2023, 94: 243-256.

[24] SAMVELYAN M, RASHID T, SCHROEDER DE WITT C, et al. The StarCraft Multi-Agent Challenge[C]// Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. 2019: 2186-2188.

[25] HAN Z R, QIAN Y H, LIU G Q. Multi Agent Communication Based on Self Attention and Reinforcement Learning[J]. Journal of Chinese Computer Systems, 2023, 44(6): 1134-1139.



ZHU Shihao, born in 1997, master. His main research interest is reinforcement learning.

MA Tinghuai, born in 1974, Ph.D, professor, Ph.D supervisor. His main research interests include data mining, social network, privacy preserving and data sharing.

（责任编辑：何杨）