基于图神经驱动自激励的协作多智能体强化学习方法

曹玉康[1)] 刘 全[1),2)] 刘洪哲[1)] 尤任阳[1)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](苏州大学江苏省计算机信息处理技术重点实验室 江苏 苏州 215006)

摘 要 在协作多智能体系统(Multi-Agent System,MAS)中,智能体之间的通信因移动、干扰或带宽受限而动态变化,导致消息丢失与网络拓扑不连通,影响协同决策效率。同时传统的探索策略缺乏针对性,智能体易陷入局部最优,无法充分覆盖环境空间。针对这些问题,提出一种协作多智能体强化学习方法(Graph-based Reinforced Exploration Multi-Agent Reinforcement Learning,GREMARL),该方法将自激励探索(Self-Motivated Exploration,SME)与图神经网络(Graph Neural Network,GNN)多智能体通信方法相结合。其中SME模块通过将状态-动作对熵增量设计为内在奖励信号,使每个智能体能够根据自身对环境未知区域的好奇心动态调整探索优先级。并且SME通过可学习权重参数来自适应衰减,保证在后期更偏重于环境外部奖励,实现从广度探索向深度利用的平滑过渡。GNN通信模块则以动态图卷积网络为基础,通过时序图神经网络端到端地学习通信拓扑的动态演化规律,确保在不同任务阶段关键信息能够沿着最优路径快速传播,而冗余或次要通道则被自动弱化。为了验证GREMARL算法的有效性,将GREMARL方法应用于星际争霸(StarCraft Multi-Agent Challenge,SMAC)与谷歌足球(Google Research Football,GRFootball)环境中。实验结果表明,该方法在SMAC等复杂任务环境中的平均胜率达到了88.8%,比目前最优算法高16.8%。通过设计消融实验,从多个方面验证了自激励探索与图神经建模对GREMARL的必要性。

关键词 深度强化学习;多智能体强化学习;自激励;多智能体探索;图神经网络

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.00029

Graph Neural Network-Driven Self-Motivated Cooperative Multi-Agent Reinforcement Learning Method

CAO Yu-Kang[1)] LIU Quan[1),2)] LIU Hong-Zhe[1)] YOU Ren-Yang[1)]

[1)](School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006)

[2)](Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu 215006)

Abstract In collaborative multi-agent system (MAS), the communication between agents dynamically changes due to mobility, interference, or bandwidth limitations, resulting in message loss and network topology disconnection, which affects the efficiency of collaborative decision-making. At the same time, traditional exploration strategies lack specificity, and intelligent agents are prone to falling into local optima, unable to fully cover the environmental space. A collaborative multi-agent reinforcement learning method, Graph-based Reinforced Exploration Multi Agent Reinforcement Learning (GREMARL), is proposed to address these challenges. This

收稿日期:2025-05-12;在线发布日期:2025-10-28。本课题得到国家自然科学基金(62376179,62176175)、新疆维吾尔自治区自然科学基金(2022D01A238)、江苏高校优势学科建设工程资助项目(PAPD)。**曹玉康**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为深度强化学习、计算机视觉、多智能体强化学习。E-mail:20244027007@stu.suda.edu.cn。**刘 全(通信作者)**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为深度强化学习、自动推理。E-mail:quanliu@suda.edu.cn。**刘洪哲**,硕士研究生,中国计算机学会(CCF)会员,主要研究领域为深度强化学习、多智能体强化学习。**尤任阳**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为深度强化学习。

中国知网 https://www.cnki.net

method combines Self-Motivated Exploration (SME) and Graph Neural Network (GNN) for multi-agent communication. The SME module uses state action entropy increment as an intrinsic reward signal to enable each agent to dynamically adjust exploration priority based on their curiosity about unknown areas of the environment. And the learnable weight parameters of SMEs are adaptively attenuated, ensuring a greater emphasis on external environmental rewards in the later stage, achieving a smooth transition from breadth exploration to depth utilization. The GNN communication module is based on a dynamic graph convolutional network, which learns the dynamic evolution law of communication topology end-to-end through a temporal graph neural network, ensuring that key information can quickly propagate along the optimal path at different task stages, while redundant or secondary channels are automatically weakened. In order to verify the effectiveness of the GREMARL algorithm, experiments were conducted in the StarCraft Multi Agent Challenge (SMAC) and Google Research Football (GRFootball) environments. The experimental results showed that the average win rate of GREMARL in the SMAC complex task environment reached 88.8%, which was 16.8% higher than the SOTA algorithm. By designing ablation experiments, the necessity of self excitation exploration and graph neural modeling for GREMARL was verified from multiple aspects.

Keywords deep reinforcement learning; multi-agent reinforcement learning; self-motivation; multi-agent exploration; graph neural network

1 引 言

作为深度强化学习[1-2]的重要分支,多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)由于涉及多个智能体,因而需要学习有效的策略,使任务的累计奖赏最大化。近年来,基于协作的多智能体强化学习(Collaborative Multi-Agent Reinforcement Learning,CMARL)因其在机器人协作[3-4]、智能交通[5-6]、无人机编队[7-8]等领域的广泛应用而备受关注。多智能体系统具备通过局部交互与协同合作完成复杂任务的潜力,使其在自适应、容错以及分布式控制等方面拥有明显优势。然而,在实际应用中,CMARL 面临的两个主要挑战使得该领域的研究进展受到限制。

首先,在多智能体中探索不足问题尤为突出[9]。传统 MARL 方法主要依赖环境提供的外部奖励信号进行学习,当奖励较为稀疏或存在延迟时,智能体很难获得充分的反馈,导致探索过程易陷入局部最优,无法有效发现全局最优策略。特别是在复杂、动态环境下,如何促使每个智能体主动发现未开发的信息区域和潜在任务目标,成为亟待解决的问题[10]。

其次,通信低效问题也是多智能体系统的核心瓶颈[11]。现有方法往往采用固定或者简单的信息传输策略,而在实际场景中,智能体之间的交互关系可能具有高度动态性和非线性特征。传统通信机制难以捕捉这种复杂性,限制了智能体在协同决策时的信息共享效率,从而影响系统整体性能。如何构建一种灵活、高效的通信机制,使各智能体在多变环境中能够实时共享和整合关键信息,是推动 MARL 发展的关键。

为了解决上述问题,本文引入了两种创新思路。首先是自激励探索(Self-Motivated Exploration,SME),该方法通过构建自激励信号促使每个智能体在环境中主动探索,从而克服传统探索策略中局部最优和反馈稀疏的问题。自激励模块不仅依托于智能体自身的状态信息,还结合了历史行为和探索多样性指标,确保在面对复杂环境时能够捕捉到更多潜在信息,从而推动策略进化。

其次是图神经网络通信(Graph Neural Network,GNN)的思想,将多智能体系统映射为图结构,其中每个智能体被视为图中的一个节点,智能体之间的交互关系则通过边权重进行量化。利用 GNN 强大的消息传递与信息聚合能力,可以有效捕捉智能体之间复杂的时空动态关系,实现高效的信息共享和协同决策。通过构造动态通信图,系统不仅能自适应地调整智能体之间的联系,还能在信息

中国知网 https://www.cnki.net

传输过程中自动识别并过滤噪声信息,提高整体决策的准确性和鲁棒性。

基于上述思路,本文提出了一种融合自激励探索和图神经网络通信的多智能体协作方法 GREMARL。该方法不仅在探索过程中引入内部激励机制,促使智能体主动拓展未知领域,而且利用图神经网络构造动态通信结构,实现高效、灵活的信息共享。本文进一步设计了联合优化算法,动态平衡各智能体的内部奖励与全局任务奖励,通过集中训练与分散执行策略(Centralized Training with Decentralized Execution,CTDE)实现系统的快速收敛和稳健协同。实验结果表明,该方法在加快收敛速度、提升任务完成率以及增强系统鲁棒性方面均优于传统方法。

本文的主要贡献可以总结为以下 3 点:

(1)提出了一种融合自激励探索与图神经网络通信的多智能体强化学习方法,有效解决了探索不足与通信低效的双重挑战。

(2)设计了联合优化算法,实现内部激励奖励与全局任务奖励之间的动态平衡,提升了系统在复杂场景下的协同决策性能。

(3)通过实验验证了该方法在多种复杂任务场景下的优越性,为大规模、多任务的多智能体系统协作提供了新思路和新方法。

本文第 2 节总结了该领域的相关研究成果;第 3 节详细介绍了本文提出的模型 GREMARL,包括模型中的组成成分、实现细节和训练过程;第 4 节为本文的实验部分,在 2 个多智能体任务上进行实验分析;最后得出结论,并介绍未来可能的研究工作。

2　相关工作

2.1　协作多智能体强化学习

多智能体强化学习是强化学习(Reinforcement Learning,RL)在 MAS 中的扩展,旨在使多个智能体通过不断试错和学习,最终形成一套协调一致的策略,以完成复杂任务[12]。

MARL 一般包含以下要素:

(1)状态空间:表示智能体所处的环境状态,可为全局可观测或部分可观测。

(2)动作空间:智能体在当前状态下可采取的所有可能动作集合。

(3)奖励函数:智能体根据动作获得的回报,可能是全局共享奖励或个体奖励。

(4)策略:智能体基于状态采取行动的映射,可以是确定性策略或随机策略。

(5)环境动态:描述智能体采取动作后,环境状态的变化情况。

CMARL 则可以从数学模型、算法范式以及通信机制 3 个方面进行阐述。

2.1.1　数学模型

CMARL 将强化学习问题建模为局部可观测马尔可夫决策过程(Decentralized Partially Observable Markov Decision Process,Dec-POMDP)。Dec-POMDP 可以描述为九元组 $G=<S,A,U,P,r,Z,O,N,\gamma>$,其中 S 表示环境的全局状态,A 表示 n 个智能体的集合,U 为动作空间。在每个时间步 t,智能体 $a\in A\equiv\{1,\cdots,n\}$ 选择一个动作 $u\in U$,形成一个联合行动 $u\in U\equiv U^n$,进而导致由状态转移函数 $P(s'/s,u)$ 表示的环境转换表示为 $S\times U\times S\to[0,1]$。所有的智能体都共享相同的奖励函数 $r(s,u):S\times U\to R$,而 $\gamma\in[0,1)$ 是一个折扣因子。

整个系统的目标是最大化全局累计回报

$$J(\pi)=\mathrm{E}\left[\sum_{t=0}^{\infty}\gamma^t(s_t,a_t^1,a_t^2,\cdots,a_t^N)\right] \tag{1}$$

其中,$\pi=\{\pi^1,\cdots,\pi^N\}$ 为各智能体的策略组合。

2.1.2　算法范式

目前基于协作多智能体强化学习的算法可以分为基于值函数分解(Value Decomposition,VD)与基于策略梯度(Policy Gradient,PG)两种。

VD 方法通过对联合 Q 值函数 $Q_{tot}=(s,a)$ 进行因式分解,降低学习难度。如 VDN(Value-Decomposition Networks,VDN)假设 $Q_{tot}=\sum Q^i(o^i,a^i)$ 保留了个体贡献的可加性,但忽略了高阶交互[13]。QMIX(Monotonic Value Function Factorization,QMIX)增加了单调性约束,$\partial Q_{tot}/\partial Q^i\geqslant 0$,通过可学习的混合网络保证 $\mathrm{argmax}_a Q_{tot}=(\mathrm{argmax}_{a^1}Q^1,\cdots,\mathrm{argmax}_{a^N}Q^N)$,在一定程度上平衡了表达与可分解能力[14]。

QPLEX(Duplex Dueling Multi-agent Q-learning,QPLEX)则通过引入 双路竞争网络结构,将联合动作值函数分解为优势函数与状态值函数,在严格保持单调性约束的同时,理论上保证了对于最优联合策略的完整表达能力。该方法避免了复杂的辅助优化目标,具有更高的样本效率与学习稳

中国知网　https://www.cnki.net

定性[15]。

基于 PG 的方法直接优化策略参数，使期望回报最大化。如 COMA(Counterfactual Multi-Agent Policy Gradients，COMA)方法通过设计对照基线 $\sum_{j\neq i}Q_{tot}(s,(a^{i}_{ref},a^{-i}))$，精确量化个体行为对全局回报的边际贡献，缓解了多智能体信用分配问题[16]。MADDPG(Multi-Agent Deep Deterministic Policy Gradient，MADDPG)将 DDPG 扩展至多智能体场景，使用集中式全局 Q 网络和局部策略网络，保证了样本效率与策略稳定性[17]。文献[18]提出的 UPDET(Universal Policy Decoupling with Transformers)通过引入 Transformer 模块解耦多智能体策略，并在多个复杂任务中表现出色。其显著特点是以注意力机制聚合局部信息，通过共享架构提升泛化能力。尽管 UPDET 在 StarCraft II 环境中的性能强劲，但在更具异构性和高动态性的场景(如 GRFootball)中仍存在适应性不足的问题。文献[19]则从值函数稳定性角度出发，重新设计了训练技巧，增强了收敛性，但其对通信建模能力有限，适用于低干扰场景。

2.1.3 通信机制

通信在 CMARL 中不仅是信息传递，更是隐式的协同图拓扑学习。近年来，GNN 提供了统一的消息传递理论，表示为

$$h_i^{(l+1)}=\sigma\Big(\sum_{j\in N(i)}f_\theta(h_i^{(l)},h_j^{(l)},\omega_{ij})\Big) \tag{2}$$

其中，$h_i^{(l)}$ 指在第 l 层时，智能体(或图节点) i 的特征向量(隐藏表示)。$N(i)$ 指节点 i 在当前通信图中的邻居集合，即所有与 i 有边相连的其他智能体索引集合。ω_{ij} 指边 (i,j) 的权重，用于量化 i 与 j 之间通信的重要性或带宽等属性，且可随训练动态更新。f_θ 为消息函数，参数为 θ，将节点 i 和邻居 j 的特征及它们的边权重映射到一条消息向量。$\sigma(\cdot)$ 指非线性激活函数，用于给聚合后的特征添加非线性变换。

尽管 CMARL 具备较强的协作与优化能力，但随着智能体数量的增加，仍然面临着一系列挑战：

(1)状态和动作空间的指数增长。

在 MAS 中，状态空间和动作空间随着智能体数量呈指数级增长，导致传统强化学习方法难以直接扩展到多智能体环境。例如，在 N 个智能体的情况下，联合状态空间 S 和联合动作空间 A 分别是

$$|S|=\prod_{i=1}^{N}|S_i|,|A|=\prod_{i=1}^{N}|A_i| \tag{3}$$

其中 S_i 和 A_i 分别表示智能体 i 的状态和动作空间维度。当 N 较大时，计算复杂度急剧上升。

(2)智能体间的非平稳性

在多智能体环境中，每个智能体的策略都会不断变化，使得环境对其他智能体而言是非平稳的。这种非平稳性导致传统单智能体 RL 方法(如 Q-learning 方法)在 MARL 中难以收敛[20]。

为了解决非平稳性问题，文献[21-22]提出了一系列方法，如集中训练-分布执行(CTDE)框架，即在训练阶段收集全局信息进行优化，而在执行阶段每个智能体独立决策。文献[23]提出马尔可夫博弈，利用博弈论建模智能体之间的动态交互，以求 Nash 均衡策略。

(3)智能体之间的信息不对称

在现实场景中，智能体往往无法访问全局状态，只能基于局部观测信息进行决策。例如，在自动驾驶场景中，每辆车只能观察到自身传感器范围内的车辆，而无法直接获取整个路网的交通状况。这种部分可观测性(Partially Observable Markov Decision Process，POMDP)进一步增加了决策难度。这种部分可观测性通常建模为部分可观测马尔可夫决策过程，使得智能体需要使用递归神经网络或注意力机制来整合历史信息[24]。

(4)通信效率与信息共享

在协作型任务中，智能体需要交换信息以进行决策，但通信成本、带宽限制和噪声干扰都会影响信息共享的质量。因此，如何在 CMARL 中设计高效的信息共享机制是当前研究热点之一。部分学者提出基于 GNN 的通信模型。智能体构建一个通信图，通过 GNN 进行信息聚合和共享，提升通信效率[25]。

2.2 自激励探索

在传统的强化学习中，智能体依赖环境提供的外部奖励进行学习。然而，在稀疏奖励环境下，智能体可能面临难以有效探索的问题，从而导致学习过程缓慢。为了解决这一问题，文献[26]提出了 SME 策略，通过设计内部奖励机制，鼓励智能体在缺乏外部奖励时进行有效的探索。SME 策略主要包括几种方法：(1)预测误差驱动的探索通过学习预测环境状态变化，利用较大的预测误差为智能体提供更高的内部奖励，激励其探索未曾到达的状态。(2)状态新颖性驱动的探索则基于状态的罕见性或新颖性来定义内部奖励，如使用随机网络蒸馏(Random Network Distillation，RND)[27]衡量当前状态的新颖

中国知网 https://www.cnki.net

性,或者通过信息增益方法估计状态的不确定性,从而鼓励智能体探索更多未知区域。(3)计数驱动的探索则在离散状态空间中,根据智能体对某一状态的访问频率来调整奖励,对于较少访问的状态给予更高奖励,促进对环境中未探测区域的探索[28]。这些策略通过有效的内部奖励机制,使智能体能够在稀疏奖励环境中加速学习,充分探索环境。

2.3　基于图建模的多智能体通信

早期的 CMARL 方法主要依赖于共享状态信息或固定规则的通信协议,但这些方法难以扩展到大规模、多智能体系统中。近年来,GNN[29]在处理复杂网络数据方面表现出了显著优势,特别是在捕捉局部邻域和全局结构信息上,将多智能体系统视为图结构,每个智能体作为图中的一个节点,边则代表智能体之间的距离、任务相关性或历史交互信息[30]。通过 GNN 的邻域信息聚合与传递,智能体能够获取全局环境的有效表达,从而提升协同决策的水平。在 GNN 模型中,智能体作为图节点,而智能体之间的交互则通过边连接,边的权重基于任务相关性、物理距离或历史交互等因素。信息聚合机制通过 GNN 的消息传递方式,促进智能体之间的信息交流与协作,显著提高系统的整体协作效率[31-32]。

2.4　融合探索与通信的研究现状

尽管自激励探索和图神经网络通信各自为多智能体系统带来了不同程度的提升,但目前大多数工作集中于单一模块的优化,如何将两种方法有机结合,既保持智能体探索的主动性,又确保协作通信的高效性,仍是当前研究的难点[33]。

3　GREMARL 方法

3.1　图神经通信模块

在多智能体协作中,智能体间的实时信息共享至关重要。为此,本模块通过构造图结构对通信进行建模。

3.1.1　图结构构造

将各个智能体视为图节点,依据任务相关性、物理距离及状态相似性构造边,并为不同边赋予不同权重。之后利用图神经网络对节点状态信息进行传递与聚合,捕捉复杂的时空交互关系,使每个智能体能获得邻域内更为准确的全局信息。

首先,每个智能体通过环境获得局部观测 o_i 与位置 p_i,之后进行图构建,输入为智能体的观测状态集合 $\{o_i\}_{(i=1)}^{N}$,位置信息集合 $\{p_i\}_{(i=1)}^{N}$,构建节点特征,表示为

$$h_i^{(0)}=MLP([o_i \parallel p_i]) \in \mathbb{R}^{(d_h)} \tag{4}$$

其中,d_h 为嵌入维度,MLP 为多层感知机。

边权重计算表示为

$$e_{ij}=\sigma(W_e \cdot [h_i^{(0)} \parallel h_j^{(0)} \parallel d_{ij}]) \tag{5}$$

其中,d_{ij} 为智能体 i 与 j 之间的欧氏距离,$\sigma(\cdot)$ 表示 sigmoid 函数,$W_e \in R^{3d_h \times 1}$,只保留每个节点的 Top-3 边。

3.1.2　动态更新

$$h_i^{hist}=GRU(h_i^{(0)},h_i^{hist}) \tag{6}$$

3.1.3　GNN 编码器进行消息传递(三层迭代)

$$m_{ij}^{(l)}=ReLU(W_m^{(l)}[h_i^{(l)} \parallel e_{ij}]) \tag{7}$$

$$h_i^{(l+1)}=LayerNorm(h_i^{(l)}+W_u^{(l)} \sum\nolimits_{j \in N(i)} m_{ij}^{(l)}) \tag{8}$$

其中,$W_m^{(l)}$ 与 $W_u^{(l)}$ 为可学习参数,$N(i)$ 表示节点 i 的邻居。

3.1.4　多头注意力增强

$$Attn(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M\right) \cdot V \tag{9}$$

其中,Q,K,V 为节点嵌入 $h_i^{(3)}$ 线性变换得到。掩码矩阵 M 由邻接矩阵定义,屏蔽非连接节点。

3.1.5　解码器与混合网络

$$Q_i=W_{q2} \cdot GELU(W_{q1} \cdot h_i^{(3)}) \in \mathbb{R}^{|A|} \tag{10}$$

其中,$W_{q1} \in \mathbb{R}^{(d_h \times 256)}$,$W_{q2} \in \mathbb{R}^{(256 \times |A|)}$,$|A|$ 为动作空间维度。

$$Q_{tot}=W_2 \cdot Relu(W_1 \cdot [Q_1,\cdots,Q_N]+b_1)+b_2 \tag{11}$$

权重 W_1 与 W_2 和偏置 b_1、b_2 由超网络生成,表示为

$$[W_1,b_1,W_2,b_2]=HyperNet(s_{\text{global}}) \tag{12}$$

超网络设计为

$$HyperNet(s_{\text{global}})=MLP(s_{\text{global}}) \odot Abs(\cdot) \tag{13}$$

其中,$Abs(\cdot)$ 确保权重非负。

3.2　自激励探索模块

为应对传统强化学习中探索不足的问题,本模块为每个智能体设计了内部激励信号。

内部奖励设计:每个智能体基于当前状态、历史轨迹和环境不确定性生成自激励奖励,使其在环境中更主动地探索未知区域。通过将自激励奖励与环境外部奖励相结合,利用 QMIX 更新智能体策略,从而达到探索与利用的平衡。

中国知网　https://www.cnki.net

内在奖励计算为

$$r_{\text{int}}^{i}=\alpha\cdot H(\pi_i(\cdot\mid s))+\beta\cdot\frac{1}{\sqrt{N(s_i)+1}} \tag{14}$$

其中，$H(\pi_i)=-\sum_a\pi_i(a\mid s)\log\pi_i(a\mid s)$ 为策略熵，$N(s_i)$ 为状态 s_i 的历史访问次数。

总奖励表示为

$$r_{\text{total}}=r_{\text{ext}}+r_{\text{int}}+\eta\cdot\sum_{i\neq j}\mathbb{I}_{\text{conflict}}(a_i,a_j) \tag{15}$$

其中，$\mathbb{I}_{\text{conflict}}$ 为冲突指示函数，η 为冲突惩罚系数。

3.3 联合训练与策略优化模块

为同时优化探索和通信两个模块，提出了一种多目标联合优化方法。(1)联合损失函数：构造包含外部任务奖励、自激励探索奖励与图神经网络通信误差的联合损失函数。(2)动态权重调整：在训练过程中根据任务进展动态调整各项奖励权重，保证系统在不同阶段既能保持充分探索，又能实现高效信息共享。

TD 误差损失表示为

$$L_{TD}=\mathbb{E}\left[Q_{\text{tot}}(s)-(r+\gamma\max_{a'}Q_{\text{tot}}(s'))^2\right] \tag{16}$$

通信一致性损失表示为

$$L_{\text{comm}}=\frac{1}{N}\sum_{i=1}^{N}D_{KL}(h_i^{enc}\mid h_i^{dec}) \tag{17}$$

策略熵正则化表示为

$$L_{\text{ent}}=-\frac{1}{N}\sum\nolimits_{i=1}^{N}\sum_a\pi_i(a\mid s)\log\pi_i(a\mid s) \tag{18}$$

总损失表示为

$$L_{\text{tot}}=L_{TD}+0.1L_{\text{comm}}+0.01L_{\text{ent}} \tag{19}$$

如图 1 所示，其为 GREMARL 的算法结构图，该算法包括智能体模块、图通信模块、编码器模块、解码器模块、自适应探索模块以及联合训练与损失计算模块。其中蓝色箭头表示环境状态和奖励流向智能体模块。绿色箭头表示通信图、编码、解码、融合及策略输出的正向数据流。橙色箭头表示自适应探索模块输出的最终决策反馈给智能体。红色箭头表示联合训练中的反向传播更新，反馈给各网络模块。

智能体模块包括观测与策略网络等内容，将各智能体的观测与内部奖励经过策略网络计算后，再传递给通信图构建模块。而在反向传播时，联合训练模块的损失通过红色箭头反馈更新智能体的网络参数。

图通信模块主要根据各智能体的位置信息、状态相似度和历史交互构建通信图，确定图的节点和边，并将构建好的图结构(节点与边)以绿色箭头传递给编码器模块。

编码器模块利用 GNN 与掩码多头注意力机制对通信图中的节点与边进行嵌入和信息聚合。编码后的嵌入信息以绿色箭头传递给解码器模块。

解码器模块汇聚编码器输出，生成更新后的 智能体表征，并计算 Q 值，以绿色箭头传递到混合网络。

混合网络融合各智能体的输出，生成全局决策信息，并将融合后的全局信息传递到自适应探索模块。

自适应探索模块根据混合网络的输出，以及各智能体内部奖励与外部奖励，计算和更新探索参数 α，决定最终动作采样策略。

联合训练与损失计算模块整合外部奖励、内部奖励、通信一致性损失以及值函数损失，计算联合损失函数。联合损失通过反向传播(红色箭头)更新 Agent 策略网络、编码器与解码器网络的参数，实现系统的持续优化。

具体算法流程如算法 1 所示。

算法 1. GREMARL 算法流程

输入：经验回放池 D、训练轮数 E、批量大小 B、折扣因子 γ、目标网络更新率 τ、初始探索率 ε

输出：训练后的策略网络参数 θ_{dec}、θ_{mix}

1. 初始化：GNN 编码器、解码器、混合网络，复制目标网络参数、经验回放池，设置优化器
2. 开始训练循环
3. 环境交互：每个智能体通过环境获得局部观测与位置(公式(3))
4. 构建动态通信图
5. 节点特征提取(公式(4))
6. 边权重计算(公式(5))
7. GNN 编码器生成节点嵌入(公式(7)，(8)，(9))
8. 解码器计算个体 Q 值(公式(10))
9. 混合网络计算全局 Q 值(公式(11))
10. 动作选择与执行：以概率 ε 随机选择动作，否则选择 argmax；执行联合动作，获取环境奖励
11. 奖励计算与存储：
12. 计算内在奖励(公式(14))
13. 将经验元组存入经验回放池
14. 参数更新：
15. 从回放池中采样批次数据
16. 计算目标 Q 值
17. 计算总损失(公式(19))
18. 反向传播更新各网络参数
19. 软更新目标网络参数
20. 探索率 ε 衰减
21. 训练结束，返回最终策略网络参数

中国知网 https://www.cnki.net

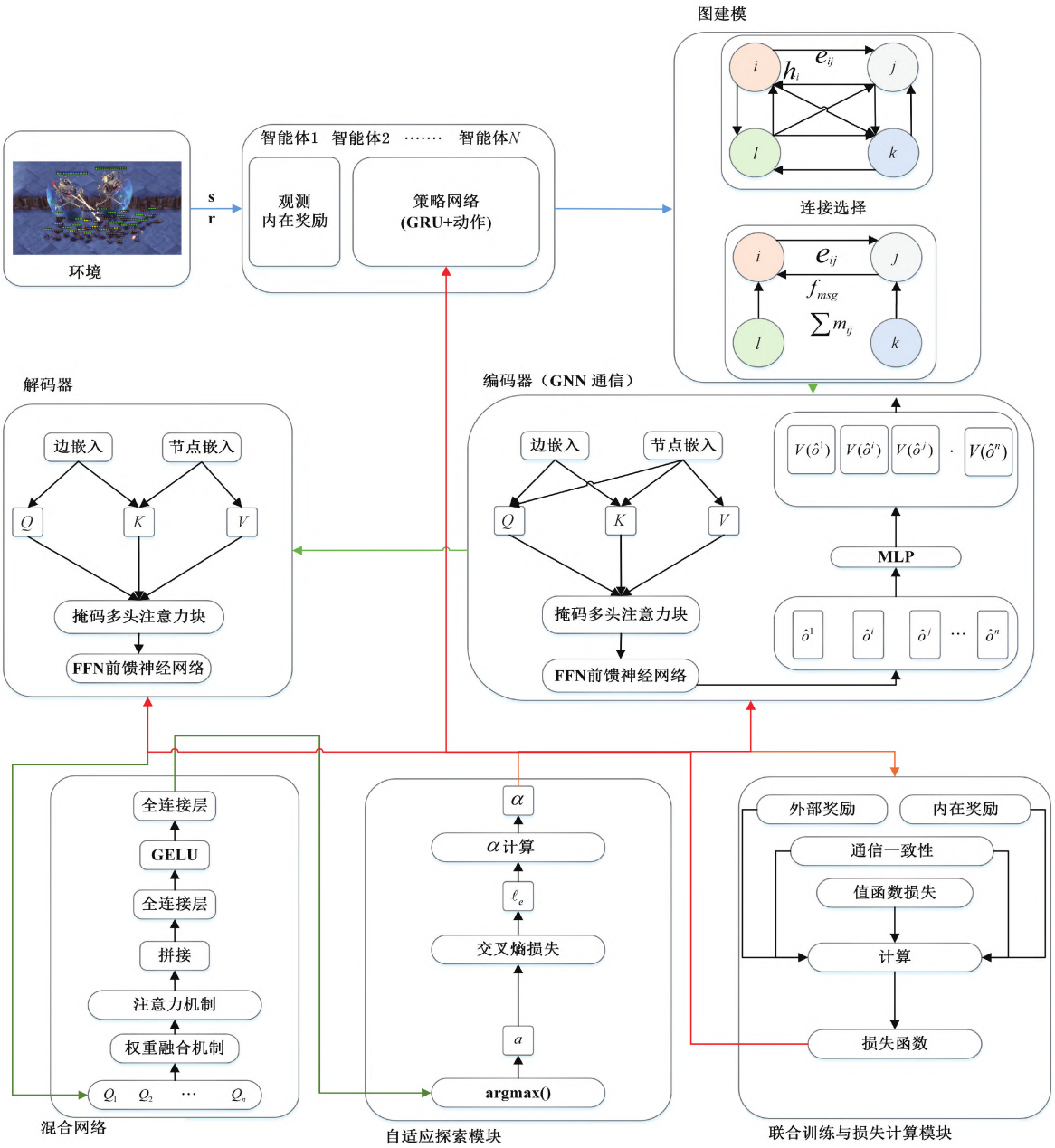

图 1　GREMARL 算法结构图

4　实　验

4.1　实验环境

实验环境使用星际争霸(StarCraft Multi-Agent Challenge,简称 SMAC)[34]、谷歌足球(Google research football,简称 GRFootball)[35]等多个任务场景,要求智能体在环境中进行信息采集与任务协同。

SMAC 环境构建了多样化的微观战斗场景,让多个智能体在局部观测条件下展开合作或对抗。这些精心设计的场景要求智能体掌握一项或多项微管理技巧,以便战胜对手。每个场景呈现两支部队之间的交锋,其初始位置、数量和单位类型均因场景而异。智能体可执行一系列离散动作,如向北、向南、向东、向西移动、攻击指定敌人或进行治疗等,而这些动作的具体范围和约束则取决于场景设置。SMAC 集成了一系列基于星际争霸Ⅱ的游戏,可以在多种复杂场景下开展多个智能体的实验,专门用于评估智能体在复杂协作任务中的表现。

GRFootball 环境在多智能体强化学习领域独具特色,其精心设计的多样化足球比赛场景允许多个智能体在有限的局部视野下实现协作与对抗。这些场景不仅真实再现了足球比赛的动态过程,还要求智能体掌握传球、射门、防守和进攻等多种细腻操作

中国知网　https://www.cnki.net

技巧，以便在比赛中占据优势。每场比赛都代表着两队之间的较量，队伍构成、球员位置和初始状态均因场景不同而变化。智能体需要执行包括带球移动、传球、射门和抢断等一系列动作，而这些动作的具体范围和效果则依据比赛规则及场景设定而有所差异。

如图 2 所示，SMAC 提供了多个挑战性任务，如“2c vs 64zg”代表了极端的不对称战争场景，考察通信效率；表 1 则展现了 GRFootball 环境所提供的多个高度动态环境，如“counterattack easy”任务中智能体需在瞬间协作决策，这与探索能力的验证目标较为契合[36]。这两个环境综合测试了本文提出方法在通信与探索两方面的改进能力。

评价指标主要考察任务完成率、收敛速度、信息传递效率及系统鲁棒性。具体而言则是包括胜率、奖励回报以及收敛速度等指标，从而全面评估多智能体强化学习方法的性能。

(a) 2c vs 64zg (b) 3m (c) 3s5z

(d) 5m vs 6m (e) 25m (f) MMM

图 2 SMAC 场景示意图

表 1 GRFootball 场景细节

名称	描述
空门射门	我方球员站在禁区内并持球，需要攻入空门
空门	我方球员站在球场中场并持球，需要攻入空门
3 对 1 与守门员	三名我方球员从禁区边缘尝试进球，一个在两侧，另一个在中间。起初，中间的球员持球并面对防守球员。场上有对方守门员
角球	标准角球情况，但角球球员可以从角旗区带球。失去控球时不会结束本回合
简单反击	4 对 1 反击，带守门员；双方其他球员都朝球跑回
困难反击	4 对 2 反击，带守门员；双方其他球员都朝球跑回
11 对 11 懒惰对手	全场 11 对 11 比赛，对方球员无法移动，但如果球靠近他们，他们可以拦截。我方中卫起初持球。本回合的最大时长为 3000 帧，而不是 400 帧

4.2 实验设置

任务设置：本文选取 QMIX、CW-QMIX[37]、QTRAN[38]、RIIT、UPDET 等作为基准算法，基准算法都是通过不同的方式分解和组合智能体的局部 Q 值函数，以实现全局最优策略的学习，从而便于在多个方面展现本文所提算法的优越性。为保证实验的公平性，本文所使用算法的环境都基于关键参数值如表 2 所示。

表 2 实验所使用的主要参数

参数名称	参数值	参数名称	参数值
批次大小	128	优化器	Adam
经验回放池容量	5000	折扣因子	0.99
隐藏层维度	64	学习率	5e-4

在相同实验环境下，每次实验对这 6 种算法进行 5 次随机种子的实验，并取平均值来比较算法的性能。本文所有实验均采用配置为 i7-14700KF CPU、NVIDIA GeForce RTX 4070 Ti GPU 和 16GB 内存的服务器作为硬件环境。软件环境方面，实验基于 Ubuntu 20.04 操作系统搭建，并采用 PyTorch 深度学习框架以保障模型训练的稳定性与高效性。所有算法参数设置均参考对应原始文献或官方代码，并在多个环境中保持一致，确保公平对比。

所有基准算法均基于公开实现。如 UPDET、QMIX 以及 QTRAN 等算法，均使用其官方 GitHub 仓库版本，并复现实验以确保一致性。

中国知网 https://www.cnki.net

4.3　对比实验

本节在 SMAC 与 GRFootball 多智能体环境上展开对比实验,旨在证明 GREMARL 相对于 5 个基准方法的有效性。对比曲线图中的实线代表五次随机种子运行结果的均值,阴影表示标准差。

4.3.1　在 SMAC 上的表现

本文选用 SMAC 中的一个简单(3m)、两个困难(8m、2c vs 64zg)以及三个超难地图(5m vs 6m、8m vs 9m、10m vs 11m)进行实验,间隔 1e4 个时间步测试一次平均胜率,绘制出对比实验的训练曲线如图 3 所示。根据图 3 中的结果,GREMARL 在六个任务上均取得了最佳胜率。

根据图 3 可以看出,在多个场景中,GREMARL 表现出较为显著的提升潜力,尤其是在训练进程的后期。在某些情况下,GREMARL 的胜率随着步数的增加,逐渐超越了其他模型。这表明,尽管 GREMARL 的早期表现较为波动,但它具有较强的适应性,能够在长期训练后实现较大的提升。特别是在 8m 和 10m vs 11m 场景中,GREMARL 表现出逐步上升的趋势,并在后期逐渐接近甚至超越其他算法,表明它能够从早期的波动中恢复过来,具备较强的长期学习能力。而从竞争力来看,尽管在大部分图表中,QMIX 和 CW-QMIX 持续占据领先地位,但 GREMARL 在一些特定场景下(如 5m vs 6m 和 8m vs 9m)展现了较强的竞争力。在这些场景中,GREMARL 在后期的训练过程中逐渐追赶并超越了其他模型,表明其具备较高的学习潜力和弹性。从算法稳定性来看,GREMARL 与其他算法(如 RIIT 和 UPDET)相比,表现出更大的波动性。尽管其初期的表现并不稳定,但这种波动反映出它在面对复杂训练环境时的高度适应性。与此相对,RIIT 和 UPDET 虽然表现出更稳定的增长,但其增长较为平缓,缺乏 GFMARL 那样的后期爆发性提升。

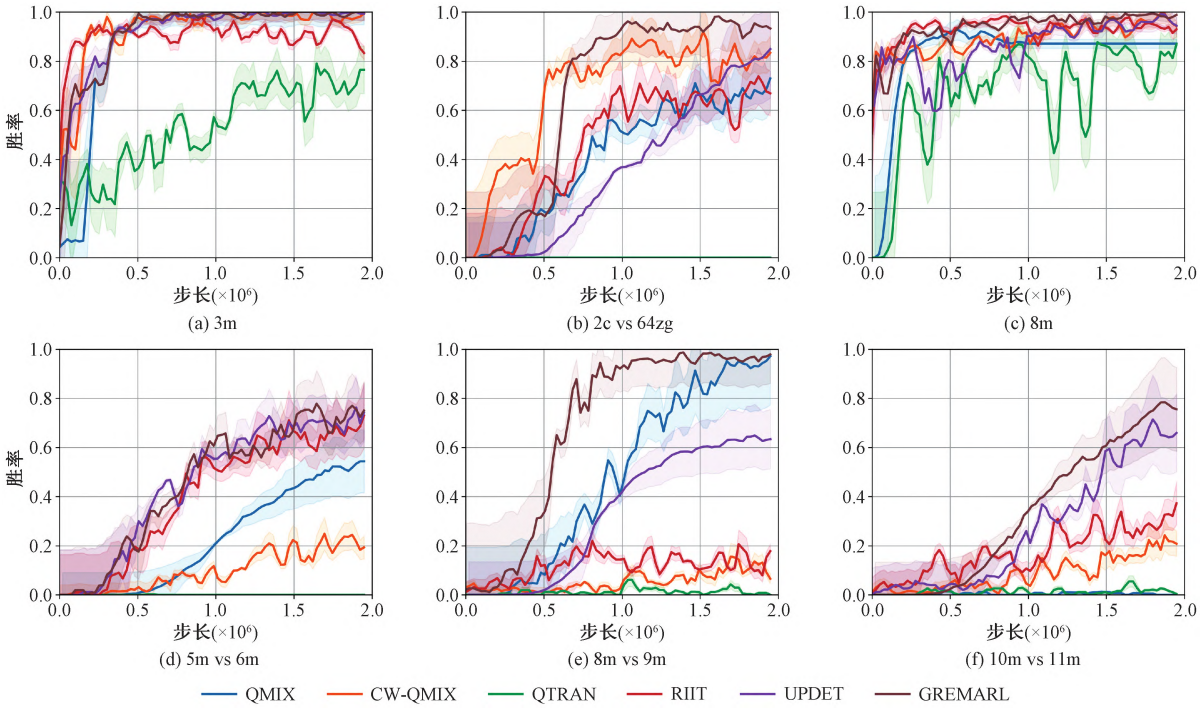

图 3　SMAC 场景结果图

如表 3 所示,本文统计了各个算法在六种地图上最后 4e5 步的平均胜率,在多个困难任务上的胜率都高于其他基准算法,说明 GREMARL 具备较强的泛化能力,可以适应各类地图。而 QMIX 与 RIIT 在非极端对抗场景下表现稳健。UPDET 次之。

表 3　算法在 SMAC 地图上最后 4e5 步的平均胜率

算法	Map					
	3m	8m	2c vs 64zg	5m vs 6m	8m vs 9m	10m vs 11m
QMIX	0.91	0.90	0.68	0.55	0.94	0.00
CW-QMIX	0.86	0.88	0.88	0.20	0.12	0.20
QTRAN	0.75	0.72	0.00	0.00	0.02	0.00
RIIT	0.89	0.88	0.65	0.72	0.16	0.33
UPDET	0.87	0.88	0.52	0.75	0.62	0.68
GREMARL	0.94	0.93	0.94	0.80	0.97	0.75

CW-QMIX 在特定地图(低干扰)上有竞争力。QTRAN 在高复杂度或对抗性地图上难以收敛。

总的来说,在 SMAC 任务上的一个简单、两个

中国知网　https://www.cnki.net

困难任务以及三个超难任务上，GREMARL 算法与 QMIX、CW-QMIX、UPDET 等基准方法相比，都展现了更好的收敛性能以及更高的胜率。

4.3.2 在 GRFootball 上的表现

如图 4 所示，在三个复杂程度不同的任务（如带守门员的攻防对抗、快速反击、传射配合）中，GREMARL 的步数表现均显著优于其他算法。这表明其对多样化场景具有更强的适应能力，尤其在需要协调多智能体合作的动态环境中（如任务 c），其优势更为突出。同时 GREMARL 能以更少的训练步数达到目标性能，学习效率更高。这种特性对计算资源受限的实际应用场景尤为重要。这也说明 GREMARL 通过引入自激励探索以及图神经网络通信方法，解决了传统算法（如 QMIX 的单调性约束、QTRAN 的优化复杂性）的局限性，从而在多智能体协作中实现更高效的策略学习。

如表 4 所示，其为算法在 GRFootball 地图上的平均回报，GREMARL 在三个不同的任务中始终表现出较高的回报值，特别是在需要较高协调性和策略的任务中，如 counterattack easy 和 3 vs 1 with keeper。与其他算法相比，GREMARL 显示了更强的可扩展性和适应能力，能够在多个任务中持续提升表现，尤其是在复杂的多智能体环境中。这表明，GREMARL 在训练 2×10^6 步之后，能够在各个任务中获得更高的平均回报。

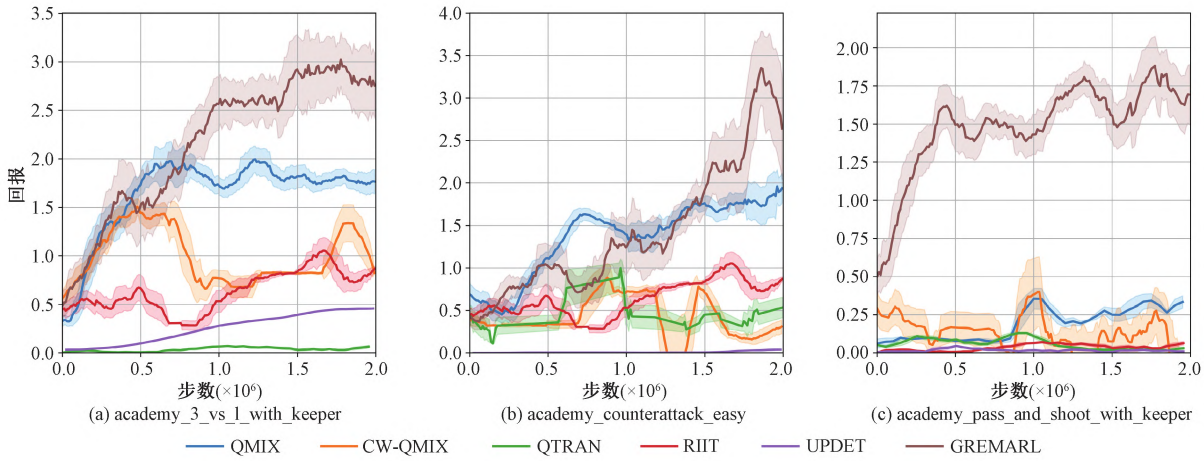

图 4 GRFootball 场景结果图

表 4 算法在 GRFootball 地图上的平均回报

算法	Map		
	3 vs 1 with keeper	Counterattack easy	pass and shoot with keeper
QMIX	2.20	2.20	0.45
CW-QMIX	0.90	0.60	0.80
QTRAN	0.05	0.02	0.01
RIIT	0.80	0.90	0.05
UPDET	0.45	0.05	0.02
GREMARL	3.10	3.80	1.80

4.4 消融实验

为了展现模型各个组成部分对 GREMARL 结构与性能的必要性，在 SMAC 与 GRFootball 上的地图上展开消融实验。在 SMAC 的 5m vs 6m 地图上，分别关闭算法的自激励探索模块与图神经网络通信模块，采用同样的实验设置，最终得到如图 5 所示的消融实验结果图。其中实线代表五次运行结果的均值，阴影表示标准差。

如图 5 所示，完整的 GREMARL 模型（蓝线）在两个模块协同作用下展现出最优性能。其增量胜率在训练初期快速上升，最终稳定收敛于 0.65，评估环境中的胜率超过 0.8，平均 Episode 回报接近 19，显示出较强的策略收敛性与环境适应性。相比之下，关闭通信模块（橙线）后，尽管保留了自激励探索机制，智能体仍可学到一定的协作策略，但因缺乏高效的信息共享与状态协调，整体收敛速度显著下降，最终胜率与回报也出现明显下滑。进一步，关闭自激励探索模块（绿线）后，智能体策略熵长期维持在较高水平，表现出探索行为无效、目标不集中、死亡率上升等问题，最终导致训练停滞在较低胜率与回报区间，难以实现策略优化。

从行为层面观察，自激励模块的移除导致智能体缺乏明确驱动，在面对复杂对抗环境时易出现无序行动或被动等待，严重影响了整体生存率和探索效率。而通信模块的缺失则使得各智能体间缺乏

中国知网 https://www.cnki.net

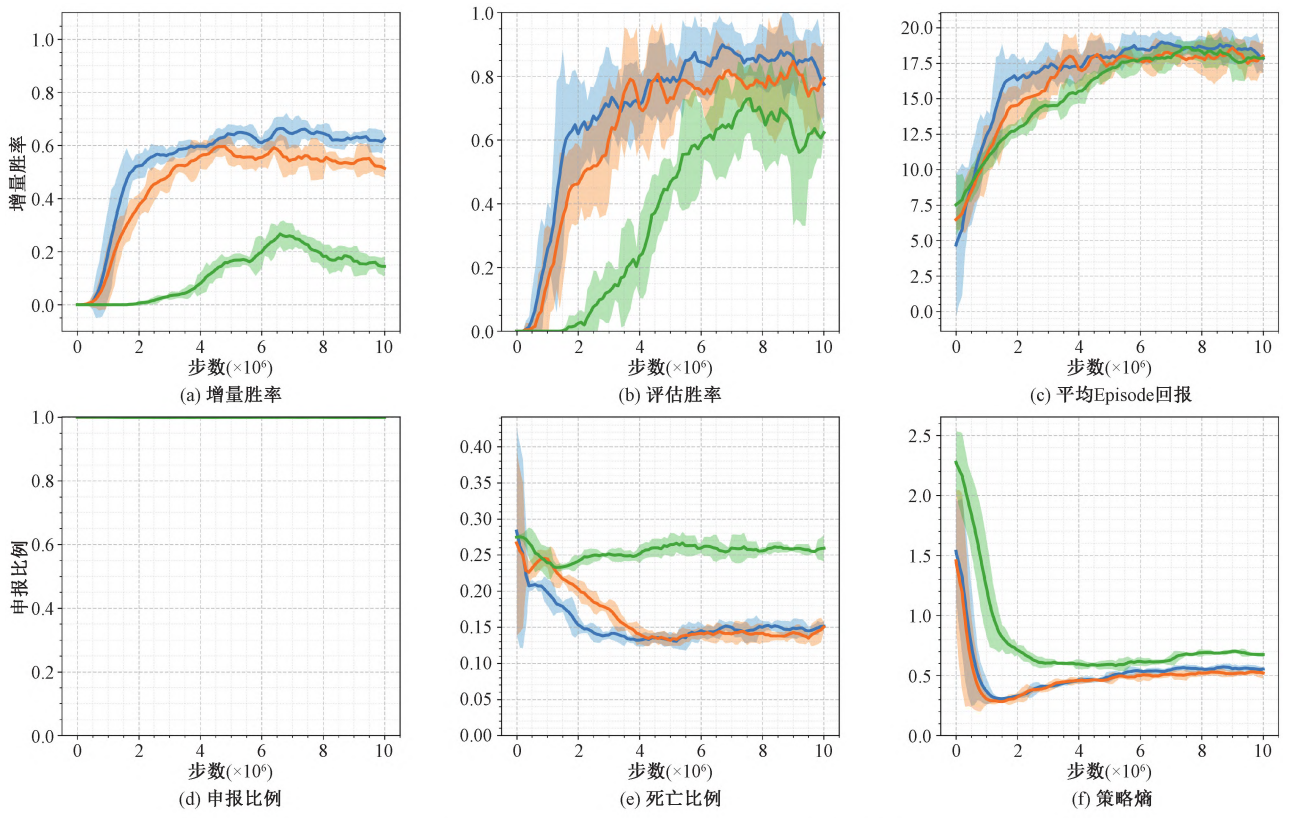

图 5　SMAC 消融实验结果图

信息协同,虽然在一定程度上仍可通过环境奖励驱动学习局部策略,但整体团队协作能力受限,表现为回报上限降低与策略更新缓慢。

此图说明,GREMARL 中的通信模块与自激励模块在功能上形成互补。通信机制主要促进智能体间的任务分工与协作协调,加速策略形成与收敛过程。而自激励机制则为智能体提供持续的探索动机和生存驱动,有效避免陷入局部最优。两者结合不仅提升了协作策略的整体水平,也显著改善了训练效率与最终性能,验证了其在多智能体强化学习中的结构必要性与实际有效性。

图 6 为 GREMARL 模型在 GRFootball 中 3 vs 1 with keeper 上的消融曲线,完整的 GREMARL 算法(蓝色曲线)在奖励获取、收益稳定性、损失收敛性等方面,显著优于“移除通信模块”(橙色曲线)和“移除自激励模块”(绿色曲线)的对比方案 。具体而言,在奖励与回报维度,GREMARL 的平均奖励、平均 Episode 回报及最大回报均更高,且收敛后波动更小,体现出更强的收益增长能力与稳定性。在损失维度,其策略损失收敛更快、波动更平缓,值损失也更快降至更低水平,反映出算法对策略优化和状态价值估计的精准性。这一结果验证了通信模块与自激励模块的协同作用。二者缺失会明显削弱算法性能,而完整集成则让 GREMARL 在复杂决策场景中表现更卓越。

4.5　可视化分析

如图 7 所示,本文收集了 GREMARL 算法在 5m vs 6m 以及 10m vs 11m 地图中梯度范数、值损失以及策略损失的变化情况,使用 t-SNE 方法[39]进行了可视化。结果表明,大部分点的策略损失和值损失都比较集中,说明算法在大部分样本上的表现比较稳定。而随着地图难度的增加,在 10m vs 11m 地图中,梯度范数分布较为分散,有更多黄色点(表示较大的梯度范数)和紫色点(表示较小的梯度范数),说明损失函数对参数变化较为敏感。

5　总　结

综上所述,在协作多智能体强化学习中,探索不足与通信低效始终制约着整体协同性能。针对这一瓶颈,本文提出的 GREMARL 方法通过融合自激励探索(SME)与图神经网络通信模块,从内而外提升探索与信息传播效率。首先通过图神经网络通信模块将智能体群视作动态图结构,节点间的连边

中国知网　https://www.cnki.net

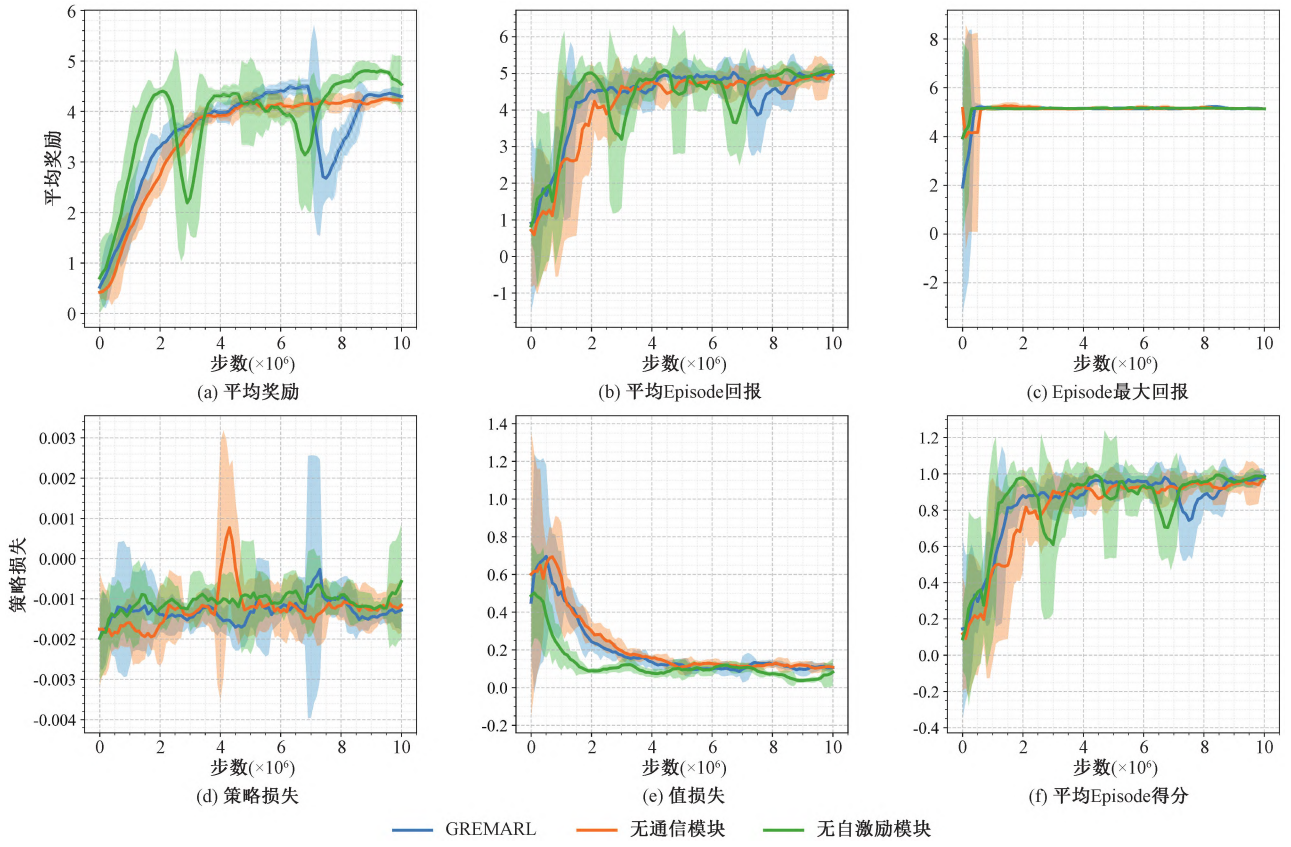

图 6 GRFootball 消融实验结果图

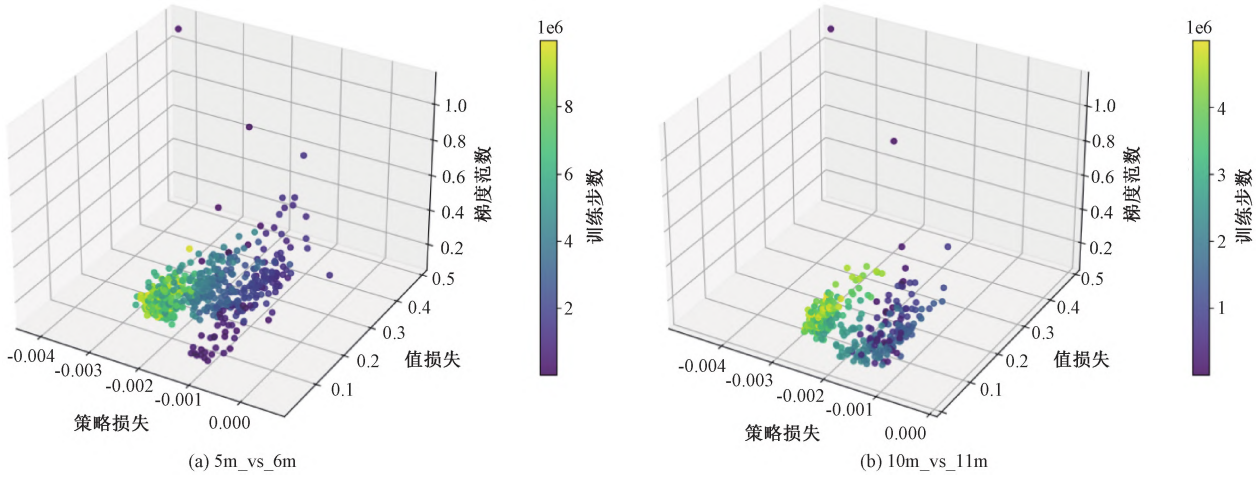

图 7 t-SNE 可视化

刻画距离、任务关联及历史交互信息，借助时序图神经网络端到端地学习最优通信拓扑，使关键信息快速聚合与分发、冗余通道自适应衰减。其次通过自激励探索方法将状态-动作对熵增量作为内在奖励，并引入历史行为与多样性指标，驱动智能体在早期广度探索、后期平滑向深度利用过渡，避免陷入局部最优。

在 SMAC 与 GRFootball 等复杂场景任务中的实验结果证明，GREMARL 在 SMAC 中的平均胜率达 88.8%，较目前最优算法提升 16.8%。消融研究进一步验证了 SME 模块与 GNN 通信的必要性。

在未来的研究中，我们将关注不同架构下的多智能体通信低效与探索不足等问题。此外，本文实验受限于硬件水平，主要考虑智能体数量较少的场景，而大规模多智能体强化学习任务下的建模通信与探索平衡同样值得关注，未来我们将继续关注该问题。

中国知网 https://www.cnki.net

参 考 文 献

[1] Sutton R S, Barto A G. Reinforcement learning: An introduction. Cambridge, USA: MIT Press, 2018

[2] Ding Shi-Fei, Du Wei, Zhang Jian, et al. Research progress on multi-agent deep reinforcement learning. Chinese Journal of Computers, 2024, 47(7):1547-1567 (in Chinese)
(丁世飞,杜威,张健,等.多智能体深度强化学习研究进展.计算机学报, 2024, 47(7):1547-1567)

[3] Gu S, Kuba J G, Chen Y, et al. Safe multi-agent reinforcement learning for multi-robot control Artificial Intelligence, 2023, 319: 103905

[4] Chen W T, Nguyen M, Li Z, et al. Decentralized Navigation of a Cable-Towed Load using Quadrupedal Robot Team via MARL arXiv preprint arXiv:2503.18221, 2025

[5] Mao F, Li Z, Lin Y, et al. Mastering arterial traffic signal control with multi-agent attention-based soft actor-critic model IEEE Transactions on Intelligent Transportation Systems, 2022, 24(3): 3129-3144

[6] Wu Q, Wu J, Shen J, et al. Distributed agent-based deep reinforcement learning for large scale traffic signal control. Knowledge-Based Systems, 2022, 241: 108304

[7] Cui J, Liu Y, Nallanathan A. Multi-agent reinforcement learning-based resource allocation for UAV networks. IEEE Transactions on Wireless Communications, 2019, 19(2): 729-743

[8] Wang B, Gao X, Xie T. An evolutionary multi-agent reinforcement learning algorithm for multi-UAV air combat. Knowledge-Based Systems, 2024, 299: 112000

[9] Hao J, Yang T, Tang H, et al. Exploration in deep reinforcement learning: From single-agent to multiagent domain. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(7): 8762-8782

[10] Calzolari G, Sumathy V, Kanellakis C, et al. D-MARL: A dynamic communication-based action space enhancement for multi agent reinforcement learning exploration of large scale unknown environments//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates, 2024: 3470-3475

[11] Xiao B, Li R, Wang F, et al. Stochastic graph neural network-based value decomposition for marl in internet of vehicles. IEEE Transactions on Vehicular Technology, 2023, 73(2): 1582-1596

[12] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[13] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296, 2017

[14] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning. Journal of Machine Learning Research, 2020, 21(178): 1-51

[15] Wang J, Ren Z, Liu T, et al. Qplex: Duplex dueling multi-agent q-learning. arXiv preprint arXiv:2008.01062, 2020

[16] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA: 2018, 32(1)

[17] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems, 2017, 30

[18] Hu S, Zhu F, Chang X, et al. Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers. arXiv preprint arXiv:2101.08001, 2021

[19] Hu J, Jiang S, Harding S A, et al. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. arXiv preprint arXiv:2102.03479, 2021

[20] Zhu Y , Shi E , Liu Z ,et al. Multi-agent reinforcement learning-based joint precoding and phase shift optimization for RIS-aided cell-free massive MIMO systems. IEEE Transactions on Vehicular Technology, 2024. DOI:10.1109/TVT.2024.3392883

[21] Taghavi, Mojtaba. Quantum computing and neuromorphic computing for safe, reliable, and explainable multi-agent reinforcement learning: Optimal control in autonomous robotics. arXiv, 2024, arXiv:2408.03884

[22] Lowe, Ryan, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in Neural Information Processing Systems, 2017, 30: 6379-90

[23] Foerster J, Nardelli N, Farquhar G, et al. Stabilising experience replay for deep multi-agent reinforcement learning//International Conference on Machine Learning. Sydney, Australia, 2017: 1146-1155

[24] Busoniu L, Babuska R, De Schutter B. Multi-agent reinforcement learning: A survey//2006 9th International Conference on Control, Automation, Robotics and Vision. Singapore, 2006: 1-6

[25] Lin Y, Li W, Zha H, et al. Information design in multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2023, 36: 25584-25597

[26] Zhao T, Chen T, Zhang B. QMIX-GNN: A graph neural network-based heterogeneous multi-agent reinforcement learning model for improved collaboration and decision-making. Applied Sciences, 2025, 15(7): 3794

[27] Houthooft, Rein, et al. VIME: Variational information maximizing exploration. Advances in Neural Information Processing Systems, 2016,29:1109-17

[28] Burda Y, Edwards H, Storkey A, et al. Exploration by random network distillation. arXiv preprint arXiv:1810.12894, 2018

[29] Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction//International Confer-

中国知网　https://www.cnki.net

ence on Machine Learning. Sydney, Australia, 2017: 2778-2787

[30] Bellemare, Marc, et al. Unifying count-based exploration and intrinsic motivation. Advances in Neural Information Processing Systems, 2016, 29:1471-79

[31] Foerster, Jakob, et al. Learning to Communicate with deep multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2016, 29:2137-45

[32] Zhang W, Liu H, Han J, et al. Multi-agent graph convolutional reinforcement learning for dynamic electric vehicle charging pricing//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA: 2022: 2471-2481

[33] Vezhnevets A S, Osindero S, Schaul T, et al. Feudal networks for hierarchical reinforcement learning//International Conference on Machine Learning. Sydney, Australia, 2017: 3540-3549

[34] Samvelyan M, Rashid T, De Witt C S, et al. The starcraft multi-agent challenge. arXiv preprint arXiv:1902. 04043, 2019

[35] Kurach K, Raichuk A, Stańczyk P, et al. Google research football: A novel reinforcement learning environment//Proceedings of the AAAI conference on artificial intelligence. New York, USA: 2020, 34(04): 4501-4510

[36] Li Z, Zhang R, Wang Z, et al. LLM-guided decision-making toolkit for multi-agent reinforcement learning. Neurocomputing, 2025: 638:1301051

[37] Rashid T, Farquhar G, Peng B, et al. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2020, 33: 10199-10210

[38] Son K, Kim D, Kang W J, et al. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning//International Conference on Machine Learning. Long Beach,USA, 2019: 5887-5896

[39] Cieslak M C, Castelfranco A M, Roncalli V, et al. t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. Marine Genomics, 2020, 51: 100723

CAO Yu-Kang, Ph. D. candidate. His main research interests include deep reinforcement learning, multi-agent reinforcement learning and computer vision.

LIU Quan, Ph. D., professor, Ph. D. supervisor. His research interests include deep reinforcement learning and automated reasoning.

LIU Hong-Zhe, master candidate. His research interests include deep reinforcement learning and multi-agent reinforcement learning.

YOU Ren-Yang, Ph. D. candidate. His research interests include deep reinforcement learning and inverse reinforcement learning.

Background

In real-world application scenarios, multi-agent systems often face issues of limited communication and insufficient exploration. To address these challenges, this paper proposes a novel collaborative multi-agent reinforcement learning framework that combines self-motivated multi-agent exploration (Self-Motivated Exploration, SME) with graph neural network-based (Graph Neural Network, GNN) multi-agent communication methods. Specifically, the paper designs an internal incentive mechanism to encourage agents to proactively explore unknown areas and uses graph neural networks to model the communication structure between agents, thereby capturing dynamic interaction relationships and promoting efficient information transmission. Experimental results show that GREMARL achieves an average win rate of 88.8% in complex SMAC environments, showcasing a significant 16.8% improvement over the suboptimal algorithm UPDET. Through carefully designed ablation studies, this paper has validated from multiple perspectives the necessity of self-motivated exploration and graph neural network modeling for GREMARL. This paper was supported by the National Natural Science Foundation of China (62376179, 62176175); The National Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A238); and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). These projects aim to enrich deep reinforcement learning theory and develop efficient algorithms to significantly enhance their computational power and applicability across diverse domains.

中国知网 https://www.cnki.net