基于双视角建模的多智能体协作强化学习方法

刘 全[1),2)] 施眉龙[1)] 黄志刚[1)] 张立华[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 在多智能体协作领域,强化学习算法通过共享智能体的局部信息来实现智能体间的协作.但共享协作机制极易引发过度协作问题,导致智能体忽视自身局部观测信息,丧失策略多样性,最终陷入低效协作的困境.为了解决该问题,本文提出基于双视角建模的多智能体协作强化学习方法(Bi-View Modeling Collaborative Multi-Agent Reinforcement Learning,简称BVM-CMARL).该方法从局部和全局两个视角对智能体进行建模,分别用于产生多样性的策略和激励协作.在局部视角最大化局部变分与自身轨迹的互信息,激励智能体的策略多样性;同时在全局视角最大化全局变分与其他智能体动作的互信息,提高智能体协作水平.最后将局部变分训练出的局部Q值与全局变分训练出的全局Q值合并,避免低效协作.将BVM-CMARL算法应用于星际争霸多智能体挑战赛(StarCraft Multi-Agent Challenge, SMAC)中的等级觅食(Level-Based Foraging, LBF)和走廊(Hallway)等环境,与QMIX、QPLEX、RODE、EOI和MAVEN等5种目前优秀的强化学习算法相比,BVM-CMARL算法具有更好的稳定性和性能表现,在SMAC上的平均胜率为82.81%,比次优算法RODE高13.42%.通过设计模型变体,在消融实验中证明了双视角建模对BVM-CMARL的必要性.

关键词 深度强化学习;多智能体系统;多智能体协作;协作建模;对比学习

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.01582

Multi-Agent Collaborative Reinforcement Learning Method Based on Bi-View Modeling

LIU Quan[1),2)] SHI Mei-Long[1)] HUANG Zhi-Gang[1)] ZHANG Li-Hua[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 recent years, there have been notable advancements in artificial intelligence technology, solidifying its crucial role in a wide array of real-world applications. Among the branches of artificial intelligence, reinforcement learning shines as a key discipline adept at tackling complex sequential decision-making challenges and playing a vital role in tasks related to control. By harnessing the progress made in neural network theory and computational power, deep reinforcement learning has revolutionized conventional reinforcement learning algorithms, smoothly integrating deep learning techniques into the decision-making frameworks of agents. For instance, Deep Q-Learning (DQN) is a prime illustration of this progress, employing a convolutional neural network to analyze visual inputs from Atari 2600 games and subsequently adjusting the policy of the reinforcement learning algorithm. Complex deep reinforcement

收稿日期:2023-09-03;在线发布日期:2024-04-19. 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资助.**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. **施眉龙**,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. **黄志刚**,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、分层强化学习. **张立华**,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、逆向强化学习.

learning tasks often entail multiple agents and are consequently formulated as multi-agent reinforcement learning, a framework that has demonstrated remarkable success across various domains, such as traffic control, sensor networks, gaming AI. In multi-agent reinforcement learning, agents can learn to collaborate through the Centralized Training with Decentralized Execution (CTDE) mechanism. In CTDE mechanism, reinforcement learning algorithms are able to realize cooperative behavior between agents through the sharing of local information between them as part of the cooperation process. As a result of this shared cooperation mechanism, complex multi-agent tasks can be solved in many fields, but the problem that arises at the same time is that excessive cooperation between the agents can lead to a conflict. There is a consequence of this in that agents begin to overlook the use of their current local observation information in cooperative efforts, losing the diversity of policy options, and eventually becoming inefficiently collaborating. Aiming at this problem, we propose a Bi-View Modeling Collaborative Multi-Agent Reinforcement Learning (BVM-CMARL) method. The method models agents from both local and global perspectives for generating a diversity of strategies and incentivizing collaboration, respectively. In the local view, the mutual information between local variation and its own trajectory is maximized, and then the agent's policy diversity is stimulated. The enhancement in agents' collaboration level is attributed to mutual information among their actions. Subsequently, a fusion of the locally trained Q value derived from local variations and the globally trained Q value derived from global variables is implemented to overcome the challenge posed by ineffective cooperation. The BVM-CMARL algorithm along with four distinguished multi-agent reinforcement learning algorithms are deployed across a spectrum of environments including the StarCraft Multi-Agent Challenge (SMAC), Level-Based Foraging (LBF), and Hallway scenarios to evaluate their efficacy and performance. The experimental findings demonstrate that the BVM-CMARL algorithm exhibits superior stability and performance in comparison to four state-of-the-art reinforcement learning algorithms, namely QMIX, QPLEX, RODE, EOI, and MAVEN. The average success rate achieved on the StarCraft Multi-Agent Challenge (SMAC) stands at 82.81%, showcasing a significant 13.42% improvement over the suboptimal algorithm RODE. Furthermore, the robustness and effectiveness of bi-view modeling are verified by ablation experiments and hyperparameter sensitivity experiments. In addition, a visualization analysis was developed and used to intuitively illustrate the role of BVM-CMARL.

Keywords　deep reinforcement learning; multi-agent system; multi-agent collaboration; collaborative modeling; contrastive learning

1　引　言

作为机器学习领域的一个重要分支,强化学习(Reinforcement Learning,RL)将任务建模为马尔可夫决策过程(Markov Decision Process,MDP),成功地解决了复杂的序贯决策问题.近年来,深度强化学习(Deep Reinforcement Learning,DRL)将深度学习引入智能体(agent)的决策过程中,给强化学习算法带来极大突破.

现实的深度强化学习任务通常涉及到多个智能体,因而被建模为多智能体强化学习(Multi Agent Reinforcement Learning,MARL).多智能体强化学习的目标是学习有效的策略,以控制多个智能体,最大化给定任务的累积奖赏.近年来,基于协作的多智能体强化学习(Collaborative Multi-Agent Reinforcement Learning, CMARL)在交通控制[1,2]、传感器网络[3,4]、游戏 AI[5,6]等领域取得了较好的表现,是多智能体系统中颇具前景的一种学习范式.在 CMARL 中,联合状态-动作空间随智能体

个数的增加呈指数式增长，其中集中式训练-分散式执行架构(Centralized Training with Decentralized Execution，CTDE)由于避免了对完整状态-动作空间的搜索，较大程度上提升了强化学习算法的效率，因而成为当前 CMARL 领域最常用的架构.

基于 CTDE 的 CMARL 算法可以分为基于值函数、基于策略梯度和基于通信 3 类算法[7]. 为了实现智能体间的协作，这 3 类算法的集中训练阶段都采取了共享局部信息的机制. 在这种共享的机制下，智能体可能会由于过度关注与环境中其他智能体的协作而忽视对自身的局部信息的利用，这种现象被称为过度协作[8]. 陷入过度协作的智能体会与其他智能体在策略上趋于一致，而这些丧失策略多样性的智能体很难在复杂的 CMARL 任务上取胜.

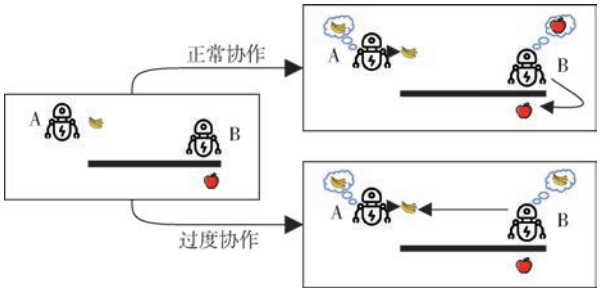


图 1 正常协作与过度协作情形示例

为了进一步说明过度协作的危害，在图 1 中的觅食场景中对比正常协作和过度协作的效果. 智能体 A 和 B 在正常协作的情形下，应该就近选取各自觅食的目标. 但在过度协作的情形下，B 忽视了对自身局部信息的利用，导致其策略丧失多样性，因此选择去与 A 协作，并捕食与其距离较远的香蕉. 该场景说明，在复杂的多智能体任务中，过度协作的效率非常低下. 因此应尽量保持各智能体策略的多样性，从而避免陷入过度协作.

为了在协作中保持策略多样性，智能体需要在策略多样性和协作间做出权衡，因此如何平衡两者至关重要. 针对这一问题，本文提出一种双视角建模的多智能体协作强化学习(Bi-View Modeling Collaborative Multi-Agent Reinforcement Learning，简称 BVM-CMARL)方法. 该方法从局部和全局两个视角对智能体进行建模，分别用于激励智能体的策略多样性和协作，以求达到二者的平衡.

具体地，根据智能体当前时刻的观测值，BVM-CMARL 首先编码出局部变分和全局变分. 在局部视角采用对比学习最大化局部变分与自身轨迹的互信息，使得局部变分表征出智能体自身的局部观测信息，进而根据局部变分训练局部 Q 网络. 同时在全局视角最大化全局变分与其他智能体采取动作的互信息，使得全局变分有感知环境中其他智能体行为的能力，随后使用全局变分训练全局 Q 网络. 最后，将偏向于学习智能体自身局部信息的局部 Q 网络与偏向于拟合其他智能体行为的全局 Q 网络合并，达到平衡智能体策略多样性与协作的目的.

本文的主要贡献可以总结为以下 3 点：

(1) 提出一种基于双视角建模的多智能体强化学习方法，显式地将智能体分解为局部视角和全局视角，分别用于激励策略多样性和提高智能体协作水平.

(2) 设计了模型变体并展开消融实验，证明本文设计的双视角架构可以提高模型的性能.

(3) 将提出的 BVM-CMARL 用于星际争霸多智能体挑战[9](StarCraft Multi-Agent Challenge，简称 SMAC)、走廊[10](Hallway)和等级觅食[11](Level-Based Foraging，LBF)3 个多智能体任务. 实验结果证明了双视角建模对 BVM-CMARL 的必要性.

本文第 2 节总结了相关研究成果. 第 3 节回顾了本文所述模型需要的基础知识. 第 4 节详细介绍了本文提出的模型 BVM-CMARL，包括模型中的组成成分、实现细节和训练过程. 第 5 节为本文的实验部分，在三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 多智能体协作强化学习

近年来，多智能体协作强化学习算法主要分为基于值函数分解、基于行动者-评论家框架和基于通信 3 类. 本节将对这 3 类算法进行简要介绍.

值函数分解的主要思想是将联合值函数分解为各智能体的效用函数，这种思想首先在价值分解网络[12](Value Decomposition Networks，VDN)上被实现. VDN 假设联合值函数来自于各个智能体的值函数的加和，从而解决了部分可观测引发的懒惰智能体(lazy agent)问题. 在其基础上，Q 混合网络[13](简称 QMIX)采用了单调混合网络实现从各个智能体的值函数到联合值函数的非线性映射. 复式双 Q 学习网络[14](Duplex Dueling Multi-Agent Q-Learning，QPLEX)采用双层 Q 网络架构以符合个体-全局最大原则(Individual-Global-Max，IGM).

多智能体深度策略梯度算法[15](Multi Agent Deep Deterministic Policy Gradient,MADDPG)是第一种基于行动者-评论家框架的多智能体强化学习算法.在 MADDPG 中,集中的评论家会根据分散的行动者收集到的轨迹计算梯度信息.针对只有全局奖赏可知的情形,反事实多智能体策略梯度[16](Counterfactual Multi-Agent Policy Gradients,简称 COMA)提出在行动者-评论家框架中引入反事实推理来完成信度分配.多智能体异策略分解策略梯度[17](Off-Policy Multi-Agent Decomposed Policy Gradients,简称 DOP)引入了值函数分解的思想,解决了连续动作空间的问题.

此外,智能体间的通信也是近年来的研究热点之一.如近似值函数分解通信[10](Nearly Decomposable Value Functions Via Communication,简称 NDQ)引入通信范式最小化和信息瓶颈来优化智能体间的通信.在其基础上,多智能体激励通信[18](Multi Agent Incentive Communication,简称 MAIC)通过显式通信来激励每个智能体生成对环境中其他智能体的激励信息,从而由克服部分可观测的环境带来不稳定问题.

2.2　多智能体强化学习中的协作与策略多样性

多智能体协作的思想来源于人类社会中的师生关系,在多智能体系统中,智能体通过拟合环境中其他智能体的行为来获得对环境的一致认知.已有诸多研究致力于激励智能体拟合其他智能体的策略,如邻域认知一致[19](Neighborhood Cognition Consistent,简称 NCC)首先采用多智能体任务的先验知识对智能体进行邻域划分,并使用变分自编码器维持邻域内所有智能体的认知一致性.Chen 等人[20]提出的协作信号构建方法(Signal Instructed Coordination,简称 SIC)则是通过最大化智能体的行为与全局策略的互信息来促进协作.角色分解[21](Roles to Decompose,简称 RODE)采用预训练的方法对智能体的角色进行分组并限制智能体在角色内进行协作.

在单智能体的情形下,策略多样性主要以探索的方式实现,现有方法包括基于变分推断[22,23]和好奇心驱动[24,25]方法.但是在多智能体强化学习中,提高智能体策略多样性的同时还必须兼顾与其他智能体的协作.多智能体变分探索[26](Multi-Agent Variational Exploration,简称 MAVEN)首先针对 QMIX 由于单调性约束而无法进行有效的探索的问题,让每个智能体的行为取决于由分层策略控制的共享潜在变量,进而实现智能体的探索.Jiang 等人[27]提出的个性涌现方法(Emergence of Individuality,简称 EOI)学习了基于局部观测的概率分类器,通过该分类器鼓励智能体访问较熟悉的观测,间接实现策略多样性.渐进式互信息协作[28](Progressive Mutual Information Collaboration,简称 PMIC)提出了一个互信息协作标准,将动作与状态之间的互信息作为伪奖励,提高对局部信息的利用效果.多样性共享[29](Celebrating Diversity in Shared,简称 CDS)认为参数共享机制给复杂的多智能体任务带来了瓶颈,因此提出基于互信息的策略多样性学习机制.相对于已有工作,本文所提方法不预先限制智能体协作的方式,而是直接将智能体分解为局部视角和全局视角,从而在协作中完成策略多样性的学习.

3　背景知识

本文针对完全协作的多智能体强化学习场景[30],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,简称 Dec-POMDP).Dec-POMDP 可以描述为九元组 $G=<I,S,A,P,R,\Omega,O,n,\gamma>$,其中 I 代表智能体集合,每个智能体 i 根据观测函数 $O(s,i)$ 从环境的真实状态 $s\in S$ 中获取其自身观测 $o_i\in\Omega$.当前时刻 n 个智能体采取的动作组成了联合动作 $a\in A^n$,联合动作与环境交互后得到奖赏 $r=R(s,a)$,进而根据转移函数 $P(s'\mid s,a)$ 转移到下一状态 s'.根据联合策略 π 可得到多智能体系统当前总 Q 值 $Q_{tot}^{\pi}(\tau,a)=\mathbb{E}_{s,a,\pi}[\sum_{t=0:\infty}\gamma^t r_t]$,其中 τ 是联合观测-动作轨迹,$\gamma\in[0,1)$是折扣因子.

此外,本文所提模型基于 CTDE 架构.在集中训练阶段,混合网络首先根据所有智能体的 Q 值计算总 Q 值,再根据总 Q 值计算 TD 损失.在分散执行阶段,混合网络则不再生效,各个智能体独立地使用已训练好的策略与环境交互.

4　BVM-CMARL

BVM-CMARL 的基本流程如图 2 所示.绿色部分为局部视角建模.智能体 i 的观测 o_i^t 经过局部编码器,生成一个高斯分布,采样出局部变分 $\rho_{loc}^{t,i}$ 后,用于学习该智能体的局部信息,随后将其解码得到局部 Q 值 Q_{loc}.局部视角建模通过最大化 $\rho_{loc}^{t,i}$ 与当前

轨迹 τ_i^t 之间的互信息，使 $\rho_{loc}^{t,i}$ 学到尽可能多的局部信息，进而激励智能体学得多样性的策略. 橙色部分为全局视角建模. 智能体 i 的观测 o_i^t 经过全局编码器，生成了 $n-1$ 个高斯分布，分别采样得到感知变分后，用于拟合其他智能体的行为. 随后使用全局解码器，将感知变分整合为一个全局变分 $\rho_{glo}^{t,i}$，进而得到全局 Q 值 Q_{glo}. 全局视角建模定义了全局损失，用于最大化感知变分与对应智能体采取的真实动作之间的互信息，使得全局变分具有感知环境中其他智能体行为的能力.

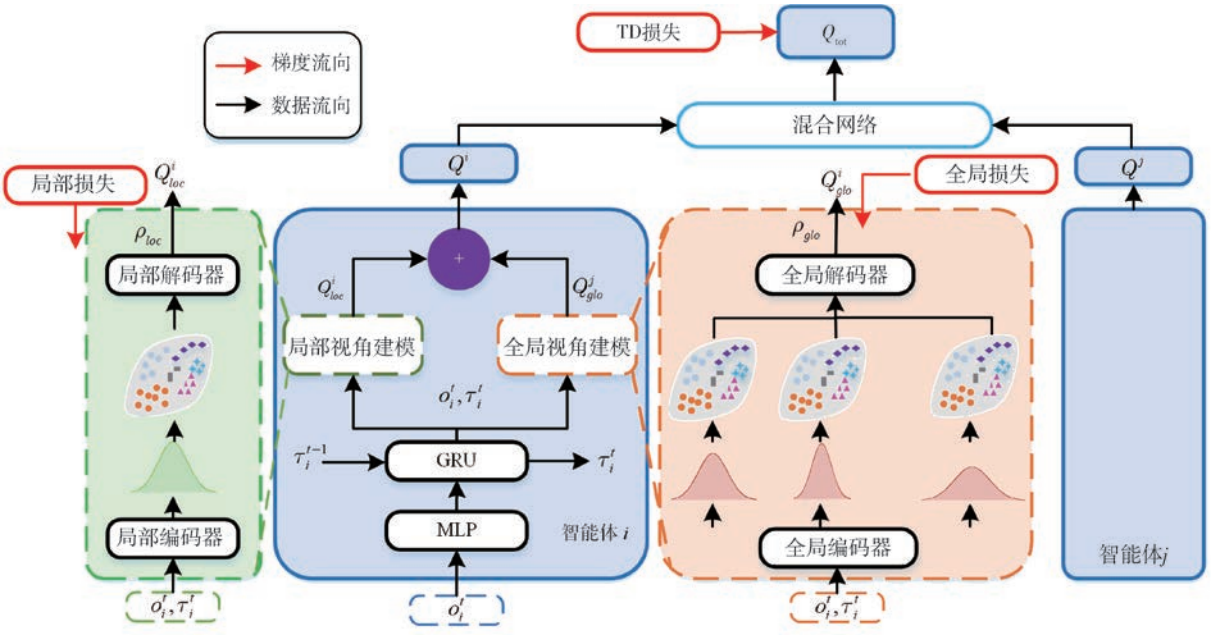


图 2 BVM-CMARL 模型框架图

完成局部和全局视角建模后，将智能体的局部和全局 Q 值合并，得到该智能体的 Q 值. 最后将所有智能体的 Q 值输入混合网络，计算总 Q 值 Q_{tot} 和对应的 TD 损失，完成整个训练过程.

在训练中，模型能否有效平衡智能体的策略多样性和协作，关键在于 ρ_{loc} 和 ρ_{glo} 应能否高效学得智能体所在视角的信息. 因此，我们提出 ρ_{loc} 和 ρ_{glo} 应满足以下两条性质：

(1)动态性：ρ_{loc} 和 ρ_{glo} 都可以随时间自适应环境的动态变化；

(2)丰富性：ρ_{loc} 和 ρ_{glo} 分别可以自动提取出丰富的局部信息和全局信息.

4.1 多智能体局部视角建模

为了使局部变分满足动态性，进而适应环境的动态变化，本文首先使用局部编码器 E_{loc} 对当前时刻的局部观测 o^t 编码，生成一个多变量高斯分布 $N(\mu_{loc}^t,\sigma_{loc}^t)$ 的均值和方差，再从 $N(\mu_{loc}^t,\sigma_{loc}^t)$ 中采样出当前时刻的局部变分 ρ_{loc}^t，如式(1)所示：

$$\begin{aligned}(\mu_{loc}^t,\sigma_{loc}^t) &= E_{loc}(o^t)\\ \rho_{loc}^t &\sim N(\mu_{loc}^t,\sigma_{loc}^t)\end{aligned} \tag{1}$$

其中，局部编码器由一个多层感知机构成. 使用当前时刻的局部观测编码出的 ρ_{loc}^t 会随着时间快速变化，因此无法提取出稳定的局部信息. 在满足动态性的同时，为了提取出相对长时间内的局部信息，局部损失考虑最大化 ρ_{loc}^t 与当前时刻的轨迹 τ^t 之间的互信息，将优化目标定为

$$\max: I(\rho_{loc}^{t,i},\tau_i^t) \tag{2}$$

其中轨迹 τ^t 使用 GRU 单元获得. 本文采用对比学习[31]来优化式(2)中无法直接计算的互信息目标. 其中正例采样于两者在经验回放池 D 中的联合分布，负例采样于两者在 D 中的边缘分布，即首先采样出 $\rho_{loc}^{t,i}$ 再独立地采样出 τ_i^t. 对于式(2)中互信息目标的优化，给出定理 1.

定理 1. 对于智能体 i，其当前时刻的局部变分 $\rho_{loc}^{t,i}$ 与轨迹 τ_i^t 之间的互信息具有证据下界，即

$$I(\rho_{loc}^{t,i},\tau_i^t) \geqslant \mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}\,\mathbb{E}_{\tau_i^-}\left[\log\frac{\exp(h(\rho_{loc}^{t,i},\tau_i^t))}{\sum\limits_{\tilde{\tau}_i\in\tau_i^-\cup\tau_i^t}\exp(h(\rho_{loc}^{t,i},\tilde{\tau}_i))}\right] \triangleq I_{\text{nce}} \tag{3}$$

其中 h 为相似性评分函数，该函数会对正例评分较高，负例评分较低.

证明. 考虑引入隐变量 C 来区分正例和负例,C 指示局部变分 $\rho_{loc}^{t,i}$ 与当前时刻的轨迹 τ_i^t 采样于联合分布 $p(\rho_{loc}^i,\tau_i)$(当 C 为 1)或边缘分布 $p(\rho_{loc}^{t,i})p(\tau_i^t)$(当 C 为 0),即

$$\begin{cases} p(\rho_{loc}^{t,i},\tau_i^t \mid C=1)=p(\rho_{loc}^{t,i},\tau_i^t) \\ p(\rho_{loc}^{t,i},\tau_i^t \mid C=0)=p(\rho_{loc}^{t,i})p(\tau_i^t) \end{cases} \tag{4}$$

考虑现有 1 个采样于联合分布 $p(\rho_{loc}^i,\tau_i)$ 的正例和 N 个采样于边缘分布的负例,得到隐变量 C 的分布

$$\begin{cases} p(C=1)=1/(N+1) \\ p(C=0)=N/(N+1) \end{cases} \tag{5}$$

对 $C=1$ 的后验概率使用贝叶斯定理,有

$$\begin{aligned} &\log p(C=1 \mid \rho_{loc}^{t,i},\tau_i^t) \\ &=\log\frac{p(C=1)p(\rho_{loc}^{t,i},\tau_i^t)}{p(C=0)p(\rho_{loc}^{t,i})p(\tau_i^t)+p(C=1)p(\rho_{loc}^{t,i},\tau_i^t)} \\ &=\log\frac{p(\rho_{loc}^{t,i},\tau_i^t)}{Np(\rho_{loc}^{t,i})p(\tau_i^t)+p(\rho_{loc}^{t,i},\tau_i^t)} \\ &\leqslant -\log N+\log\frac{p(\rho_{loc}^{t,i},\tau_i^t)}{p(\rho_{loc}^{t,i})p(\tau_i^t)} \end{aligned} \tag{6}$$

由于局部变分 $\rho_{loc}^{t,i}$ 与当前时刻轨迹 τ_i^t 的互信息定义为

$$I(\rho_{loc}^{t,i},\tau_i^t)=\mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}\left[\log\frac{p(\rho_{loc}^{t,i},\tau_i^t)}{p(\rho_{loc}^{t,i})p(\tau_i^t)}\right] \tag{7}$$

对式(6)两边取期望,整理得到

$$\begin{aligned} &I(\rho_{loc}^{t,i},\tau_i^t) \\ &\geqslant \log N+\mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}[\log p(C=1 \mid \rho_{loc}^{t,i},\tau_i^t)] \end{aligned} \tag{8}$$

考虑一个相似性评分函数 h,当 h 具有足够的表征能力时,最优相似性评分函数满足

$$h^*(\rho_{loc}^{t,i},\tau_i^t)=p(C=1 \mid \rho_{loc}^{t,i},\tau_i^t) \tag{9}$$

因此,有

$$\begin{aligned} &I(\rho_{loc}^{t,i},\tau_i^t) \\ &\geqslant \log N+\mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}[\log p(C=1 \mid \rho_{loc}^{t,i},\tau_i^t)] \\ &=\log N+\mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}[\log h^*(\rho_{loc}^{t,i},\tau_i^t)] \\ &\geqslant \log N+\mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}[\log h^*(\rho_{loc}^{t,i},\tau_i^t)- \\ &\quad \log\sum_{\tilde{\tau}_i\in\tau_i^-\cup\tau_i^t}\exp(h(\rho_{loc}^{t,i},\tilde{\tau}_i))] \\ &=\log N+I_{nce}(h^*) \\ &=\log N+\max_h I_{nce}(h) \\ &\geqslant \log N+I_{nce}(h) \\ &\geqslant I_{nce} \end{aligned} \tag{10}$$

证毕.

根据定理 1 我们得到了证据下界 I_{nce},因此确定局部视角的局部对比损失为最大化 I_{nce}

$$\begin{aligned} &\mathcal{L}_{loc}(\theta_{loc}) \\ &=-\sum_N \mathbb{E}_{(\rho_{loc}^{t,i},\tau_i^t)\sim D}\,\mathbb{E}_{\tau_i^-}\left[\log\frac{\exp(h(\rho_{loc}^{t,i},\tau_i^t))}{\sum_{\tilde{\tau}_i\in\tau_i^-\cup\tau_i^t}\exp(h(\rho_{loc}^{t,i},\tilde{\tau}_i))}\right] \end{aligned} \tag{11}$$

其中 θ_{loc} 表示局部视角的所有参数. 随后将智能体 i 的局部变分 $\rho_{loc}^{t,i}$ 经过一层局部解码器 D_{loc}^i 后得到局部 Q 值 $Q_{loc}^i(\cdot,\rho_{loc}^{t,i},\tau_i^t)$.

4.2 多智能体全局视角建模

对于任意智能体 i,为了使其生成的全局变分满足动态性,BVM-CMARL 方法使用当前时刻的局部观测 o_i^t 并行地编码出个高斯分布,进而采样出 $n-1$ 个感知变分 $\{z_{i\to1}^t,z_{i\to2}^t,\cdots,z_{i\to N}^t\}$. 每个感知变分 $z_{i\to j}^t$ 用于学习智能体 i 对智能体 j 行为的感知,即

$$\begin{aligned} (\mu_{glo}^{t,i\to j},\sigma_{glo}^{t,i\to j})&=E_{glo}^{i\to j}(o_i^t) \\ z_{i\to j}^t &\sim N(\mu_{glo}^{t,i\to j},\sigma_{glo}^{t,i\to j}) \end{aligned} \tag{12}$$

其中全局编码器由一个多层感知机构成,为了使 $z_{i\to j}^t$ 具有感知其他智能体行为的能力,本文设置全局视角的目标为最大化每一个感知变分 $z_{i\to j}^t$ 与对应的智能体 j 采取的真实动作 a_j^t 之间互信息 $I(z_{i\to j}^t,a_j^t \mid o_j^t)$. 对于该互信息目标的优化,本文给出定理 2:

定理 2. 对于任意智能体 i,$z_{i\to j}^t$ 是 i 对其他智能体 j 的感知变分,a_j^t 是智能体 j 采取的真实动作,则互信息 $I(z_{i\to j}^t,a_j^t \mid o_j^t)$ 有证据下界

$$\begin{aligned} &I(z_{i\to j}^t,a_j^t \mid o_i^t) \\ &\geqslant \mathbb{E}_D[-D_{KL}(p(z_{i\to j}^t \mid o_i^t)\,\|\,q_\xi(z_{i\to j}^t \mid o_i^t,a_j^t))] \end{aligned} \tag{13}$$

其中,$q_\xi(z_{i\to j}^t \mid o_i^t,a_j^t)$ 是变分近似器,D 代表经验回放池.

证明. 根据互信息的定义,有

$$\begin{aligned} &I(z_{i\to j}^t,a_j^t \mid o_i^t) \\ &=\mathbb{E}_{z_{i\to j}^t,a_j^t,o_i^t}\left[\log\frac{p(z_{i\to j}^t \mid a_j^t,o_i^t)}{p(z_{i\to j}^t \mid o_i^t)}\right] \\ &=\mathbb{E}_{z_{i\to j}^t,a_j^t,o_i^t}\left[\log\frac{q_\xi(z_{i\to j}^t \mid a_j^t,o_i^t)}{p(z_{i\to j}^t \mid o_i^t)}\right] \\ &+\mathbb{E}_{a_j^t,o_i^t}[D_{KL}(p(z_{i\to j}^t \mid a_j^t,o_i^t)\,\|\,q_\xi(z_{i\to j}^t \mid a_j^t,o_i^t))] \end{aligned} \tag{14}$$

考虑到 KL 散度的非负性,有

$$\begin{aligned} &I(z_{i\to j}^t,a_j^t \mid o_i^t) \\ &\geqslant \mathbb{E}_{z_{i\to j}^t,a_j^t,o_i^t}\left[\log\frac{q_\xi(z_{i\to j}^t \mid a_j^t,o_i^t)}{p(z_{i\to j}^t \mid o_i^t)}\right] \\ &=\mathbb{E}_D[-D_{KL}(p(z_{i\to j}^t \mid o_i^t)\,\|\,q_\xi(z_{i\to j}^t \mid o_i^t,a_j^t))] \end{aligned} \tag{15}$$

证毕.

根据定理 2 得到了 $I(z_{i\to j}^t, a_j^t \mid o_j^t)$ 的证据下界，因此设置全局视角的损失函数为

$$\mathcal{L}_{glo}(\theta_{glo}) = \sum_{i\neq j} \mathbb{E}_D[D_{KL}(p(z_{i\to j}^t \mid o_i^t) \parallel q_\xi(z_{i\to j}^t \mid o_i^t, a_j^t))] \tag{16}$$

其中 θ_{glo} 表示全局视角的所有参数.

为了整合智能体的每一个感知变分，本文考虑采用注意力机制

$$\begin{cases} q_i = W_q o_i^t \\ k_{i,j} = W_k z_{i\to j}^t \\ v_{i,j} = W_v z_{i\to j}^t \end{cases} \tag{17}$$

其中，注意力机制中的查询 q_i 由智能体的观测 o_i^t 计算，键 $k_{i,j}$ 和值 $v_{i,j}$ 由全局感知变分 $z_{i\to j}^t$ 计算，W_q，W_k 和 W_v 均为注意力参数. 最后通过注意力函数合并所有感知变分，并经过一层全局解码器，得到智能体 i 的全局变分 ρ_{glo}^i

$$\rho_{glo}^i = D_{glo}^i\left(\sum_{j\neq i} Attn(q_i, k_{i,j}, v_{i,j})\right) \tag{18}$$

得到全局变分后，使用全局变分得到全局 Q 值 $Q_{glo}^i(\cdot, \rho_{glo}^{t,i}, \tau_i^t)$，完成全局视角建模的过程.

4.3 总目标

经过局部和全局视角建模后，我们得到了局部 Q 值 Q_{loc} 和全局 Q 值 Q_{glo}，将两者进行线性组合后得到每个智能体的 Q 值

$$Q^i = \lambda_Q Q_{loc}^i + (1-\lambda_Q) Q_{glo}^i \tag{19}$$

其中 λ_Q 是折扣因子，其作用是平衡两个视角的 Q 值的重要性. 而由于 BVM-CMARL 服从 CTDE 范式，为了计算总 TD 损失 $\mathcal{L}_{TD}$，需要将每个智能体的 Q 值输入混合网络，得到总 Q 值 Q_{tot}，根据 Q_{tot} 计算出的 TD 损失为

$$\mathcal{L}_{TD}(\theta) = \mathbb{E}_D[r + \gamma \max_{a'} Q_{tot}(\tau', a; \theta^-) - Q_{tot}(\tau, a; \theta)]^2 \tag{20}$$

其中，θ^- 是目标网络的参数. 本文采用 QMIX 作为混合网络，QMIX 也可以用其他网络代替. 除 TD 损失以外，结合前文所述的局部损失和全局损失，BVM-CMARL 的总目标定义为

$$\mathcal{L}(\theta) = \mathcal{L}_{TD}(\theta) + \lambda_{\mathcal{L}} \mathcal{L}_{loc}(\theta_{loc}) + (1-\lambda_{\mathcal{L}}) \mathcal{L}_{glo}(\theta_{glo}) \tag{21}$$

其中，$\lambda_{\mathcal{L}}$ 是折扣因子，其作用是平衡两个视角的损失函数的重要性.

BVM-CMARL 算法的训练流程如算法 1 所示，其中 7-8 行和 9－14 行分别表示局部视角建模和全局视角建模的学习过程. 在第 15 行，智能体对局部视角生成的局部 Q 值和全局视角生成的全局 Q 值进行线性组合以获取 Q 值，进而选择动作.

算法 1. BVM-CMARL 算法

1. 输入：每个智能体的动作 $a_i \in A$ 和观测 $o_i \in O$
2. 输出：Q_{tot}
3. 参数初始化
4. FOR 每个时间步 t DO
5. FOR 智能体 $i \in I$ DO
6. 使用 GRU 单元生成自身轨迹 τ_i^t
   // 局部视角建模：
7. 根据式(1)采样局部变分 $\rho_{loc}^{t,i}$
8. 获取局部 Q 值 $Q_{loc}^i(\cdot, \rho_{loc}^{t,i}, \tau_i^t)$
   // 全局视角建模
9. 根据式(12)采样感知变分 $\{z_{i\to 1}^t, z_{i\to 2}^t \cdots, z_{i\to N}^t\}$
10. $q_i = W_q o_i^t$
11. $k_{i,j} = W_k z_{i\to j}^t$，for $j \in I$
12. $v_{i,j} = W_v z_{i\to j}^t$，for $j \in I$
13. $\rho_{glo}^j = D_{glo}^j\left(\sum_{j\neq i} Attn(q_i, k_{i,j}, v_{i,j})\right)$，FOR $j \in I$
14. 获取全局 Q 值 $Q_{glo}^i(\cdot, \rho_{glo}^{t,i}, \tau_i^t)$
15. 线性组合获取 Q 值
    $Q^i = \lambda_Q Q_{loc}^i + (1-\lambda_Q) Q_{glo}^i$
16. 根据 Q^i 选取动作 a_i^t
17. 将 a_i^t 和 τ_i^t 存储到经验回放池
18. IF 当前处于集中训练阶段 DO
19. 将 Q^i 输入混合网络以获取 Q_{tot}
20. 根据式(21)优化目标
21. END IF
22. END FOR
23. END FOR

5 实 验

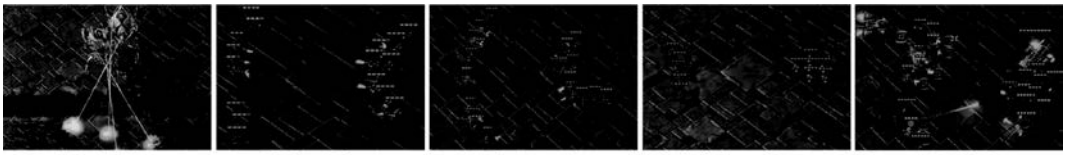
本节首先展开对比实验，以验证 BVM-CMARL 相对其他方法的有效性. 此外，进行消融实验和超参数敏感性实验来证明 BVM-CMARL 算法的有效性和鲁棒性. 最后，为了直观地展示双视角建模的效果，本文进行了可视化分析.

5.1 实验环境

本文采用的实验环境包括 SMAC、Hallway 和 LBF，它们的特点分别如下所示：

SMAC：如图 3 所示，SMAC 包含一系列星际争霸Ⅱ游戏，旨在评估智能体在解决复杂协作任务时的性能. 对比实验选取了三个难地图和两个超难地

图,消融实验选取了六个简单地图,具体细节如表 1 所示.智能体的动作空间由移动、攻击、治疗、停止和空操作 5 个离散动作组成.在每个地图中,己方算法控制的智能体与内置游戏 AI 控制的敌方智能体作战.当任意一方的智能体全部阵亡时结束,游戏取胜的条件是消灭所有敌方智能体.



(a) 2c vs 64zg　(b) 8m vs 9m　(c)10m vs 11m　(d) 5m vs 6m　(e) MMM2

图 3　选取的部分 SMAC 场景

表 1　SMAC 地图细节

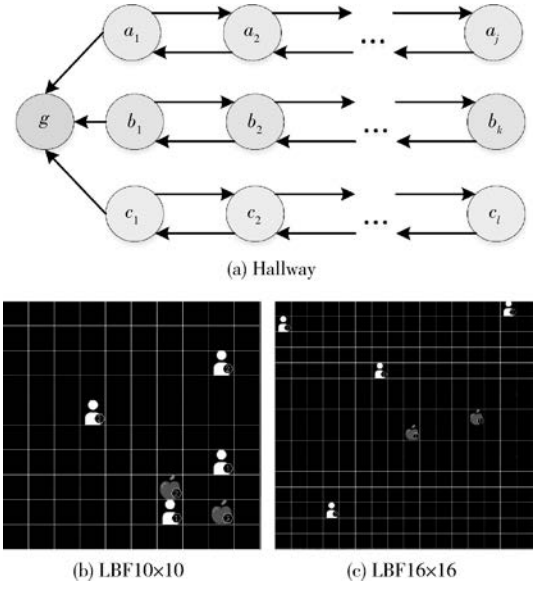
地图	难度	己方阵容	敌方阵容
1c3s5z	简单	1 个巨人 3 个伴随者 5 个跳虫	1 个巨人 3 伴随者 5 个跳虫
2s3z	简单	2 个伴随者 3 个跳虫	2 个伴随者 3 个跳虫
3m	简单	3 个海军陆战队	3 个海军陆战队
8m	简单	8 个海军陆战队	8 个海军陆战队
so many baneling	简单	7 个跳虫	32 个异虫
3s5z	简单	3 个伴随者 5 个跳虫	3 个伴随者 5 个跳虫
2c vs 64zg	难	2 个巨人	64 个跳虫
8m vs 9m	难	8 个海军陆战队	9 个海军陆战队
10m vs 11m	难	10 个海军陆战队	11 个海军陆战队
5m vs 6m	超难	5 个海军陆战队	6 个海军陆战队
MMM2	超难	1 个治疗者 2 个掠夺者 7 个海军陆战队	1 个治疗者 2 个掠夺者 8 个海军陆战队

Hallway:如图 4(a)所示,3 个智能体分别在 3 条走廊上随机初始化.每个智能体只能观测到其当前所在位置,并选择移动方位.只有当所有智能体同时到达终点时才被视为获胜.为了增大获胜难度,设置不同的走廊长度为 2、6 和 10.

LBF:如图 4(b)和(c)所示,LBF 是一个部分可观测的格子世界觅食游戏.每个情节开始时,智能体和食物的等级随机初始化.智能体的动作空间包括移动、装载食物和空操作 3 个动作.只有当一组智能体同时选择装载某个食物,且其等级之和大于待装载的食物等级时,才能成功获取该食物,并得到对应的奖赏.在本文选取的两个 LBF 环境中,4 个智能体分别在 10×10 和 16×16 的格子世界环境中觅食,食物设置为两份,观测范围设置为 2×2.

5.2　实验设置

本文选取了 QMIX、QPLEX、RODE、EOI 和 MAVEN 作为基准方法.其中 QMIX、QPLEX 和



(a) Hallway

(b) LBF10×10　(c) LBF16×16

图 4　LBF 和 Hallway 环境

RODE 是优秀的基于值函数的算法,EOI 和 MAVEN 是提升个体多样性的方法.

在实现 BVM-CMARL 算法时,各个智能体的观测值首先被输入一个全连接层和 64 维 GRU,以生成当前时刻轨迹,混合网络采用了 QMIX.为保证实验的公平性,所有算法采用的参数均与标准库 PyMARL 保持一致,关键参数值如表 2 所示.

表 2　实验使用的参数

参数名称	参数值	参数名称	参数值
批次大小	32	优化器	RMSprop
经验回放池容量	5000	折扣因子	0.99
隐层维度	64	学习率	5e-4

在相同实验环境下,每次实验对这四种算法进行 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 Intel Xeon E5-2680 v4 CPU、2 块 NVIDIA Tesla P40 GPU 和 128 GB 内存的服务器作为硬件环境.

5.3 对比实验

本节在 SMAC 和 Hallway 上展开对比实验，旨在证明 BVM-CMARL 相对于五个基准方法的有效性. 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 5 所示. 根据图 5 中的结果，BVM-CMARL 在六个任务上均取得了最佳胜率.

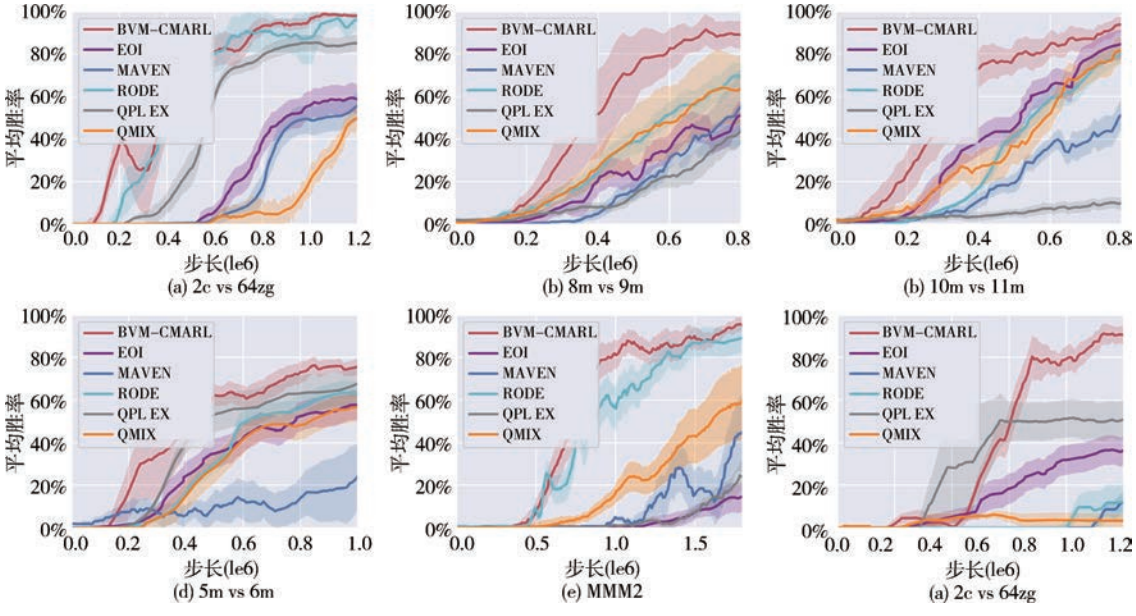


图 5 SMAC 任务对比实验训练曲线

根据图 5 中结果，MAVEN 在大多数任务上都没有学会持续击败敌方智能体的策略，证明单纯地将智能体的行为限制在由分层策略控制的共享潜在变量上并不能有效避免过度协作. 值得注意的是，大多数算法在 2c vs 64zg 上都能在 60 万步内达到 60%以上的胜率，BVM-CMARL 提升幅度仍然不大. 本文认为，这是由于该地图相对简单，我方智能体能较快击败敌方智能体，所以过度协作带来的影响较小. 而随着地图难度的增大，其他基准方法难以在给定的时间步内取胜，此时 BVM-CMARL 缓解过度协作的优势得以显现. 例如，BVM-CMARL 在难地图 8m vs 9m，10m vs 11m 和超难地图 MMM2 上都达到了 90%左右的胜率，训练过程中的胜率优于三种基准方法. 表 3 详细记录了每个算法在 SMAC 上最后 40 万步的平均胜率，实验结果显示，本文所述方法的平均胜率为 82.81%，比次优算法 RODE 高 13.42%.

表 3 对比实验结果

算法	最后 4e5 步平均胜率(%)
BVM-CMARL	**82.81**
EOI	43.67
MAVEN	35.22
RODE	69.39
QPLEX	42.88
QMIX	46.61

在 Hallway 任务上，由图 5(f)可知 QMIX、RODE 和 MAVEN 胜率极低，而 QPLEX 虽然训练前期效果超过 BVM-CMARL，但最终胜率仍然很低，只能达到 50%左右的胜率，而本文方法 BVM-CMARL 可以在 120 万步内达到 90%以上的胜率.

BVM-CMARL 可以和多种值分解方法结合. 因此，我们探究了本文方法与不同值分解方法结合后表现是否提升. 选取了 VDN 和 QPLEX 两种最常见的值分解方法进行对比，将结合后的方法分别记为 BVM-VDN 和 BVM-QPLEX，在六个环境中平均胜率如图 6 所示.

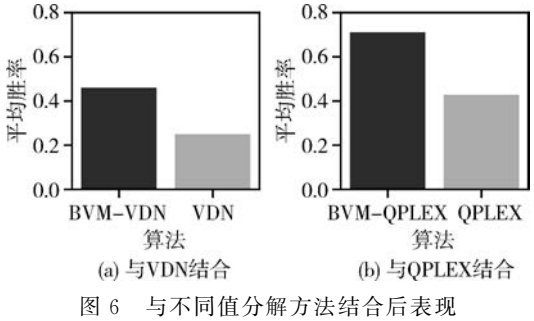


图 6 与不同值分解方法结合后表现

根据图 6 中的结果，BVM-CMARL 与各种值分解方法结合后，性能相对于原方法有较大提高. 实验结果验证了 BVM-CMARL 的鲁棒性，并进一步说明了 BVM-CMARL 可以不依赖值分解模块，以用于一般性任务，例如混合协作和竞争任务.

5.4　消融实验

为了探究模型中各个组分对本文所提方法的必要性,本节在 SMAC 的 6 个简单地图和两个 LBF 地图上展开消融实验.

为了证明双视角建模的必要性,在设计变体时保持全局视角不变,即每个智能体的 Q 值仅由全局 Q 值构成,旨在观察 BVM-CMARL 去掉局部视角后的表现. 如图 7(a)—(f)所示,该单视角模型(Single-View Modeling Collaborative Multi-Agent Reinforcement Learning, SVM-CMARL)在 SMAC 上的胜率和收敛速度上均有所下降,平均胜率下降了 13.97%. 而对于 LBF 任务,在较简单的 LBF-10×10地图上,获得的奖赏下降了 16.67%,但仍可以获得 0.6 左右的平均奖赏. 这是由于 LBF-10×10 地图状态空间较小,智能体能较快获得食物,所以智能体不会轻易陷入过度协作. 在较难的 LBF-16×16 中 BVM-CMARL 可以获得 0.673 的奖赏,但是去掉局部视角后只能获得 0.4 左右的奖赏,此时 BVM-CMARL 避免过度协作的优势得以显现. 因此,实验证明了双视角建模对本文方法的必要性.

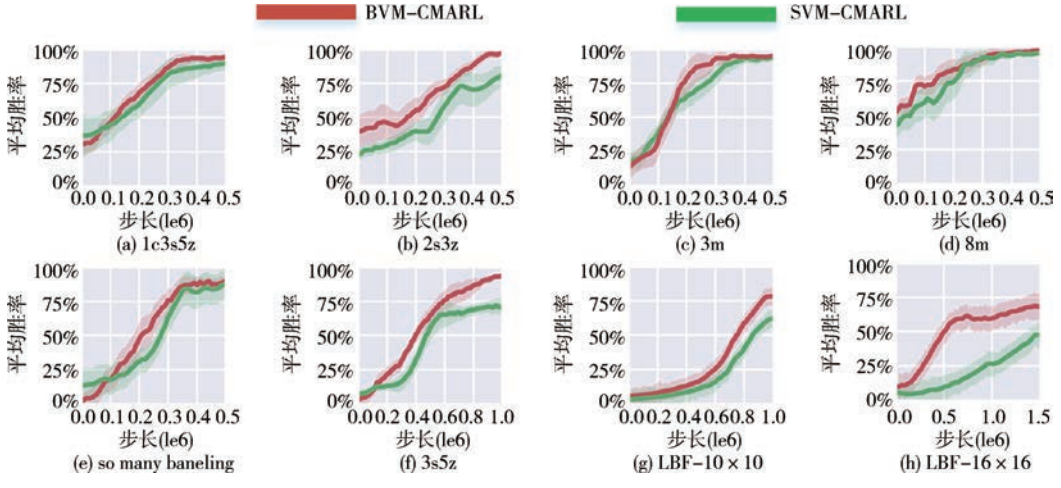


图 7　消融实验结果

在整合感知变分时,本文采用了注意力机制获取全局变分. 为了探究注意力机制对本文方法的影响,本文在无注意力机制的情形下进行实验,即直接将所有感知变分的拼接作为全局信息,实验结果如图 8 所示. 根据实验结果,注意力机制对本文方法的影响较显著,提升平均性能 9.36%,证明了注意力机制的有效性.

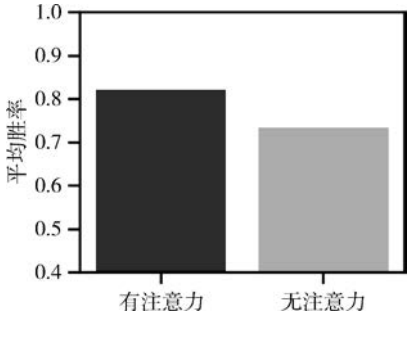


图 8　注意力机制对实验效果的影响

5.5　超参数敏感性实验

本文采用了两个超参数 λ_Q 和 $\lambda_{\mathcal{L}}$ 来平衡两个视角的重要性. 为探究这两个超参数对模型效果的影响,对于每个参数值,在 5m vs 6m 地图上用不同的参数独立运行了 5 次实验,实验结果如图 9 所示.

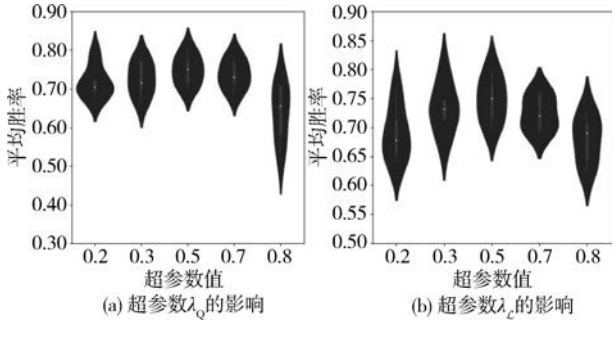


图 9　超参数对实验效果的影响

根据图 9 中的结果,λ_Q 和 $\lambda_{\mathcal{L}}$ 均为取 0.5 时效果最好,表现为超参数的取值越接近 0.5 则平均胜率和稳定性越好. 由于 λ_Q 和 $\lambda_{\mathcal{L}}$ 这对参数体现了两个视角在训练时的重要性,实验结果说明两个视角的训练权重相同时,模型的效果最好,进一步说明了局部信息和全局信息平衡的重要性.

5.6　可视化分析

为了说明双视角建模平衡智能体策略多样性和

协作的效果，本节考虑对 BVM-CMARL 训练出的局部和全局变分进行可视化分析. 具体地，我们选取 Hallway 任务中三条不同走廊的智能体的局部和全局变分进行分析. 由 5.3 节图 5 可知，BVM-CMARL 在 Hallway 任务上的胜率基本能在 110 万步内收敛，因此本节在 110 万步后随机选取了的 5 个连续时间步，使用 PCA 将这 5 个连续时间步的局部和全局变分降维到 3 维空间，结果如图 10 所示.

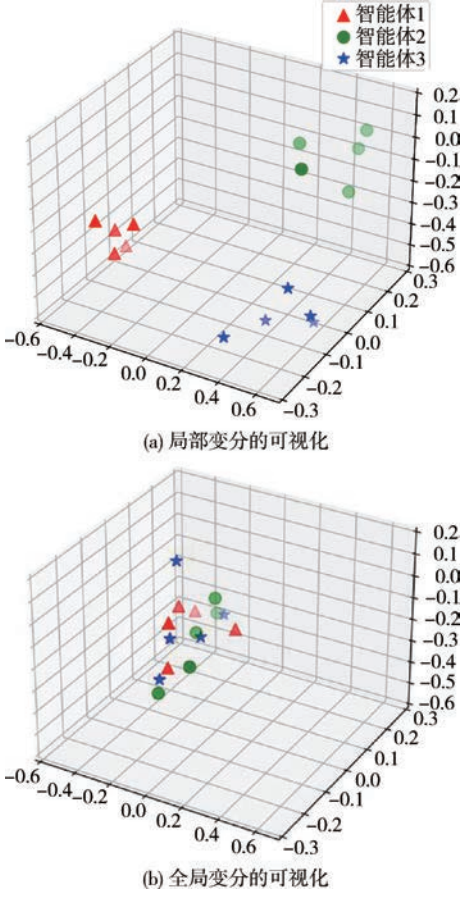


图 10 BVM-CMARL 在 Hallway 上可视化

由图 10(a)可知，三个智能体的局部变分在低维空间中距离较远，这是由于局部变分学到了较多样化的局部信息，有助于激励智能体的策略多样性. 而在图 10(b)中，三个智能体的全局变分在空间中距离较近，可以认为所有智能体的全局变分学到了较一致的全局信息，进而能有效促进协作. 而通过两者的结合，智能体能在协作的同时保持策略多样性，避免陷入过度协作.

6 总 结

在基于 CTDE 架构的多智能体强化学习中，过度协作是导致智能体的协作效率低下的一个重要原因. 已有方法大多通过预先限制智能体的协作来避免过度协作，同样导致了智能体的协作效率低下. 为了解决这一问题，本文所提 BVM-CMARL 方法不限制智能体的协作，而是直接完成对智能体不同视角的建模，从而实现智能体策略多样性与协作的平衡. 该方法首先将 Q 值分解到全局视角和局部视角，在局部视角，采用对比学习方法最大化局部变分与自身轨迹的互信息，使得局部变分表征出丰富的局部信息，进而根据局部变分训练局部 Q 网络；同时在全局视角，最大化全局变分与其他智能体动作的互信息，使得全局变分获取感知全局环境中其他智能体行为的能力，随后使用全局变分训练全局 Q 网络. 实验结果显示，本文所述方法的平均胜率为 82.81%，比次优算法 RODE 高 13.42%. 并且在 SMAC 的简单任务和 LBF 中展开的消融实验证明了双视角建模对 BVM-CMARL 的必要性.

在未来的研究中，我们将关注 DTDE 架构下智能体策略多样性与协作平衡的问题. 此外，本文实验主要考虑智能体数量少于 20 的场景，而大规模多智能体强化学习任务[32-34]中出现的过度协作现象同样值得关注，在未来研究中我们将进一步探讨该问题.

参 考 文 献

[1] Yang S, Yang B, Kang Z, Deng L. IHG-MA: Inductive heterogeneous graph multi-agent reinforcement learning for multi-intersection traffic signal control. Neural Networks, 2021, 139(1): 265-277

[2] Wang T, Cao J, Hussain A. Adaptive traffic signal control for large-scale scenario with cooperative group-based multi-agent reinforcement learning. Transportation Research Part C: Emerging Technologies, 2021, 125(1): 103046

[3] Sahraoui M, Bilami A, Taleb-Ahmed A. Schedule-based cooperative multi-agent reinforcement learning for multi-channel communication in wireless sensor networks. Wireless Personal Communications, 2022, 122(4): 3445-3465

[4] Li X, Hu X, Zhang R, Yang L. Routing protocol design for underwater optical wireless sensor networks: A multiagent reinforcement learning approach. IEEE Internet of Things Journal, 2020, 7(10): 9805-9818

[5] Mguni DH, Wu Y, Du Y, Yang Y, Wang Z, Li M, Wen Y, Jennings J, Wang J. Learning in nonzero-sum stochastic games with potentials// Proceedings of the International Conference on Machine Learning. 2021. 7688-7699

[6] Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P.

Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature, 2019,575(7782):350-354

[7] Wong A, Bäck T, Kononova AV, Plaat A. Deep multiagent reinforcement learning: Challenges and directions. Artificial Intelligence Review, 2022,56(6):1-34

[8] Lanctot M, Zambaldi V, Gruslys A, Lazaridou A, Tuyls K, Pérolat J, Silver D, Graepel T. A unified game-theoretic approach to multiagent reinforcement learning // Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2017,30

[9] Samvelyan M, Rashid T, Schroeder De Witt C, Farquhar G, Nardelli N, Rudner TG, Hung C-M, Torr PH, Foerster J, Whiteson S. The starcraft multi-agent challenge // Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. Vancouver, Canada, 2019. 2186-2188

[10] Wang T, Wang J, Zheng C, Zhang C. Learning nearly decomposable value functions via communication minimization // Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[11] Papoudakis G, Christianos F, Schäfer L, Albrecht SV. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. arXiv preprint arXiv: 2006.07869, 2020

[12] Sunehag P, Lever G, Gruslys A, Czarnecki WM, Zambaldi V, Jaderberg M, Lanctot M, Sonnerat N, Leibo JZ, Tuyls K. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296, 2017

[13] Rashid T, Samvelyan M, Schroeder C, Farquhar G, Foerster J, Whiteson S. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning // Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018. 4295-4304

[14] Wang J, Ren Z, Liu T, Yu Y, Zhang C. QPLEX: Duplex dueling multi-agent q-learning. arXiv preprint arXiv: 2008.01062, 2020

[15] Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments // Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2017,30

[16] Foerster J, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients // Proceedings of the AAAI Conference on Artificial Intelligence. 2018

[17] Wang Y, Han B, Wang T, Dong H, Zhang C. DOP: Off-policy multi-agent decomposed policy gradients // Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[18] Yuan L, Wang J, Zhang F, Wang C, Zhang Z, Yu Y, Zhang C. Multi-agent incentive communication via decentralized teammate modeling // Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 9466-9474

[19] Mao H, Liu W, Hao J, Luo J, Li D, Zhang Z, Wang J, Xiao Z. Neighborhood cognition consistent multi-agent reinforcement learning // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. 7219-7226

[20] Chen L, Guo H, Du Y, Fang F, Zhang H, Zhang W, Yu Y. Signal instructed coordination in cooperative multi-agent reinforcement learning // Proceedings of the Distributed Artificial Intelligence: Third International Conference. Shanghai, China, 2022. 185-205

[21] Wang T, Gupta T, Mahajan A, et al. RODE: Learning roles to decompose multi-agent tasks. arXiv preprint arXiv:2010.01523, 2020

[22] Eysenbach B, Gupta A, Ibarz J, Levine S. Diversity is all you need: Learning skills without a reward function // Proceedings of the International Conference on Learning Representations. Montreal, Canada, 2018

[23] Sharma A, Gu S, Levine S, Kumar V, Hausman K. Dynamics-aware unsupervised discovery of skills. arXiv preprint arXiv:1907.01657, 2019

[24] Pathak D, Agrawal P, Efros AA, Darrell T. Curiosity-driven exploration by self-supervised prediction // Proceedings of the International Conference on Machine Learning. Los Angeles, USA, 2017. 2778-2787

[25] Burda Y, Edwards H, Storkey A, Klimov O. Exploration by random network distillation. arXiv preprint arXiv: 1810.12894, 2018

[26] Mahajan A, Rashid T, Samvelyan M, Whiteson S. Maven: Multi-agent variational exploration // Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019, 7613-7624

[27] Jiang J, Lu Z. The emergence of individuality // Proceedings of the International Conference on Machine Learning, 2021. 4992-5001

[28] Li P, Tang H, Yang T, Hao X, Sang T, Zheng Y, Hao J, Taylor ME, Wang Z. PMIC: Improving multi-agent reinforcement learning with progressive mutual information collaboration. arXiv preprint arXiv:2203.08553, 2022

[29] Li C, Wang T, Wu C, et al. Celebrating diversity in shared multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2021, 34(1): 3991-4002

[30] Oliehoek FA, Amato C. A concise introduction to decentralized pomdps. Switzerland: Springer, 2016

[31] Bai C, Wang L, Han L, et al. Dynamic bottleneck for robust self-supervised exploration. Advances in Neural Information Processing Systems, 2021, 34(1): 17007-17020

[32] Liu Quan, Zhai Jian-Wei, Zhang Zong-Zhang, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018,41(1):1-27(in Chinese)
(刘全,翟建伟,章宗长等. 深度强化学习综述. 计算机学报, 2018,41(1):1-27)

[33] Chai Lai, Zhang Ting-Ting, Dong Hui, et al. Multi-agent deep reinforcement learning algorithm based on partitioned buffer replay and multiple process interaction. Chinese Jour-

nal of Computers, 2021,44(6):1140-1152(in Chinese)
(柴来,张婷婷,董会等.基于分区缓存区重放与多线程交互的多智能体深度强化学习算法.计算机学报,2021,44(6):1140-1152)
[34] Li Jing-Chen, Shi Hao-Bin, Huang Guo-Sheng. A multi-agent reinforcement learning method based on self-attention mechanism and policy mapping recombination. Chinese Journal of Computers, 2022,45(9):1842-1858(in Chinese)
(李静晨,史豪斌,黄国胜.基于自注意力机制和策略映射重组的多智能体强化学习算法.计算机学报,2022,45(9):1842-1858)

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

SHI Mei-Long, M. S. candidate. His main research interests focus on deep reinforcement learning.

HUANG Zhi-Gang, Ph. D. candidate. His research interests include deep reinforcement learning and hierarchical reinforcement learning.

ZHANG Li-Hua, Ph. D. candidate. His research interests include deep reinforcement learning and inverse reinforcement learning.

Background

Real-world reinforcement learning tasks often involve multiple agents and are formulated as cooperative multi-agent reinforcement learning. CMARL aims to learn an efficient policy that controls multiple agents and maximizes the cumulative reward from the given task. The mainstream CMARL architecture is CTDE. However, a concern in CTDE is over-collaboration, which makes the agents lose policy diversity.

In order to maintain policy diversity in collaboration of CTDE, agents need to make trade-offs between policy diversity and collaboration, so it is important to balance the two. To address this issue, this paper proposes Bi-View Modeling Collaborative Multi-Agent Reinforcement Learning (BVM-CMARL). In this method, the local and global view is modeled to stimulate policy diversity and collaboration and achieve a balance between the two.

Supported by National Natural Science Foundation of China (62376179, 62176175); Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A238); Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).