

引文格式:王焱森,王莉. 非全知环境中的多智能体双向通讯顺序决策模型[J]. 太原理工大学学报, 2026, 57(2): 387-396.
WANG Yaosen, WANG Li. Bidirectional communication-based sequential decision-making model for multi-agent systems in partially observable environments [J]. Journal of Taiyuan University of Technology, 2026, 57(2): 387-396.

非全知环境中的多智能体双向通讯 顺序决策模型

王焱森, 王莉

(太原理工大学 人工智能学院, 山西 太原)

摘要:【目的】多智能体系统(MAS)在多个领域取得了广泛应用,尤其是在工业控制和自动化领域。然而,在局部观测环境中,多智能体系统因信息缺失与动态策略交互面临非平稳性挑战,导致协作效率低下与学习收敛困难。提出一种基于双向通讯的多智能体顺序决策模型,通过序列化决策与双向信息交互缓解上述问题。【方法】首先,将传统并行决策转化为智能体依次决策的序列过程,降低策略冲突;其次,设计双向通讯模块,结合前向动作意图传递与反向注意力驱动的观测信息提取,增强全局感知能力;同时引入决策调度模块,动态评估智能体状态价值以优化决策顺序。实验基于星际争霸多智能体挑战(SMAC)环境,谷歌足球(GRF)环境,涵盖异构、同质及复杂对抗场景。【结果】结果表明,本文方法在胜率与收敛速度上显著优于基线算法,消融实验验证了双向通讯与决策调度模块的有效性。

关键词:多智能体强化学习;非平稳性问题;顺序决策架构;双向通信

中图分类号:TP183 **文献标识码:**A

DOI: 10.16355/j.tyut.1007-9432.20250288

文章编号: 1007-9432(2026)02-0387-10

Bidirectional Communication-Based Sequential Decision-making Model for Multi-agent Systems in Partially Observable Environments

WANG Yaosen, WANG Li

(College of Artificial Intelligence, Taiyuan University of Technology, Taiyuan, Shanxi, China)

Abstract:【Purposes】In recent years, multi-agent systems (MAS) have achieved extensive applications across various domains, particularly in industrial control and automation. However, in partial observability environments, MAS face non-stationarity challenges caused by information incompleteness and dynamic policy interactions, leading to inefficient cooperation and learning convergence difficulties. In this paper, a bidirectional communication-based sequential decision-making model for multi-agent systems is proposed to address these issues through serialized decision processes and bidirectional information exchange. 【Methods】First, traditional parallel decision-making was transformed into a sequential process where agents make decisions successively to reduce policy conflicts. Second, a bidirectional communication module was designed, which integrates forward action intention propagation with reverse attention-driven observation extraction to enhance global perception ca-

收稿日期: 2025-05-06; 修回日期: 2025-07-22

基金项目: 煤系低渗储层“一巷采三气”全生命周期智能高效抽采基础研究(U22A20167)

第一作者: 王焱森, 硕士, (E-mail) 1670721877@qq.com.

通信作者: 王莉, 博士, 教授, 主要研究方向为人工智能、大数据, (E-mail) wangli@tyut.edu.cn.

pabilities. Additionally, a decision scheduling module was introduced to dynamically evaluate agent state values for optimizing decision sequences. Experiments were conducted in the StarCraft Multi-Agent Challenge and Google Research Football environments, covering heterogeneous, homogeneous, and complex adversarial scenarios. **【Results】**Results demonstrate that the proposed method significantly outperforms baseline algorithms in win rate and convergence speed, and ablation studies validate the effectiveness of both bidirectional communication and decision scheduling modules.

Keywords: multi-agent reinforcement learning; non-stationarity problem; sequential decision-making architecture; bidirectional communication

近年来,合作型多智能体的研究逐渐从游戏领域被应用到许多工业控制以及生活的场景当中,例如无人机集群自主避障,关键节点的搜寻与网络瓦解问题^[1],自动驾驶当中的智能车的控制与调度问题,以及核聚变的自动化控制问题^[2]。目前,现存的多智能体强化学习算法已经在多智能体合作领域取得了惊人的成就^[3]。

在分布式局部观测多智能体系统中,每个智能体根据各自的局部观测状态独立决策。在这种场景下,环境的状态转移呈现出非平稳特性。首先,在传统多智能体强化学习体系下,智能体并发决策。每个智能体在尝试学习其最佳策略,而其他智能体也在与环境进行交互,这种互动会导致环境持续发生动态变化^[4]。决策环境的不断变化以及智能体并发决策难以控制,使得强化学习算法无法收敛到纳什均衡。智能体互相影响和策略调整的循环使得多智能体强化学习过程呈现非平稳特性。其次,在非全知的环境中,智能体缺乏对全局环境尤其是其他智能体的行为的感知。感知缺陷使得智能体难以及时感知环境的变化并对自身策略做出调整^[5]。常规的通信机制通过传递智能体观测信息来缓解这一问题。但是在传统的通信方法中,未经处理的观测信息在一定程度上对智能体决策造成干扰,而且仅仅传递观测信息无法让智能体互相理解对方意图并做出更好的配合^[6]。为了解决上述两种问题,我们提出了非全知环境下的双向通信的顺序决策模型。该模型通过两个创新点解决传统方法的局限性:

1)引入半中心化的顺序决策架构,利用状态价值评估动态调度决策顺序,赋予高价值智能体优先决策权以平衡自主性与协作效率;

2)设计双向通信机制,正向传递动作意图实现预判协作,反向通过注意力机制筛选关键观测信息减少干扰。实验证明该模型在SMAC任务中有效缓解了并发决策导致的非平稳性问题,性能优于基准方法。

1 相关工作

1.1 观测信息补全的方法

非全知环境下智能体无法获取全局信息,通过通信的方式智能体可以交换观测以及意图信息从而缓解非平稳问题。近年来,众多研究围绕MARL中的通信展开,并取得了一系列进展。

YUN et al^[7]提出了一种注意力交换网络GAXNet (Generalized Attention-based exchange Network), GAXNet采用星形拓扑和注意力权重交换机制,结合URLLC指标优化通信质量。文献[8]设计了LSC (Leader-based Simplified Communication)。相较于GAXNet的固定拓扑结构,LSC通过领导者选举和对等传播两阶段通信,实现拓扑动态调整。SUN et al^[9]学者提出了双层多智能体通信框架T2MAC(Two-Tiered Multi-Agent Communication)。该方法利用门控注意力与贝叶斯信任模型实现目标导向的稀疏通信。

传统的通信方法中,智能体传递的信息冗余且缺乏动作信息,尤其是在对合作要求高的场景下不能很好地给智能体的决策提供帮助。

1.2 多智能体强化学习框架

目前通过“中心化训练-去中心化执行”的强化学习框架来解决智能体并发决策导致的非平稳问题,文献[10]提出的MAPPO(Multi-Agent Proximal Policy Optimization)通过对传统PPO算法进行调整,共享Critic网络和分散式策略网络平衡全局优化与局部稳定性。

中心化训练去中心化执行(CTDE)方法通过集中训练共享全局信息,优化多智能体协作策略,缓解非平稳性并提升执行鲁棒性,但其计算成本随联合动作空间维度指数增长,且静态信息机制难以适应动态环境或策略偏移,全局噪声可能干扰学习效率。

1.3 序列化的解决方案

目前有研究将多智能体并发决策问题转化为

序列预测问题,这类方法可以对全局轨迹建模,有效缓解了由于智能体并发决策以及智能体信息缺失引发的非平稳性问题。CHEN et al^[11]提出了Decision Transformer方法,该方法将强化学习问题重新定义为序列建模任务,通过序列生成直接建模全局轨迹,避免因智能体策略相互影响导致的动态偏差。MENG et al^[12]提出的Offline Pre-trained Multi-Agent Decision Transformer(MADT),利用Transformer架构对智能体的观测序列进行建模,实现了在离线数据上预训练通用策略,通过预训练学习多智能体共享表征,提升对其他智能体策略分布的鲁棒性,从而提升在非平稳环境下的鲁棒性。CAI et al^[13]提出的基于Transformer的多智能体强化学习方法COMAT,通过结合图神经网络和自注意力机制,有效捕捉异构多机器人系统中复杂的交互动态,实现了策略在不同团队规模和机器人能力组合上的良好适应性,降低非平稳性干扰。LI et al^[14]提出了双向动作依赖协同多智能体Q学习(ACE)方法,该方法通过建模智能体间的双向动作依赖关系,利用协同Q值分解机制来优化联合策略,从而提升多智能体协作效率。其优点是能够有效捕捉智能体间的复杂交互,并在非全知环境下实现更稳定的策略学习。

现阶段序列的方法要求智能体拥有全局视野,不符合实际应用场景,且大部分方法依赖离线预训练数据,且高度依赖历史数据的分布质量。若数据集存在策略差异过大或噪声干扰,模型可能过度拟合低质量行为,导致策略泛化能力下降。此外序列化的方案依赖固定时间窗口的历史信息,难以捕捉长程动态交互,在策略复杂的场景下表现不佳。

2 问题定义

本文以局部观测马尔可夫决策过程(POMDPs)为基础模型进行研究。在局部观测马尔可夫决策过程(POMDPs)中,POMDPs用元组表示为 $(N, S, \{A^i\}_{i \in N}, P, R, \{Z^i\}_{i \in N}, O, \gamma)$,其中 $N = \{1, 2, \dots, n\}$ 为智能体集合, S 代表状态空间, A^i 表示智能体 i 的动作空间^[15]。 $A := A^1 \times A^2 \times \dots \times A^N$ 表示联合动作空间,并且 $P: S \times A \times S \rightarrow [0, 1]$ 表示从任意状态 $s \in S$ 经由动作 $a \in A$ 转变到状态 $s' \in S$ 的状态转移函数, $R: S \times A \times S \rightarrow \mathbb{R}$ 为奖励函数,用于确定智能体从 (s, a) 过渡到 s' 时获得的即时奖励, Z 为智能体观测的合集。由观测函数 $O(s, i)$ 得到的符号 $o^i \in Z^i$ 为智能体 i 在全局状态 s 时的局部观测值,

$\gamma \in [0, 1]$ 代表折扣因子。将POMDPs简化为 G 来描述状态转移,其中 $G = \{(s^t, a^t, r, o)\}$;本文将原始POMDPs中的 G 转换为序列多智能体MDP中的 $\tilde{G}$ 。

$$\tilde{G} = \{(s^t, o_1^t, \{h^i\}_{i \in N, i \neq 1}, a_1^t, 0, e_{a1}^t), \\ (e_{a1}^t, o_2^t, \{h^i\}_{i \in N, i \neq 2}, a_2^t, 0, e_{a1:2}^t) \dots \\ (e_{a1:n-1}^t, o_n^t, \{h^i\}_{i \in N, i \neq n}, a_{1:n}^t, r(s^t, a^t), s^{t+1})\}. \quad (1)$$

其中, $(s^t, o_1^t, \{h^i\}_{i \in N, i \neq 1}, a_1^t, 0, e_{a1}^t)$ 为序列决策顺序中第一个智能体的相关表示, s^t 为 t 时刻所有智能体所处的状态。由于智能体1为第一个决策的智能体,因此元组中的初始状态为 s^t 。 o_1^t 为智能体的观测, $\{h^i\}_{i \in N, i \neq 1}$ 为智能体1通过反向信息提取机制提取到的其他智能体的观测, a_1^t 为智能体1在 t 时刻采取的动作,0为智能体1获得的奖励值,原因为处于中间状态的智能体并未将动作提交到环境中。当智能体1完成决策后会将自身的动作信息和自身观测信息通过正向消息传递机制传递给下一个智能体,同时到达中间状态 e_{a1}^t 。中间状态是一个虚拟的状态,在顺序决策过程中用该状态划分不同智能体的决策过程,同时中间状态也用来记录正向消息传递中的观测信息以及动作信息,是正向消息传递机制的基础。

在 $(e_{a1}^t, o_2^t, \{h^i\}_{i \in N, i \neq 2}, a_2^t, 0, e_{a1:2}^t)$ 中, e_{a1}^t 为顺序决策的中间状态,包含了前序所有智能体的动作信息以及观测信息。当智能体2完成决策后状态转移为 $e_{a1:2}^t$,其中下标 $a1:2$ 表示智能体1到智能体2都已经完成决策。

$(e_{a1:n-1}^t, o_n^t, \{h^i\}_{i \in N, i \neq n}, a_{1:n}^t, r(s^t, a^t), s^{t+1})$ 一项中 $e_{a1:n-1}^t$ 为智能体1到智能体 $n-1$ 都完成决策的状态。最后一个智能体完成决策后会将所有智能体的动作 a^t 一并提交到环境,同时会获取到环境的奖励 $r(s^t, a^t)$,并且状态转移到 s^{t+1} 。本文的目标为在非全知的环境下学习到一种平稳的强化学习算法,使得多智能体系统收益最大化。

3 模型框架

针对非全知多智能体强化学习场景下的智能体并发决策问题,本文在多智能体环境中增加了一个全局控制的决策调度模块,将并发模式转化为顺序模式,解决并发决策冲突的不可控问题。针对非全知多智能体强化学习场景下,在顺序决策过程中智能体无法获取到完整的全局信息的问题,本文在顺序决策的过程中引入双向通信机制以补充智能体对重要友方智能体的观测信息,弥补局部观测的

盲区,利于协同决策。

3.1 局部观测下的顺序决策模型

本文构建了基于双向通信机制的顺序决策架构(如图1所示)。具体来说,在 t 时刻,环境状态为 s^t ,所有智能体将自身的观测 o^t 提交到决策调度模块。决策调度模块根据所有智能体的观测对智能体的状态进行评估得到决策顺序。智能体将按照决策顺序在多智能体系统内部进行顺序决策。在决策过程中,系统采用双向通信机制传递其他智能体的动作信息以及观测信息。以智能体 m 为例,智能体 m 在决策开始前经由正向通信机制获取智能体 $m-1$ 的动作信息以及观测信息并到达中间状态 $e_{a1:m}^t$ 。在进入中间状态后,智能体 m 会通过反向的

信息网络提取后续的智能体观测中对自己有利的信息。智能体 m 会根据自身的观测 o_m^t ,反向通信获取的信息以及前向通信传递来的动作信息和观测信息进行决策得到动作 a_m^t 。在智能体 m 决策完成后环境状态 s^t 不发生转移,中间状态从 $e_{a1:m-1}^t$ 转移到 $e_{a1:m}^t$ 。由于没有发生真实的状态转移,中间状态的智能体不会获得奖励信息。当最后一个智能体决策完成后,该智能体会将所有智能体的动作提交到环境中。交互环境根据智能体集群的动作和状态给出奖励 r^t ,此时状态从 s^t 转移到 s^{t+1} 。经由上述过程,我们成功将传统多智能体强化学习决策过程转化为局部观测下带通讯的多智能体序列决策过程。

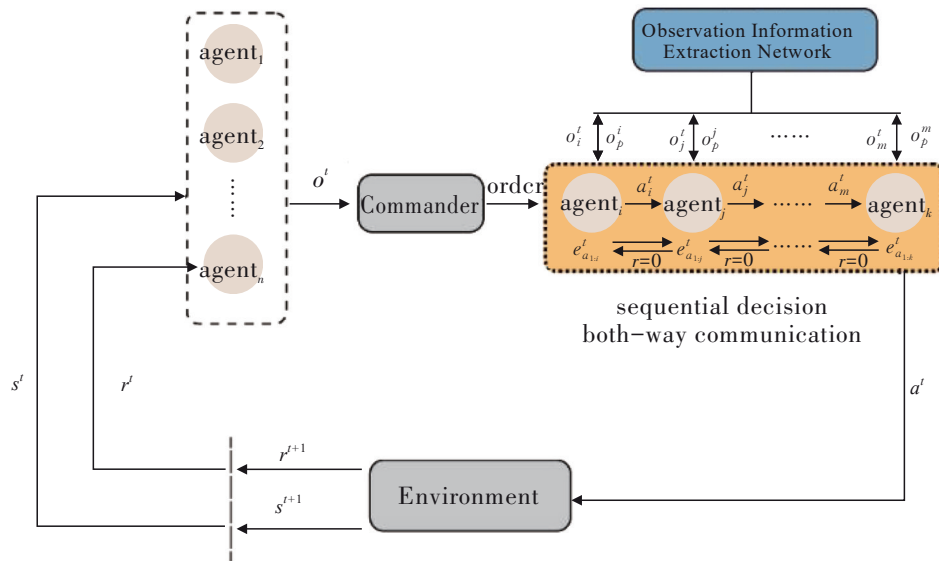


图1 双向通信的顺序决策模型

Fig.1 Sequential decision model for bidirectional communication

3.2 顺序决策

群体决策中总是隐含一定的博弈策略,在局部感知环境中,群体智能体的决策实际上有主次之分,这样就构成了一定的顺序决策机制,状态价值越高的智能体决策顺序会靠前,价值低的智能体会根据当前状况调整自己策略。基于此,我们提出决策调度模块,通过评估各智能体状态价值来确定决策顺序,将状态价值较高的智能体置于决策顺序前列,赋予其更大主动权,同时为后续智能体决策提供引导。

具体来说, $[agent_1, agent_2, \dots, agent_n]$ 会观测自身所处的环境得到观测 o 。所有智能体将自己的观测值 o 传递给决策调度模块。决策调度模块会根据智能体的观测进行状态评估得到状态评估分数集合 $U = \{u'_1, u'_2, \dots, u'_n\}$ 。该网络根据评估分数输

出智能体的决策顺序 $[agent_i, agent_k, \dots, agent_m]$ 。

在决策调度模块的计算中,每个智能体的局部状态表示为 s_i^t ,局部状态包含多种信息,例如智能体自身的属性、对周围环境的观测等。对于智能体 i ,根据贝尔曼方程,目标价值 $U_{\text{target}}^i(s_i^t)$ 可以表示为:

$$U_{\text{target}}^i(s_i^t) = r_i(s_i^t, a_i^t) + \gamma E_{s_i^{t+1} \sim P(s_i^{t+1}|s_i^t, a_i^t)} [U_{\theta_i}(s_i^{t+1})]. \quad (2)$$

其中, $r_i(s_i^t, a_i^t)$ 是智能体 i 在状态 s_i^t 采取动作 a_i^t 获得的即时奖励, γ 是折扣因子, $P(s_i^{t+1}|s_i^t, a_i^t)$ 是状态转移概率, s_i^{t+1} 是下一个状态。本文通过训练一个共享神经网络 $U_{\theta}(s, i)$ 来近似每个智能体的目标价值 U_{target} ,输入为全局状态 s 和智能体的索引 i ,输出为智能体 i 的状态价值。顺序决策过程如图2所示。

为了在中间状态 $e_{a1:i}^t$ 下做出决策,本文评估智能体 $i+1$ 的动作空间 A^{i+1} 中采取动作转移到的所

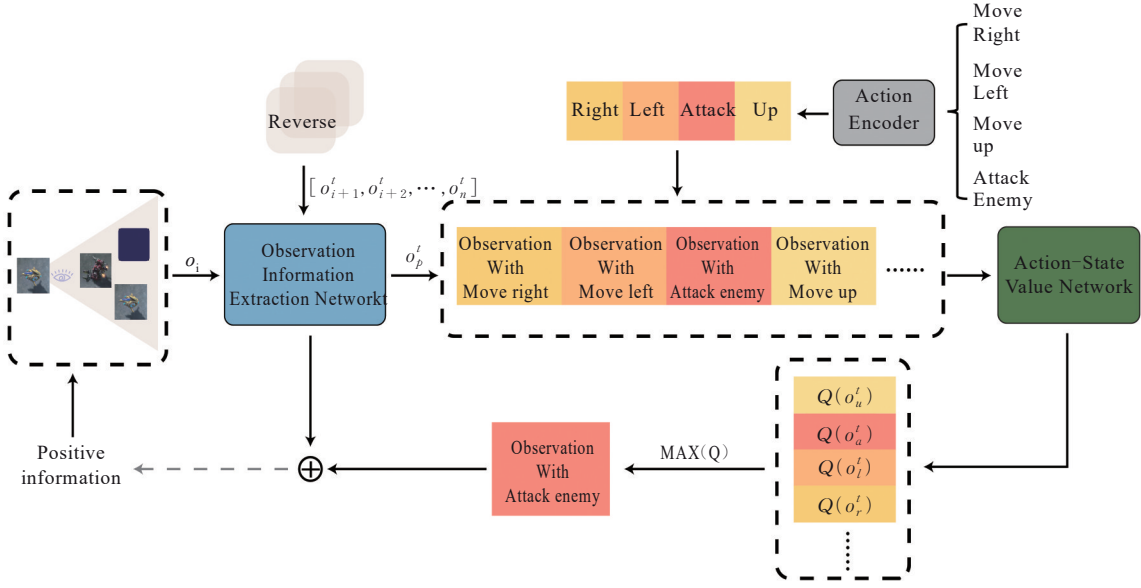


图 2 顺序决策过程

Fig.2 Sequential decision process

有可能的下一个中间状态 $e_{a1:i+1}^t$ ($i < n$) 或 e_{a1}^{t+1} ($i = n$), 并根据 $a_i^t = \operatorname{argmax}_{a_i} Q(e_{a1:i+1}^t, a_{i+1}^t)$ ($i < n$) 或者 $a_i^t = \operatorname{argmax}_{a_i} Q(e_{a1}^{t+1})$ ($i = n$) 选择动作。价值函数 Q 源自单智能体强化学习的 Bellman 方程, 在中间状态下估算后续智能体状态价值并选最高值对应的动作, 由于终态才有真实奖励, 因此 Q 函数分中间状态和终态计算形式, 如公式 3 所示。其中 t 表示时间步, $e_{a1:i}^t$ 是在 t 时刻, 前面 i 个智能体依次执行动作 $(a_1^t, a_2^t, \dots, a_i^t)$ 后的观测状态。 a_{i+1}^t 表示智能体 $i+1$ 采取的动作。 $Q(e_{a1:i}^t, a_{i+1}^t)$ 是在 t 时刻下, 前面 i 个智能体执行动作 $(a_1^t, a_2^t, \dots, a_i^t)$ 且智能体 $i+1$ 执

行动作 a_{i+1}^t 后的状态-动作值函数, 用于评估在该状态下采取该动作的价值。 $r(e_{a1:n}^t)$ 是在 t 时刻, 所有 n 个智能体都执行完动作 $(a_1^t, a_2^t, \dots, a_i^t)$ 后获得的奖励。 γ 是折扣因子, 用于衡量未来奖励的重要程度。 $i < n$ 时智能体处于中间决策状态, 不会接收到环境的奖励, 当 $i = n$ 时智能体会获得环境的奖励 $r(e_{a1:n}^t)$ 。

$$\hat{Q}(e_{a1:i}^t) = \begin{cases} \max_{a_{i+1}^t} \gamma Q(e_{a1:i+1}^t, a_{i+1}^t), & \text{if } i < n \\ \max_{a_{i+1}^t} (r(e_{a1:n}^t) + \gamma Q(e_{a1}^{t+1})), & \text{if } i = n \end{cases} \quad (3)$$

反向传播过程中的损失函数为:

$$L = \begin{cases} \frac{1}{2} (Q(e_{a1:i}^t, a_i^t) - \gamma \max Q(e_{a1:i+1}^t, a_{i+1}^t))^2 & \text{if } i < n \\ \frac{1}{2} (Q(e_{a1:n}^t, a_n^t) - (r(e_{a1:n}^t) + \gamma \max Q(e_{a1}^{t+1}, a_1^{t+1})))^2 & \text{if } i = n \end{cases} \quad (4)$$

当 $i < n$ 时, 由于智能体处于中间状态, 中间状态的转移不会获得奖励, 因此损失由 $Q(e_{a1:i}^t, a_i^t)$ 与 $\gamma \max Q(e_{a1:i+1}^t, a_{i+1}^t)$ 计算得到。当 $i = n$ 时, 智能体会获得环境给出的奖励 $r(e_{a1:n}^t)$, 因此损失函数计算时会加上对应的奖励值。通过最小化损失, 智能体能够学习到更准确的 Q 值估计, 让当前动作的价值评估与考虑后续智能体最优动作选择后的预期价值相符。

3.3 基于注意力机制的双向通信模块

针对顺序决策过程中智能体无法获取完整的全局信息的问题, 我们提出了顺序决策框架下的双向通信模块, 通过正向通信传递智能体的动作信

息、反向通信提取其他智能体的观测信息补全全局信息。动作信息使得智能体可以理解其他智能体的行为并且做出配合, 反向通信中的观测信息给智能体提供了更多的环境信息, 可以更好地辅助自身做出有利于全局的决策从而提升决策的效率和稳定性。

正向通信机制: 正向通信是智能体在顺序决策中将自身观察和动作信息编码后传递给后续智能体的方式。正向通信机制通过编码、拼接、评估、传递四个步骤实现, 在实现过程中高度依赖顺序决策过程。正向通信在顺序决策中传递的核心内容是智能体的动作-观测编码, 其本质是当前智能体的

动作信息与局部观测的融合表达。具体来说,假设智能体 i 的所有的动作为 $\{a_1, a_2, \dots, a_n\}$, 首先, 动作编码器将智能体的所有动作编码生成 $\{m_1, m_2, \dots, m_n\}$, $m_i = f(a_i)$, 其中 f 为动作编码函数。随后, 将 m_i 和当前观测 o_i^t 进行拼接得到动作-观测编码。将动作-观测编码输入到动作价值网络中, 网络输出每个动作-观测对应的价值, 智能体从中选择价值最大的动作提交到联合动作集合中。并把动作-观测编码传递给后续决策的智能体 $i+1$, 完成正向消息传递。后续智能体将继续按照此流程继续进行正向消息传递。

基于注意力的反向通信: 在顺序决策过程中智能体通过注意力机制从后续智能体提取信息, 如图 3 所示, 整合为有利于当前智能体决策的信息, 从而对全局环境建模。具体来说, 当智能体 i 处于中间状态 $e_{a1,i-1}^t$ 时, 智能体会获取后续智能体的观测表征 $\{o_{i+1}^t, \dots, o_n^t\}$ 。反向信息提取网络从观测表征 o_i^t 中生成查询向量 q_i^t , 键向量 k_i^t 以及值向量 v_i^t , 经由计算得到增强表征 o_p^i , o_p^i 为智能体 i 增强表征, p 为 plus 的缩写。 o_p^i 包含了智能体利用注意力机制提取出的和自身观测相关性高的表征信息, 是智能体根据其他智能体的观测对全局状态的还原。注意力机制 (Q, K, V) 的计算如公式 5、公式 6 所示。为建模全局环境, 建立解码比对监督, 将智能体提取融合的增强表征输入解码器, 用于全局观测, 有效引导表征学习过程。

$$Q, K, V = \text{MLP}_{Q,K,V}([o_i^t, o_{i+1}^t, \dots, o_n^t]). \quad (5)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (6)$$

3.4 算法复杂度分析

决策调度模块的状态价值评估采用共享神经网络, 时间复杂度为 $O(N \cdot F)$, 其中 N 为智能体数量, F 为特征维度。双向通信机制中, 正向通信的动作-观测编码复杂度为 $O(N \cdot A)$, 其中 A 为动作空间维度, 反向注意力机制的复杂度为 $O(N^2 \cdot d_k)$, 其中 d_k 为键向量维度。当智能体数量从 $N=10$ 增至 $N=100$ 时, 单次决策的时间复杂度从 $O(10F + 10A + 100d_k)$ 提升至 $O(100F + 100A + 10\,000d_k)$ 。为应对大规模场景, 未来可引入分层决策架构, 将智能体划分为子群体 (每个子群体 ≤ 20 个智能体), 使复杂度降至 $O\left(\left(\frac{N}{G}\right)^2 \cdot d_k\right)$ (G 为子群体数量), 在保持协作效率的同时降低计算开销。

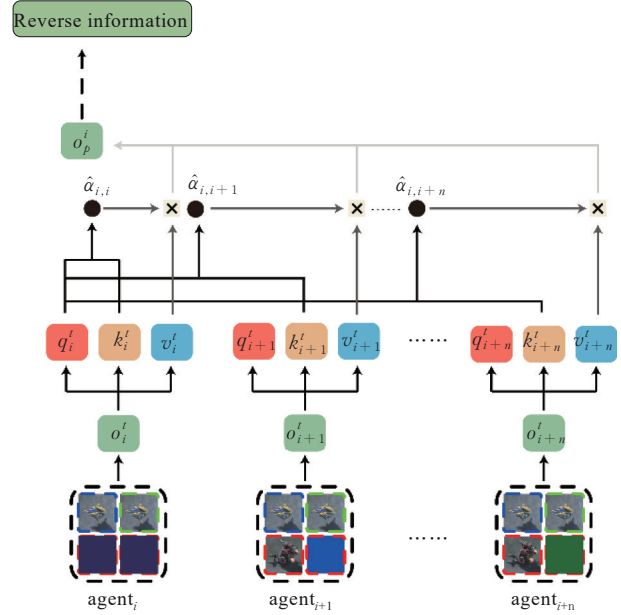


图 3 基于注意力的反向通信

Fig.3 Attention-based reverse communication

4 实验结果与分析

4.1 实验环境

为验证算法在复杂局部观测场景下的有效性, 实验基于星际争霸多智能体挑战 (StarCraft Multi-Agent Challenge, SMAC) 环境开展。该环境支持异构单位协同、大规模离散动作空间和动态对抗场景。在不同的任务中, 智能体无法观测到整个战场环境。本文设定智能体观测范围为半径为 3 的圆形区域, 智能体攻击范围为半径为 6 的圆形区域, 智能体可以在视野范围外发动攻击。该设定对智能体间信息交换能力以及智能体合作能力提出了更高的要求。

实验环境分为以下四类。

1) 微操作技巧环境 (2c_vs_64zg、3s_vs_5z、6h_vs_8z): 该环境要求智能体集火同一个目标, 集火时间偏差或定位错误可能导致游戏失败, 对智能体合作精度提出了更高的要求;

2) 同构不对称环境 (5m_vs_6m、8m_vs_9m): 需智能体在团队中承担不同的角色, 例如部分单位承担诱敌任务, 其余单位执行包抄。该环境对智能体对任务的分配提出了更高的要求;

3) 异构不对称环境 (3s5z_vs_3s6z、MMM2): 该环境下, 不同单位类型需形成互补策略 (如远程单位掩护近战突袭), 同时协调行动优先级。异质单位能力差异导致策略耦合度复杂, 例如治疗单位需实时预判队友行动轨迹以规避误判。该环境需要智能体全面考虑整个战场环境, 对其他智能体的行

为做出配合;

4)特定单位对抗环境(corridor):这些环境通常涉及特定类型的单位在特定地形或条件下的对抗。

智能体需要利用地图中的特定元素(如墙体)来制定战术,以在数量或种类劣势下保持战斗优势。

具体任务设置如表 1 所示。

表 1 SMAC 实验环境设置
Table 1 SMAC experimental environments

Name	Ally Units	Enemy Units	Type	Name
2c_vs_64zg	2 Colossi	64 Zerglings	micro-trick: positioning	2c_vs_64zg
3s5z_vs_3s6z	3 Stalkers &. 5 Zealots	3 Stalkers &. 6 Zealots	heterogeneous&-asymmetric	3s5z_vs_3s6z
3s_vs_5z	3 Stalkers	5 Zealots	micro-trick: kiting	3s_vs_5z
5m_vs_6m	5 Marines	6 Marines	homogeneous &. asymmetric	5m_vs_6m
6h_vs_8z	6 Hydralisks	8 Zealots	micro-trick: focus fire	6h_vs_8z
8m_vs_9m	8 Marines	9 Marines	homogeneous &. asymmetric	8m_vs_9m
corridor	6 Zealots	24 Zerglings	micro-trick: wall off	corridor
MMM2	1 Medivac , 2 Marauders &. 7 Marines	1 Medivac, 3 Marauders &. 8 Marines	heterogeneous &. asymmetric	MMM2

此外本文在 GRF(Google Research Football)环境进行相关实验。与 SMAC 相比,GRF 2020 提供了一个更具挑战性的环境,动作空间更大且奖励稀疏。在 GRF 中,智能体需要协调时间和位置来组织进攻,只有得分才能获得奖励。在本文实验中,智能体控制左队除守门员之外的球员,而游戏内置引擎控制右队球员。本文在两个具有挑战性的场景中进行实验:academy_3_vs_1_with_keeper, academy_counterattack_hard^[16]。

未验证模型在通用环境的能力,本文在 multiagent_mujoco 的环境下进行实验。

实验采用 DI-engine^[17] 框架实现,对比基线包括 QMIX^[18]、VDN、ACE 和 MASIA 等算法。

QMIX 通过混合网络非线性组合智能体 Q 值,利用单调性约束确保中心化训练与去中心化执行的一致性,缓解因策略动态耦合导致的非平稳性;

VDN 基于动作价值独立假设,以线性分解联合 Q 值简化计算,为非全知环境中的协作提供稳定策略依据,降低环境动态变化的干扰;

MASIA 通过自监督目标聚合策略相关消息,过滤冗余信息并增强决策相关性,减少智能体策略间的动态冲突;

ACE 则借助双向动作依赖建模和高效网络表示,显式捕捉智能体间行为关联性,在复杂联合动作空间中提升协作鲁棒性,从而增强对非平稳环境的适应能力。

所有实验重复 5 次,训练步数为 2M,结果以 95% 置信区间展示。

4.2 对比实验

图 4 展示了各算法在典型任务中的胜率曲线与

收敛速度。本文方法在复杂异构场景中表现显著优于基线。

MMM2 为混合编队对抗,需通过治疗链维持生存并针对敌方单位类型制定战术。在该任务中本文方法最终胜率达 92.1%,较 ACE(85.7%)、QMIX(78.4%)分别提升 6.4% 与 13.7%。在该实验中顺序决策和正向消息传递使得智能体能够更好地分享自己的意图,智能体之间可以更好地协作,且智能体通过反向通信机制获取其他整体的观测可以使得智能体更早发现敌对目标,治疗单位更早发现需要治疗的队友。3s5z_vs_3s6z 任务为异构单位对抗,敌我单位数量不对称,需协调高机动单位与近战单位的协同进攻。本文方法在 1.2M 步时胜率稳定至 89.3%,而 MASIA 仅 72.5%。该任务中决策调度模块可以更好地将一些高机动单位作为优先决策,显著提升了战术灵活性。

2c_vs_64zg 任务中,友方为 2 个巨像,敌方为 64 个跳虫。核心挑战在于通过高效的 AOE 伤害与位置控制抵御虫海冲击。本文方法胜率 96.7%,较 QMIX(82.3%)提升 14.4%。通信机制使智能体间共享动作和位置信息,更易于形成交叉火力覆盖。corridor 任务中,本文方法将胜率从 68.4% 提升至 89.4%。

在 5m_vs_6m 任务中,本文方法胜率为 87.6% 与 VDN(85.2%)接近,因同质单位意图相似性削弱了通信增益,但决策调度模块仍通过动态排序优化集火顺序,减少无效动作频率。本文方法在 6h_vs_8z 任务中仅需 450k 步即达到 95% 胜率,较 ACE(680k 步)提速 51.2%。决策调度模块通过优先决策高价值单位,加速策略优化进程。

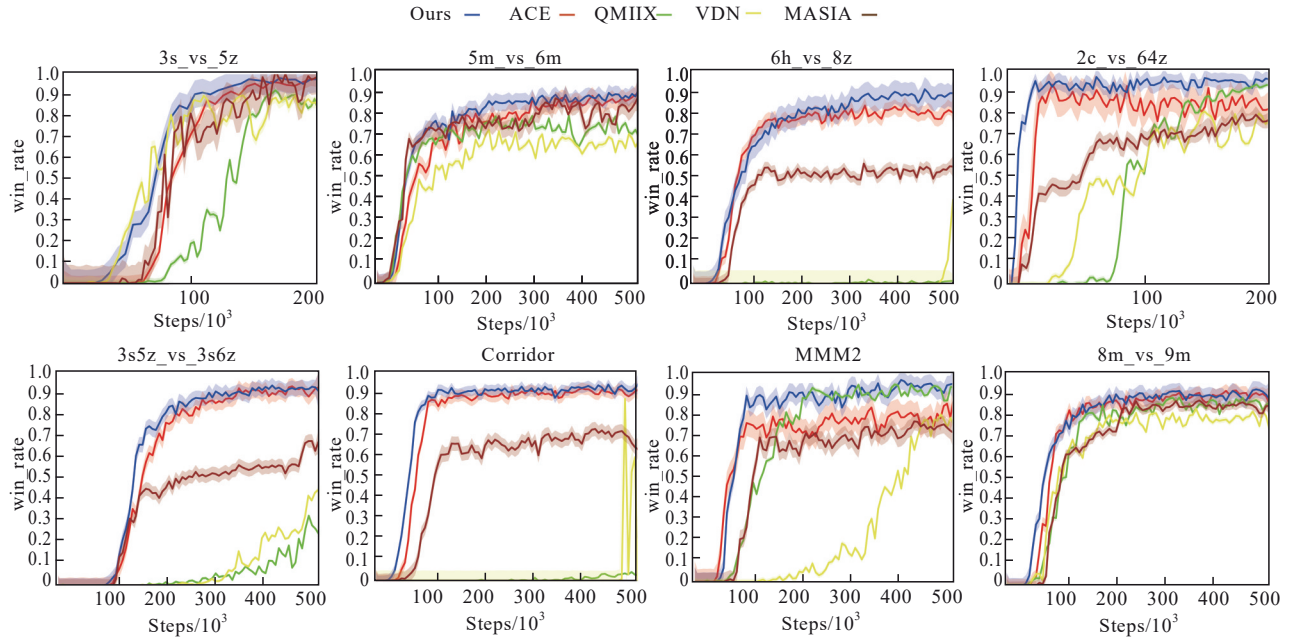


图 4 SMAC 实验结果图

Fig.4 SMAC experimental results

图 5 展示了在 GRF 的环境下本文方法与基线方法的性能对比。在这两个场景中,本文方法均超

越了基线方法。由于 GRF 对于智能体合作以及对内通信要求更高,因此模型性能差异较大。

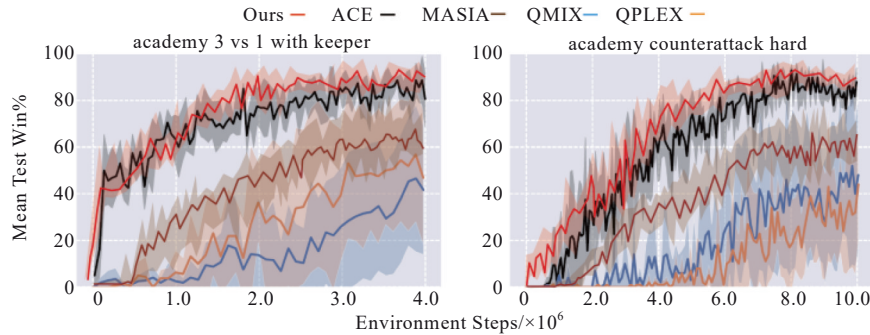


图 5 GRF 实验结果图

Fig.5 GRF experimental results

图 6 展示了在 multiagent_mujoco 的环境下本文方法与基线方法的性能对比。

在这两个场景中,本文方法均超越了基线方法,表明该方法在通用场景仍能起到很好的作用。

4.3 消融实验

为了更好地理解为什么本文的算法优于其他算法,我们分别消去决策调度模块和双向通讯模块。实验结果如图 7 所示,决策调度模块通过评估智能体状态价值动态调整决策顺序,赋予高价值智能体优先决策权。这一机制在异构场景中尤为重要。实验数据显示,加入决策调度模块后,MMM2 任务的胜率从 78.3% 提升至 92.1%,收敛速度加快约 40%。此外,在 6h_vs_8z 任务中胜率稳定在 95% 以上,显著优于 ACE 算法。

双向通讯通过前向传递动作意图和反向提取后续观测信息,显著缓解了局部观测下的信息缺失问题。具体而言,反向消息传递利用注意力机制整合后续智能体观测信息,增强当前智能体的状态表征。例如,在 3s5z_vs_3s6z 任务中,双向通讯使该任务的胜率从 72.5% 提升至 89.3%。前向消息传递则编码历史动作与观测信息,为后续智能体提供决策引导,在 corridor 任务中,胜率提高 21%。

5 结束语

本文聚焦于局部观测环境中多智能体系统面临的非平稳性挑战,提出了一种创新的基于双向通讯的多智能体顺序决策模型,有效提升了多智能体协作效率与学习收敛性。

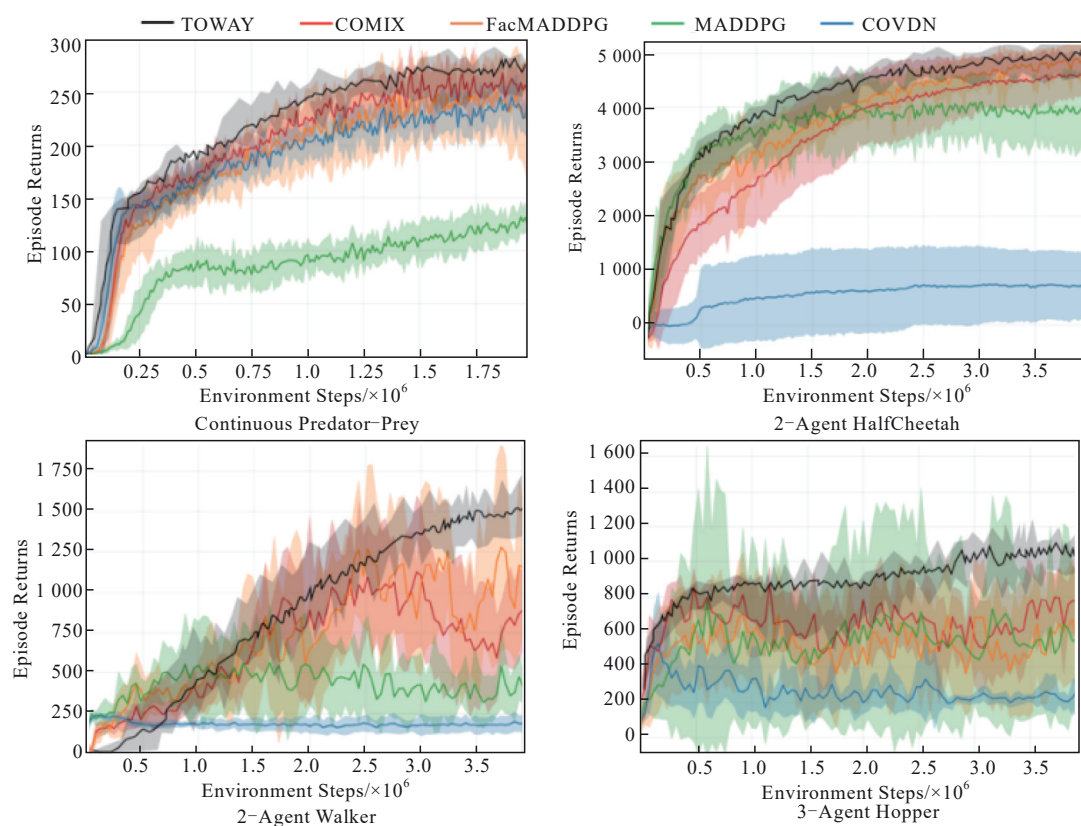


图 6 multiagent_mujoco 实验结果图

Fig.6 Experimental results of multiagent_mujoco

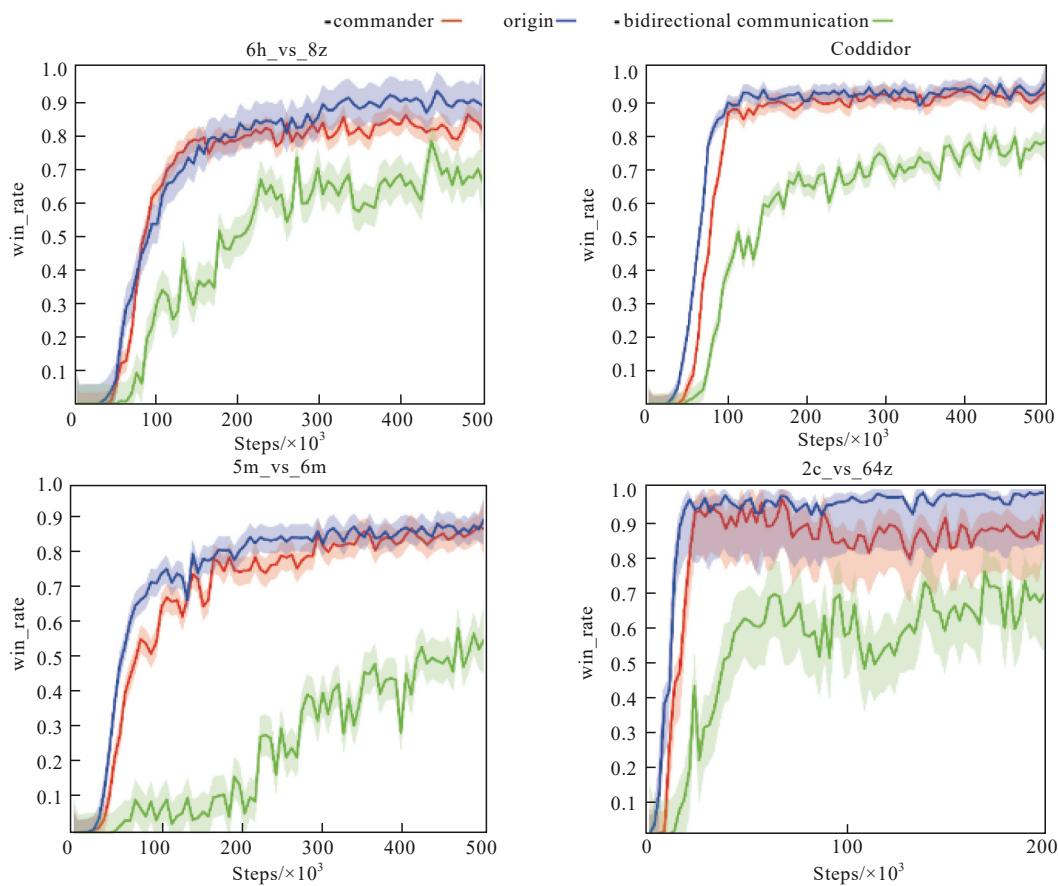


图 7 消融实验结果

Fig.7 Results of ablation experiments

在局部观测环境下,多智能体系统存在信息缺失和动态策略交互问题,导致环境非平稳,传统并发决策模式难以收敛到纳什均衡,智能体也因感知局限难以调整策略。针对这些问题,本文提出的模型包含两个核心部分:局部观测下的顺序决策模型和基于注意力机制的双向通信模块。前者通过决策调度模块评估智能体状态价值,将状态价值高的智能体置于决策前列,把并发决策转换为顺序决策,减少策略冲突;后者通过正向通信传递动作信息、反向通信提取观测信息,弥补智能体的局部观测缺陷,

增强全局感知能力,辅助智能体做出更优决策。

本文基于星际争霸多智能体挑战(SMAC)环境以及谷歌足球环境(GRF)开展实验,对比QMIX^[22]、VDN^[23]、ACE^[18]和MASIA^[11]等基线算法。结果显示,本文方法在复杂异构场景中表现卓越,在对应的实验环境中,胜率和收敛速度均显著优于基线算法。消融实验进一步验证了决策调度模块和双向通讯模块的有效性,它们分别通过动态调整决策顺序和缓解信息缺失问题,提升了模型性能。

参考文献:

- [1] FAN C, ZENG L, SUN Y, et al. Finding key players in complex networks through deep reinforcement learning[J]. *Nature Machine Intelligence*, 2020, 2(6): 317-324.
- [2] DEGRAVE J, FELICI F, BUCHLI J, et al. Magnetic control of tokamak plasmas through deep reinforcement learning[J]. *Nature*, 2022, 602(7897): 414-419.
- [3] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]// *Advances in Neural Information Processing Systems (NeurIPS)*. Long Beach, USA: Carran Associates Inc, 2017: 6379-6390.
- [4] HERNÁNDEZ-LEAL P, KARTAL B, TAYLOR M E. A survey and critique of multi-agent deep reinforcement learning[J]. *Autonomous Agents and Multi-Agent Systems*, 2019, 33(6): 750-797.
- [5] PAPOUDAKIS G, CHRISTIANOS F, RAHMAN A, et al. Dealing with non-stationarity in multi-agent deep reinforcement learning[EB/OL]. arXiv preprint arXiv: 1906. 04737.
- [6] FOERSTER J, NARDELLI N, FARQUHAR G, et al. Stabilising experience replay for deep multi-agent reinforcement learning[C]// *Proceedings of the 34th International Conference on Machine Learning*, 2017: 1146-1155.
- [7] YUN W J, LIM B, JUNG S, et al. Attention-based reinforcement learning for realtime UAV semantic communication[C]// *2021 17th International Symposium on wireless Communication System (ISWCS)*. Berlin, Germany: IEEE, 2021: 1-6.
- [8] CHARAPKO A, AILIJANG A, DEMIRBAS M. Pigpaxos: Devouring the communication bottlenecks in distributed consensus[C]// *Proceedings of the 2021 International Conference on Management of Data*. Visual Event: ACM, 2021: 235-247.
- [9] SUN C, ZANG Z, LI J, et al. T2MAC: Targeted and trusted multi-agent communication through selective engagement and evidence-driven integration[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press, 2024, 38(13), 15154-15163.
- [10] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 24611-24624.
- [11] CHEN L, LU K, RAJESWARAN A, et al. Decision transformer: Reinforcement learning via sequence modeling[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 15084-15097.
- [12] MENG L, WEN M, LE C, et al. Offline pre-trained multi-agent decision transformer[J]. *Machine Intelligence Research*, 2023, 20(2): 233-248.
- [13] CAI Y, HE X, GUO H, et al. Transformer-based multi-agent reinforcement learning for generalization of heterogeneous multi-robot cooperation[C]// *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and systems*. Abu Dhabi, United Arab Fmirates: IEEE, 2014: 13695-13702.
- [14] LI C, LIU J, ZHANG Y, et al. ACE: Cooperative multi-agent Q-learning with bidirectional action-dependency[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. Washinton, DC: AAAI Press, 2023, 37(7), 8536-8544.
- [15] OLIEHOEK F A, AMATO C. A concise introduction to decentralized POMDPs. Springer International Publishing[M]. Cham: Springer Interctional Publishing, 2016.
- [16] KURACH K, RAICHUK A, STAŃCZYK P, et al. Google research football: A novel reinforcement learning environment[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. New York, NY: AAAI Press, 2020, 34(4): 4501-4510.
- [17] CONTRIBUTORS OPENDILAB. DI-engine: OpenDILab decision intelligence engine[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtual Event: AAAI Press, 2021, 35(13), 11956-11964.
- [18] RASHID T, SAMVELYAN M, DE WITT C S, et al. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. *Journal of Machine learning Research*, 2020, 21(178): 1-51.

(编辑:贾丽红)