计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009705] [CSTR: 32024.14.csa.009705]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

双注意力记忆多智能体强化学习①

马裕博, 周长东, 张志文, 杨培泽, 张　博

(大连海事大学 人工智能学院, 大连 116026)
通信作者: 张　博, E-mail: bzhang@dlmu.edu.cn

摘　要: 多智能体协同在强化学习研究领域占据重要地位, 旨在深入探讨智能体如何通过相互协作实现共同目标. 大部分协作多智能体算法注重合作的构建, 但忽略了个体策略的强化. 为解决上述问题, 本文提出一种 BiTransformer 记忆 (BTM) 在线强化学习模型, 该模型不仅考虑多智能体之间的协同, 还利用记忆模块辅助个体决策. BTM 由双注意力编码器和双注意力解码器组成, 分别用于个体策略的增强和多智能体系统的协作. 在双注意力编码器中, 受人类的决策经验依赖的启发, 提出记忆注意力模块为当前决策提供历史决策经验. 与传统利用 RNN 的方法不同, BTM 为每一个提供的是一个显式历史决策经验库, 而非隐藏单元. 此外, 提出融合注意力模块, 在历史决策经验的辅助下处理当下的局部观测信息, 从而获取环境中最具决策价值的信息, 进一步提高智能体个体的决策能力. 在双注意力解码器中, 本文提出了决策注意力模块和合作注意力模块两个模块, 通过综合考虑其他已经做出决策智能体与当前智能体的合作收益以及带有历史决策经验的局部观察, 从而促进历史决策辅助下的多智能体潜在合作的形成. 最终本文在星际争霸中的多个场景下对 BTM 进行了测试, 取得了 93% 的平均胜率.

关键词: 多智能体协同; 在线强化学习; 局部观测; 历史决策经验; 合作收益; 个体策略增强

引用格式: 马裕博,周长东,张志文,杨培泽,张博.双注意力记忆多智能体强化学习.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9705.html

BiTransformer Memory for Multi-agent Reinforcement Learning

MA Yu-Bo, ZHOU Chang-Dong, ZHANG Zhi-Wen, YANG Pei-Ze, ZHANG Bo

(College of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China)

Abstract: Multi-agent collaboration plays a crucial role in the field of reinforcement learning, focusing on how agents cooperate to achieve common goals. Most collaborative multi-agent algorithms emphasize the construction of collaboration but overlook the reinforcement of individual decision-making. To address this issue, this study proposes an online reinforcement learning model, BiTransformer memory (BTM), which not only considers the collaboration among multiple agents but also uses a memory module to assist individual decision-making. The BTM model is composed of a BiTransformer encoder and a BiTransformer decoder, which are utilized to improve individual decision-making and collaboration within the multi-agent system, respectively. Inspired by human reliance on historical decision-making experience, the BiTransformer encoder introduces a memory attention module to aid current decisions with a library of explicit historical decision-making experience rather than hidden units, differing from the conventional RNN-based method. Additionally, an attention fusion module is proposed to process partial observations with the assistance of historical decision experience, to obtain the most valuable information for decision-making from the environment, thereby enhancing the decision-making capabilities of individual agents. In the BiTransformer decoder, two modules are proposed: a decision attention module and a collaborative attention module. They are used to foster potential cooperation among agents by considering the collaborative benefits between other decision-making agents and the current agent, as

① 收稿时间: 2024-05-22; 修改时间: 2024-06-17; 采用时间: 2024-07-04; csa 在线出版时间: 2024-10-31

well as partial observations with historical decision-making experience. BTM is tested in multiple scenes of StarCraft, achieving an average win rate of 93%.

Key words: multi-agent collaboration; online reinforcement learning; partial observation; historical decision-making experience; collaborative benefit; individual policy enhancement

在强化学习领域[1–3], 多智能体协同[4–6]是一个探索如何使多个智能体在环境中共同作用以达到目标的复杂子领域. 在线多智能体强化学习[7,8]是合作多智能体强化学习 (MARL) 的一种重要算法, 可以有效地处理非平稳性. 当前, 大部分在线 MARL 算法[9,10]通过实时策略迭代[11]、经验回放[12]以及模型自适应[13]等机制, 让智能体能够与环境交互同时学习策略的即时更新和优化. 其中 MADDPG[14]将 DDPG[15]算法扩展到了多智能体系统中, 相较于传统多智能体强化学习的分散训练方式, 它采用集中训练分散执行 (CTDE)[16–18]的方式, 充分考虑多个智能体之间的潜在合作意向, 实现了协同策略的演进. 然而由于其不能反映每个智能体对团队的贡献度, 学者们提出了 COMA[19], 基于反事实差异评估特定智能体的没有采用特定行动会对结果带来什么影响, 来衡量每个智能体的具体贡献. 但是, COMA 的信用分配方法是基于全局状态, 与大部分 MARL 环境的局部观察性相悖. MAPPO[20]不仅维持了信用分配的效能, 还通过更为灵活的状态表征手段优化了智能体基于局部观察下的决策过程. 最近几年注意力模型在深度学习领域取得了显著的成效, MAAC[21]算法将注意力机制引入到多智能体强化学习中, 显著提升了算法在处理多智能体复杂交互中的学习效率和策略适应性.

尽管目前的方法在某种程度上取得了一定的成效, 然而, 大多数 MARL 算法[22–24]主要集中在宏观层面的协作形成上, 而忽略了在个体层面增强智能体策略对团队策略的演进支持的重要性. 此外, 在多智能体系统 (MAS) 中, 如何动态学习适应每一次决策过程中与队友合作的程度对 MAS 最终的性能也有重要的影响[25,26].

为了解决上述问题, 本文提出了一个 BiTransformer 记忆模型 (BTM). 该模型包括双注意力编码器 (BTE) 和双注意力解码器 (BTD) 两个部分, 其中 BTE 由记忆注意力模块 (MF) 和融合注意力模块 (FF) 组成. MF 依据生活中的经验, 阴天要带伞这种行为是基于过去经验的结果. 因此, 基于这一现象, MF 引入了历史决策经验模块, 专注于对历史决策经验的分析、存储, 支持决策经验的持续更新和有效保留, 并为每个智能体配备了定长的记忆空间, 利用先入先出的方式进行更新. 最终实现决策经验辅助下的局部观测的特征提取, 这是本文实现个体决策增强的重要的一步. FF 利用注意力分析当前观测为己方团队通信分配机制, 从而为对当前决策过程更具价值的决策个体和环境信息分配更高的关注度. 在 BTD 中, 本文考虑到智能体如果能够察觉到其他智能体的行为更有利于智能体做出决策, 提出了动态策略调整机制, 首先由合作注意力模块 (CF) 对已经做出决策的“前辈”智能体的行为策略进行分析, 通过注意力机制计算并分配与不同“前辈”智能体的合作强度, 动态调整智能体的策略进而获取最大的合作收益. 最后在决策注意力模块 (PF) 中, 通过合作收益与交互状态获取决策序列.

1 背景

1.1 多智能体在线强化学习

在线多智能体强化学习[27,28]可以转换成局部可观测的马尔可夫决策过程$<N,O,A,R,P,\gamma>$[29], 其中$N=\{1,\cdots,n\}$为智能体集合. $O=O_1\times O_2\times\cdots\times O_n$表示联合观测空间, 即局部观测空间的集合. $R:O\times A\to[-R_{\max},R_{\max}]$为智能体联合奖励函数. $P:O\times A\times O\to[0,1]$表示状态转移函数. 智能体根据以下协议与环境交互: 在$t\in T$时间步, 第$i\in N$个智能体获得一个观察值O_t^i, 并根据策略π_i来选择动作a_t^i. 每一个时间步结束, 所有智能体会获得一个整体奖励为$R(O^t,A^t)$和下一时刻状态O^{t+1}. 最终目标是使得智能体学习协作策略, 使得累积回报$\sum_{t=0}^{\infty}\gamma^t R(O^t,A^t)$最大化, 其中$\gamma\in[0,1)$表示折扣因子.

1.2 广义优势估计

优势函数$A(s,a)$定义为选择动作a在给定状态s下相较于平均策略行为所带来额外的期望汇报, 具体如式 (1) 所示:

$$A(s,a)=Q(s,a)-V(s) \tag{1}$$

其中, $Q(s,a)$表示动作价值函数, 即在当前状态s下采

取动作a后所能获得的预期回报; $V(s)$为状态价值函数, 表示在给定策略下, 从当前状态s开始的预期回报. 广义优势估计 (GAE)[30]通过加权累积来平衡期望回报的方差和估计偏差, 可以更有效地估计优势函数. 其核心是在估计优势函数时, 引入了参数λ用于调节多步时间差分 (TD) 估计, 具体如式 (2) 所示:

$$\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V \tag{2}$$

其中, $\delta_t^V = R_t + \gamma V(s_{t+1}) - V(s_t)$是 TD 残差, R_t是时间t的奖励, λ是超参数, 用于控制 TD 估计的偏差和方差之间平衡.

1.3 注意力机制

注意力机制[31]包括查询 (q)、键 (k) 和值 (v). 对于给定的q, 注意力机制通过计算与k之间的相似度或相关性, 并将该相关性作为权重分配给v, 然后这些权重将用于加权汇总对应的值, 以此产生加权的输出, 这个输出即为模型应当“关注”的信息. 具体计算如式 (3) 所示:

$$Attention(q,k,v) = Softmax\left(\frac{qk^{\mathrm{T}}}{\sqrt{d_k}}\right)v \tag{3}$$

其中, d_k表示q和k的维度.

2 方法

本文提出的 BTM 旨在强化智能体自身决策能力的同时促进多智能体合作的形成, 整体架构如图 1 所示. 该模型由 BTE 和 BTD 组成, 其中 BTE 通过引入记忆网络处理历史决策经验$\{M_i\}_{i=1:n}$, 并以此作为辅助分析局部观测$\{O_i\}_{i=1:n}$, 从而得到交互状态$\{\hat{O}_i\}_{i=1:n}$以及交互状态状态价值$\{V(\hat{O}_i)\}_{i=1:n}$. 解码器则通过评估“前辈”智能体决策$\{a_t^i\}_{i=1:j-1}$与当前智能体j的合作收益, 动态调整当前智能体的策略, 并最终与交互状态结合, 得到最终决策.

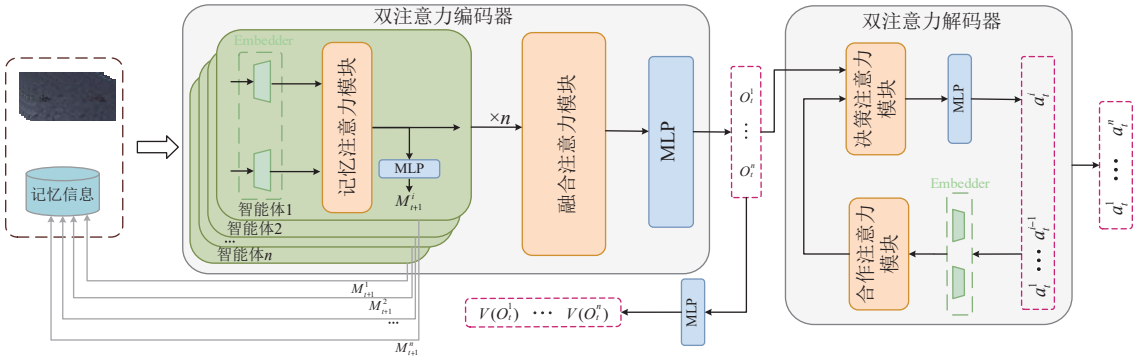

图 1 BTM 结构示意图

2.1 双注意力编码器

BTE 负责处理分析输入的观测以及历史决策经验, 以生成更具决策价值的交互状态. BTE 增强了智能体处理环境信息的能力, 提高了智能体的个体决策, 为后续解码器提供了有效输入.

记忆注意力模块: 本文设计了记忆注意力模块, 该网络利用注意力机制在历史决策经验$M_t^i = \{m_{t-1}^i, m_{t-2}^i, \cdots, m_{t-k+1}^i\}$的辅助下对当前观测$O_t^i$进行分析, 具体推理如式 (4) 所示:

$$m_t^i = \psi(\kappa(O_t^i, M_t^i)) \tag{4}$$

其中, κ表示 MF, ψ表示全连接层.

此外, 该模块为每个智能体开辟了固定的记忆空间, 并采用队列先入先出的方式进行更新, 具体如图 2 所示. 当记忆空间大小$k=3$时, m_{t-1}^i从$t-1$时刻参与训练, 直到$t+2$时刻才被m_{t+2}^i替换, 这确保了对历史决策经验的有效保留.

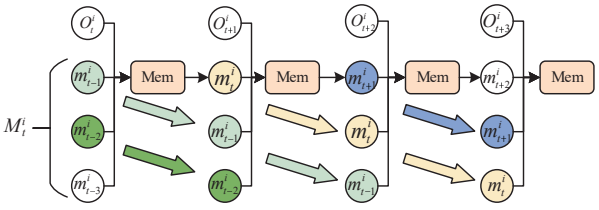

图 2 记忆模块框架图

融合注意力模块: 本文设计了融合注意力模块, 旨在促进智能体之间的交流并为智能体的决策过程提供更具价值的环境信息, 融合注意力模块综合处理所有智能体的融合历史决策经验的观测, 提高了智能体的交流同时, 为智能体做决策提供了具有重要决策价值

的交互状态$\{\hat{O}_i\}_{i=1:n}$.

最终, 本文通过最小化经验贝尔曼误差来训练BTE, 如式 (5) 所示:

$$L_{\mathrm{BTE}}(\varphi)=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=0}^{T-1}[R(O^t,A^t)+\gamma V_{\bar{\varphi}}(\hat{O}^i_{t+1})-V_{\varphi}(\hat{O}^i_t)]^2 \tag{5}$$

其中, $V_{\varphi}(\hat{O}^i_t)$表示当前交互观测$\hat{O}^i_t$下的状态价值函数, 其中φ是状态价值函数的参数. $V_{\bar{\varphi}}(\hat{O}^i_{t+1})$表示下一个交互观测$\hat{O}^i_{t+1}$下的目标状态价值函数, 其中$\bar{\varphi}$是目标状态价值函数的参数. T表示环境运行的时间步数.

BTE 的方程如式 (6) 所示:

$$\hat{L}^t_i=\chi(L^t_1,L^t_2,\cdots,L^t_n;\varphi) \tag{6}$$

其中, χ表示 BTE, $L^i_t=(O^i_t,M^i_t)$, $\hat{L}^i_t=(\hat{O}^i_t,M^i_{t+1})$, φ为 BTE 的参数.

2.2 双注意力解码器

BTD 的核心为利用 BTE 计算的交互状态$\{\hat{O}_i\}_{i=1:j}$与已有的“前辈”智能体的动作$\{a^i_t\}_{i=1:j-1}$, 来生成当前智能体动作a^j_t. 这一过程确保了每个智能体的动作选择不只是基于环境的当前观测, 还融入了智能体间的协作信息和先前的决策经验, 从而使得每个智能体学会更具协同性和适应性的策略.

合作注意力模块: 对第j个智能体而言, 由合作注意力模块分析“前辈”智能体的动作$\{a^1_t,\cdots,a^{j-1}_t\}$, 得到当前智能体与“前辈”智能体的合作收益, 从而为其决策过程提供了重要参考. 以战斗状况为例, 当发现已有两个智能体正在针对同一目标进行攻击, 智能体会较大概率选择此目标, 因为与这两个智能体形成的协同作战有望实现更优越的战果. 值得注意的是, 在为了保证合作注意力模块输入的一致性, 我们在后$n-j+1$个智能体的位置上执行了补 0 操作.

决策注意力模块: 通过决策注意力模块分析与“前辈”智能体的合作收益来处理交互状态, 生成当前智能体的行动决策. 经过n轮循环迭代, 最终形成智能体的联合动作$\{a^i_t\}_{i=1:n}$.

为了训练 BTD, 最小化以下基于裁剪的 PPO 目标, 如式 (7) 所示:

$$L_{\mathrm{BTD}}(\theta)=-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\min(r^i_t(\theta)\hat{A}_t,\ \mathrm{clip}(r^i_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t) \tag{7}$$

其中, $r^i_t(\theta)=\pi(a^i_t|\hat{O}_t,\hat{a}^{i-1}_t)/\pi_{\mathrm{old}}(a^i_t|\hat{O}_t,\hat{a}^{i-1}_t)$, $\hat{a}^{i-1}_t$表示前$i-1$个智能体所采取的动作. $\pi(a^i_t|\hat{O}_t,\hat{a}^{i-1}_t)$表示第$i$个智能体在观测$\hat{O}_t$, 前$i-1$个智能体采取$\hat{a}^{i-1}_t$的情况下采取动作$a^i_t$的概率. π_{old}表示上一次的策略. ε是一个超参数, 本文设置为 0.2, 用于裁剪 PPO 的幅度. $\mathrm{clip}(x,a,b)=\max(\min(x,b),a)$将$x$限制在区间$[a,b]$之间. 优势函数$\hat{A}_t$计算如式 (2) 所示. BTD 的推理过程如式 (8) 所示:

$$a^j_t=\xi(\hat{O}^i_t,a^i_t;\theta)_{i=1:j-1} \tag{8}$$

其中, ξ表示 BTD, θ为 BTD 的参数.

BTM 采用端到端的训练模式, 具体来说, 利用 BTE 将局部观测和历史决策经验映射到高维空间, 随后 BTD 将高维特征映射到动作空间完成决策. 在每一轮后, 利用 TD 算法来计算损失, 然后使用 Adam 优化器更新模型参数, 训练过程如算法 1 所示.

算法 1. BTM 算法

输入: 轮数K, 每一轮的步数T, 批次数大小B, 智能体的数量n.
输出: 智能体的动作$\{a^i_t\}_{i=1:n-1}$.

初始化: 编码器参数φ, 解码器参数θ, 经验回访池D, 历史决策经验$\{M_i\}_{i=1:n}$
1) For k=0 to K−1 do
2) For t=0 to T−1 do
3) 从环境中获得一系列观测值$\{O_i\}_{i=1:n}$
4) 获取历史决策经验$\{M^t_i\}_{i=1:n}$
5) 式 (6) 计算每个智能体的$\hat{O}^i_t$与M^i_{t+1}
6) For j=1 to n do
7) 式 (8) 计算智能体动作a^j_t
8) End for
9) 执行动作$\{a^i_t\}_{i=1:n-1}$并得到奖励$R(O_t,a_t)$
10) $(O_t,a_t,M_t,R(O_t,a_t))$放入经验回放池D
11) End for
12) 从D中随机抽取B个样本子集
13) 式 (6) 计算$\{V(\hat{O}_i)\}_{i=1:n}$
14) 式 (5) 计算 BTE 的损失$L_{\mathrm{BTE}}(\varphi)$
15) 式 (2) 计算优势函数$\hat{A}_t$
16) BTD 利用$\{O_i\}_{i=1:n}$和$\{a^i_t\}_{i=1:j-1}$生成策略a^j_t
17) 式 (7) 计算$L_{\mathrm{BTD}}(\theta)$
18) $\varphi\leftarrow\mathrm{Adam}(L_{\mathrm{BTE}}(\varphi)),\ \theta\leftarrow\mathrm{Adam}(L_{\mathrm{BTD}}(\theta))$
19) End for
20) Return $\{a^i_t\}_{i=1:n-1}$

3 实验

在本节中, 评估了 BTM 在星际争霸中的表现. 首先将 BTM 与典型的 MARL 模型进行了对比实验. 其次, 对 BTM 的各个组件进行了消融实验, 以验证 BTM 的有效性.

3.1 星际争霸多智能体挑战

星际争霸多智能体挑战 (SMAC)[32]是暴雪星际争霸上进行的多智能体协同强化学习环境. SMAC 具有庞大的状态、观测空间和动作空间, 对多智能体协同算法的泛化性和适应性提出了挑战. 图 3 展示了 SMAC 中部分场景.

图 3 SMAC 场景图

3.2 性能研究

图 4 展示了在 SMAC 的不同场景下, BTM 与基线方法的胜率对比折线图. 实验结果为在不同随机种子的情况下进行 10 次实验取中位数为评价指标. 值得注意的是, HAPPO 和 HATRPO 仅在难度简单的 2s_vs_1sc 场景上表现出色. 而在难度较高的 5m_vs_6m 等场景中, HAPPO 和 HATRPO 方法之所以在复杂场景表现不佳, 源于它们采用的共享参数训练机制, 这导致所有智能体在能力上无法明显区分, 与独立网络的训练方法相比, 它们在信用分配上显得过于简化, 影响了决策精度和适应性. 相较于同样使用注意力机制的 TransQMix, BTM 在收敛速度和效果上都更加优秀, 证明了历史决策经验的必要性. 在测试场景中, BTM 的训练曲线相对于 MAPPO 等来说增长较慢, 这是因为训练初期历史决策经验不足, 导致其性能提升缓慢. 然而, 随着历史决策经验的逐步融合, BTE 的优势逐渐显现, 最终模型性能超越了 MAPPO 等其他基线模型. BTM 在 5m_vs_6m 场景中取得了次优的表现, 仅低于 EMU, 但是在其他场景中, 尤其是在像 6h_vs_8z 这类困难地图中, BTM 的效果远高于 EMU, 因为使用了前辈智能体的合作收益的情况下, 增强了模型的泛化性.

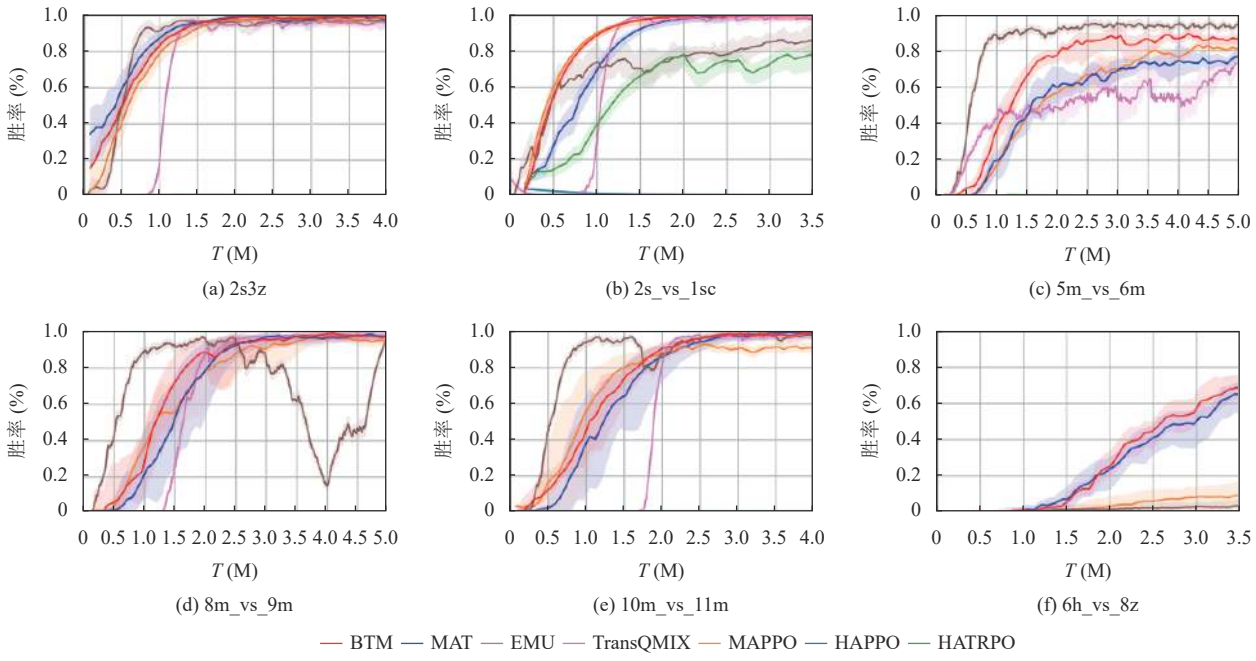

图 4 SMAC 场景的测试胜率

此外, 本文不仅与在线模型进行了比较, 还与离线模型进行比较, 以全面评估 BTM 的性能, 结果如表 1 所示. BTM 的总体胜率以 93% 的平均胜率优于其他的 10 种方法. 除 HAPPO 和 HATRPO 外, 其他模型在简单场景下表现良好, 但随着场景难度的增加, 智能体间的协作和各自的独立决策变得更为关键. BTM 利用记忆模块管理历史决策经验, 提高智能体在复杂环境中的决策. 同时, 处理“前辈”智能体的动作收益增强了智能体的决策能力, 提升了协作效率. 通过这些模块, 使得 BTM 在极具挑战性的场景 6h_vs_8z 中实现了高达 67% 的胜率, 相较于其他模型的平均胜率高出 55%.

3.3 BTM 性能分析

本节计算了 BTM 的浮点运算次数 (FLOPs) 和参

数量, 由于不同的场景会产生不同的观测输入, 导致模型的参数量与复杂度有差异性, 因此表 2 中展示的结果是不同场景的平均值. 最终 BTM 的 FLOPs 和参数量分别为 1.12×10^7 和 1.86×10^5, 具体如表 2 所示.

表 1 SMAC 场景测试胜率 (%)

算法类别	模型	场景类别						
		2s3z	2s_vs_1sc	10m_vs_11m	5m_vs_6m	8m_vs_9m	6h_vs_8z	平均胜率
在线算法	BTM	100	100	100	90	99	67	93
	EMU[34]	98	86	100	56	96	5	74
	MAT[12]	100	100	100	76	90	78	91
	MAPPO[18]	100	100	100	81	99	18	83
	HAPPO[9]	0	0	0	0	0	0	0
	HATRPO[9]	0	17	0	0	0	0	3
离线算法	TransQMix[17]	100	100	97	50	88	0	73
	QMIX[22]	100	100	94	72	88	5	77
	VDN[11]	100	98	94	68	90	5	76
	Qtran[23]	100	85	59	70	88	4	68

表 2 BTM 的 FLOPs 与参数量

网络结构	浮点数运算	参数量
BTE	9.3×10^6	1.27×10^5
BTD	1.9×10^6	5.89×10^4
BTM	1.12×10^7	1.86×10^5

3.4 消融实验

为了证明 BTM 模块的有效性, 在 5m_vs_6m、8m_vs_9m 和 6h_vs_8z 的场景上做以下两个实验: (1) 将 BTM 与 BT (移除 BTM 中记忆注意力模块)、BTMbase (移除 BTM 中解码器模块, 给每一个智能体分配一个全连接层) 和 BTbase (在 BT 的基础上, 移除解码器模块, 给每个智能体分配一个全连接层); (2) 探究记忆块的长度对模型性能的影响. 两个实验的结果均采用不同随机种子并且进行 10 次训练取平均值, 结果如图 5 所示. BT 的 SMAC 平均胜率显著下降, 这种性能下降可能是由于模型缺乏足够的历史决策经验, 导致智能体做出了有偏差的决策. 与此同时, 当使用 BTMbase 时, SMAC 的平均胜率也出现了明显下降. 这种效果变差可能是因为智能体并没有考虑到与其他智能体之间的合作, 导致了胜率的下降. BTbase 整体模型的效果下降了较大幅度这表明这两个模块对模型性能的贡献是相互叠加的, 去掉它们会导致更大程度的性能下降. 如表 3 所示, 没有记忆注意力模块和 BTD 的时候, 模型的平均胜率分别降低 10% 和 7.6%. 这证明了这两个模块对于模型性能的重要性. 当同时去掉这两个模块时, 整体模型的效果下降 27.6%. 同时, 也分析了不同的时间步长度对模型的影响. 当时间步长度为 1 时, 胜率降低了 22%, 是因为模型训练过程中所能够利用的决策经验相对较少, 所导致的性能变差. 而当时间步长度为 5 时, 胜率降低 19%, 这因为模型在处理更长的决策经验时需要更多的时间来学习有效的信息, 而且过多的冗余信息会对模型的性能带来负面影响. 总体而言, 通过调整时间步长度, 可以看到在时间步长度为 3 时获得了最佳效果, 相较于其他两个长度的时间步, 胜率分别提高 22% 和 19%.

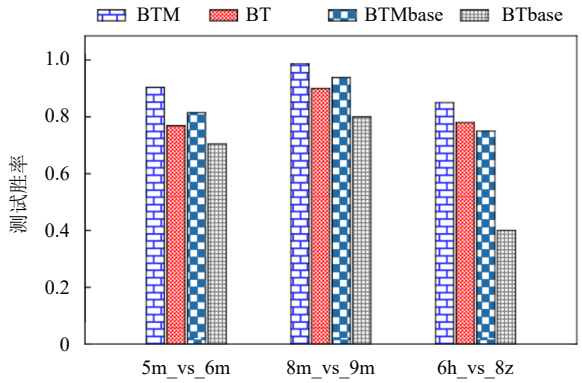

图 5 BTM 组件消融实验示意图

表 3 消融实验 (%)

消融类别	实验配置	场景类别		
		5m_vs_6m	8m_vs_9m	6h_vs_8z
模型结构	BTM	90	99	67
	BT	76	90	64
	BTMbase	82	94	61
	BTbase	71	80	40
时间步长度	1步	75	81	52
	3步	90	99	67
	5步	82	85	50

4 结束语

本文提出了 BiTransformer 记忆在线强化学习模型, 来解决多智能体协同问题. BTM 由编码器和解码器组成, 首先, 利用记忆注意力模块分析的智能体的局部观测和历史决策经验, 提取出智能体观测中有助于决策优化的信息. 然后, 将这些信息传入融合注意力模块中, 在历史决策经验的基础上, 加强智能体间的交流, 得到交互状态. 利用合作注意力模块分析当前智能体与“前辈”智能体动作的合作收益, 从而做出更有助于团队合作的决策. 最后由决策注意力模块综合合作收益与交互状态生成智能体的决策. 在 SMAC 的不同场景下平均胜率比其他方法高 33%. 鉴于这是一种新的方法, 它为进一步研究提供了几个开放的途径. 例如, 可以通过引入更复杂的记忆更新策略来提高模型对复杂环境的建模能力未来的工作将专注于增强智能体之间的协同性, 同时提高模型的解释性和可解释性. 这些努力将有助于推动多智能体系统在复杂任务中取得更优秀的表现.

参考文献

1 林谦, 余超, 伍夏威, 等. 面向机器人系统的虚实迁移强化学习综述. 软件学报, 2024, 35(2): 711–738. [doi: 10.13328/j.cnki.jos.007006]

2 Li SE. Deep reinforcement learning. Reinforcement Learning for Sequential Decision and Optimal Control. Singapore: Springer, 2023. 365–402. [doi: 10.1007/978-981-19-7784-8_10]

3 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.

4 Guo J, Chen YH, Hao YH, *et al*. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 114–121. [doi: 10.1109/CVPRW56347.2022.00022]

5 Avalos R. Exploration and communication for partially observable collaborative multi-agent reinforcement learning. Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2022. 1829–1832.

6 Oroojlooy A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning. Applied Intelligence, 2023, 53(11): 13677–13722. [doi: 10.1007/s10489-022-04105-y]

7 Sharma PK, Fernandez R, Zaroukian E, *et al*. Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training. Proceedings of the 2021 Conference on Artificial Intelligence and Machine Learning for Multi-domain Operations Applications III. SPIE, 2021. 117462K. [doi: 10.1117/12.2585808]

8 Zhang KQ, Yang ZR, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In: Vamvoudakis KG, Wan Y, Lewis FL, *et al*., eds. Handbook of Reinforcement Learning and Control. Cham: Springer, 2021. 321–384. [doi: 10.1007/978-3-030-60990-0_12]

9 李茹杨, 彭慧民, 李仁刚, 等. 强化学习算法与应用综述. 计算机系统应用, 2020, 29(12): 13–25. [doi: 10.15888/j.cnki.csa.007701]

10 Li C, Wang T, Wu C, *et al*. Celebrating diversity in shared multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2021, 34: 3991–4002.

11 周毅, 刘俊. 融合强化学习的多目标路径规划. 计算机系统应用, 2024, 33(3): 158–169. [doi: 10.15888/j.cnki.csa.009418]

12 Peng CY, Kim M, Zhang Z, *et al*. VDN: Virtual machine image distribution network for cloud data centers. Proceedings of the 2012 IEEE INFOCOM. Orlando: IEEE, 2012. 181–189. [doi: 10.1109/INFCOM.2012.6195556]

13 Wen MN, Kuba JG, Lin RJ, *et al*. Multi-agent reinforcement learning is a sequence modeling problem. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2024. 1201.

14 Jaakkola T, Singh SP, Jordan MI. Reinforcement learning algorithm for partially observable Markov decision problems. Proceedings of the 7th International Conference on Neural Information Processing Systems. Denver: MIT Press, 1994. 345–352.

15 Kwon D, Jeon J, Park S, *et al*. Multiagent DDPG-based deep learning for smart ocean federated learning IoT networks. IEEE Internet of Things Journal, 2020, 7(10): 9895–9903. [doi: 10.1109/JIOT.2020.2988033]

16 Shakya AK, Pillai G, Chakrabarty S. Reinforcement learning algorithms: A brief survey. Expert Systems with Applications, 2023, 231: 120495.

17 Gallici M, Martin M, Masmitja I. TransfQMix: Transformers for leveraging the graph structure of multi-agent reinforcement learning problems. Proceedings of the 2023 International Conference on Autonomous Agents and

Multiagent Systems. London: International Foundation for Autonomous Agents and Multiagent Systems, 2023. 1679–1687.

18 Yu LL, Li KY, Huo SX, *et al*. Cooperative offensive decision-making for soccer robots based on bi-channel Q-value evaluation MADDPG. Engineering Applications of Artificial Intelligence, 2023, 121: 105994. [doi: 10.1016/j.engappai.2023.105994]

19 Foerster J, Farquhar G, Afouras T, *et al*. Counterfactual multi-agent policy gradients. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 2974–2982. [doi: 10.1609/aaai.v32i1.11794]

20 Yu C, Velu A, Vinitsky E, *et al*. The surprising effectiveness of PPO in cooperative multi-agent games. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1787.

21 Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning. Proceedings of the 36th International Conference on Machine Learning. Long Beach: ICML, 2019. 2961–2970.

22 陈妙云, 王雷, 盛捷. 基于值分布的多智能体分布式深度强化学习算法. 计算机系统应用, 2022, 31(1): 145–151. [doi: 10.15888/j.cnki.csa.008237]

23 Son K, Kim D, Kang WJ, *et al*. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. Proceedings of the 36th International Conference on Machine Learning. Long Beach: ICML, 2019. 5887–5896.

24 Rashid T, Samvelyan M, De Witt CS, *et al*. Monotonic value function factorisation for deep multi-agent reinforcement learning. The Journal of Machine Learning Research, 2020, 21(1): 178.

25 马佩鑫, 程钰, 侯健, 等. 基于多智能体深度强化学习的协作导航应用. 计算机系统应用, 2023, 32(8): 95–104. [doi: 10.15888/j.cnki.csa.009200]

26 Geng MH. Scaling up cooperative multi-agent reinforcement learning systems. Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. Auckland: International Foundation for Autonomous Agents and Multiagent Systems, 2024. 2737–2739.

27 Shen GC, Wang Y. Review on Dec-POMDP model for MARL algorithms. In: Jain JC, Kountchev R, Hu B, *et al*., eds. Smart Communications, Intelligent Algorithms and Interactive Methods. Singapore: Springer, 2022. 29–35. [doi: 10.1007/978-981-16-5164-9_5]

28 Zhang Z, Ong YS, Wang DQ, *et al*. A collaborative multiagent reinforcement learning method based on policy gradient potential. IEEE Transactions on Cybernetics, 2021, 51(2): 1015–1027. [doi: 10.1109/TCYB.2019.2932203]

29 Lauri M, Hsu D, Pajarinen J. Partially observable Markov decision processes in robotics: A survey. IEEE Transactions on Robotics, 2023, 39(1): 21–40. [doi: 10.1109/TRO.2022.3200138]

30 Raileanu R, Fergus R. Decoupling value and policy for generalization in reinforcement learning. Proceedings of the 38th International Conference on Machine Learning. Berlin: ICML, 2021. 8787–8798.

31 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

32 Samvelyan M, Rashid T, de Witt CS, *et al*. The StarCraft multi-agent challenge. Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. Montreal: AAMAS, 2019. 2186–2188.

(校对责编: 孙君艳)