·研究论文·

文章编号：1003－1251(2026)04－0001－07

基于深度强化学习的多智能体射击游戏研究

梁嘉欣，苗好田，李博由，姜月秋

（沈阳理工大学 信息科学与工程学院，沈阳 110159）

摘 要：为解决多智能体射击游戏训练样本效率低、训练不稳定、奖励函数设计困难等问题，改进了MA-POCA(multi-agent posthumous credit assignment)算法，提出了基于时间衰减的分层奖励机制。首先基于Unity3D搭建训练环境，实现智能体与环境的交互，再采用射线传感器及Unity API构建观测系统并设计混合动作空间，实现智能体的自主决策；然后采用基于时间衰减的分层奖励机制改进的MA-POCA算法构建模型，解决长期任务中的信用分配问题，再通过时空注意力机制实现记忆检索，提高战术连续性。仿真实验结果显示，经过3 000万步训练，智能体实现了从个体作战到高级团队协作，掌握了交叉火力等战术行为。优化后的算法显著提高了智能体的战术同步率，可为游戏AI和机器人协作等领域的深入研究提供重要参考。

关 键 词：多智能体；MA-POCA算法；射击游戏；强化学习

中图分类号：TP391.9 **文献标志码**：A **DOI**：10.3969/j.issn.1003－1251.2026.04.001

Research on Multi-agent Shooting Games Based on Deep Reinforcement Learning

LIANG Jiaxin，MIAO Haotian，LI Boyou，JIANG Yueqiu

（Shenyang Ligong University，Shenyang 110159，China）

Abstract：To solve the problems of low sample efficiency，unstable training，and difficulty in designing reward functions in multi-agent shooting games，the MA-POCA (multi-agent posthumous credit assignment) algorithm was improved，and a hierarchical reward mechanism based on time decay was proposed. Firstly，a training environment was built based on Unity3D to enable interaction between the intelligent agent and the environment. Then，a radiation sensor and Unity API were used to construct an observation system and design a hybrid action space to achieve autonomous decision-making of the intelligent agent. Then，the MA-POCA algorithm improved by a hierarchical reward mechanism based on time decay was used to construct a model to solve the credit allocation problem in long-term tasks. The spatiotemporal attention mechanism was then used to achieve memory retrieval and improve tactical continuity. The simulation experiment results show that after 30 million steps of training，the intelligent agent has achieved from individual combat to advanced team collaboration，and mastered tactical behaviors such as cross firepower. The optimized algorithm significantly improves the tactical synchronization rate of the intelligent agent，which can provide important references for in-depth research in fields such as game AI and robot collaboration.

Key words：multi-agent；MA-POCA algorithm；shooting games；reinforcement learning

收稿日期：2025－06－23

基金项目：沈阳理工大学引进高层次人才科研支持计划项目(1010147001225)；教育部供需对接就业育人项目(2023122570529)；辽宁省属本科高校基本科研业务费专项资金资助项目(SYLUGXTD07)

作者简介：梁嘉欣(1999—)，女，硕士研究生；姜月秋(1975—)，通信作者，女，教授，博士，博士生导师。

近年来,人工智能技术以惊人的速度迭代演进,深刻重塑着游戏产业的格局。多智能体系统作为人工智能领域的关键分支,在游戏领域的应用呈现出蓬勃发展的态势[1]。无论是策略类游戏,还是角色扮演游戏,多智能体系统凭借其模拟复杂交互行为的能力,为玩家带来了更加真实、动态的游戏体验,同时也成为游戏开发者提升作品竞争力的重要技术手段。

射击游戏作为游戏市场中最受欢迎的游戏类型之一,以紧张刺激的战斗节奏、丰富多样的战术策略和激烈对抗的竞技体验吸引着大量玩家。在这类游戏中,智能体不仅需要具备快速反应能力,能够在瞬息万变的战场环境中迅速做出判断和决策,还需展现出高超的协作水平,与队友紧密配合以应对敌方威胁[2]。例如,在团队对战模式中,智能体需根据队友位置、敌方分布情况及战场局势,合理分配火力、选择掩护位置,并制定进攻或防守策略。因此,智能体的决策能力需满足极高的要求。传统游戏 AI 主要依赖预先编写的规则和脚本,通过固定的逻辑判断控制智能体的行为,这种方式虽然在一定程度上能够实现基本的游戏功能,但在面对复杂多变的游戏场景时,其局限性日益凸显[3]。例如,当游戏环境出现新的元素或敌方采取非常规战术时,传统 AI 往往无法做出有效应对,导致游戏体验的真实性和趣味性大打折扣。强化学习算法的兴起促进了游戏 AI 的发展,其通过使智能体在环境中不断进行“试错”,根据环境反馈的奖励信号逐步调整策略,实现从经验中学习的能力。这种学习方式赋予了智能体更强的灵活性和适应性,使其能够在复杂、动态的游戏环境中自主探索最优决策。

目前,强化学习在游戏 AI 领域的应用已取得突破性进展,并逐渐发展成为游戏智能决策系统的核心技术范式之一。作为一种通过环境交互学习最优策略的机器学习方法,强化学习在游戏 AI 开发、自动化测试、动态难度平衡及程序化内容生成等多个维度展现出独特的优势和应用潜力。早期的强化学习研究主要聚焦于离散状态空间的简单控制问题[4],随着深度强化学习框架的提出与发展,现代智能体已经能够在高维连续状态空间和复杂动作空间的游戏环境中实现超越专业人类选手的表现。

深度强化学习在策略类游戏中取得了重大突破:AlphaGo[5]结合深度网络与蒙特卡洛树搜索,攻克了围棋这一复杂博弈难题,其改进版 AlphaZero[6]通过纯自我对弈,在多种棋类中达到超人类水平,为后续非完美信息博弈研究奠定了基础[7]。深度强化学习在复杂游戏场景中亦表现优异:OpenAI 的 Five 系统[8]采用多智能体强化学习,在 Dota2 中击败了职业战队;DeepMind 的 AlphaStar[9]结合分层学习与模仿学习,攻克了《星际争霸Ⅱ》的复杂环境问题,展现出处理多智能体协作等核心问题的能力。

尽管强化学习在游戏领域取得了显著成果,但现有研究仍面临若干关键性挑战:深度强化学习算法普遍存在样本效率低下的问题,在复杂游戏环境中往往需要数百万次的环境交互才能收敛[10];多智能体强化学习的训练过程具有内在的非平稳性,易导致策略震荡和收敛困难[11];奖励函数的设计高度依赖领域知识,不当的奖励塑形可能导致策略崩溃[12]。此外,现有研究多集中于棋类或即时战略游戏,针对第一人称射击等快节奏对抗性游戏的多智能体强化学习研究则相对匮乏。

为此,本文构建一个具有物理真实性的战术射击游戏环境,重点研究智能体在动态战场环境中的射击精度、战术机动性和团队协作能力等强化学习训练方法。提出一种基于时间衰减的分层奖励机制改进的 MA-POCA(multi-agent posthumous credit assignment)算法,推进强化学习在实时对抗游戏中的实际应用,并为相关领域的研究提供可复用的技术框架和基准测试环境。

1 强化学习框架

强化学习是继有监督学习和无监督学习之后的又一重要研究方向,该方法借鉴生物学的学习原理,在不断感知、分析、再感知的基础上通过不断尝试寻求累积预期收益最大化的策略[13],其基本框架如图 1 所示。

图 1 强化学习框架

Fig. 1 Reinforcement learning framework

现实问题往往涉及多个智能体之间的交互,单个智能体的强化学习框架不宜直接使用。为此,研究者引入马尔可夫博弈来扩展经典马尔可夫决策过

程,以刻画多智能体环境下的协同或竞争关系[11]。

多智能体情况下,采用马尔可夫博弈算法建模。马尔可夫博弈也称随机博弈(stochastic game, SG),可由多元组 $<\boldsymbol{S},\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n,R_1,R_2,\cdots,R_n,f,\gamma>$ 表示,其中 n 为环境中智能体的数量,$\boldsymbol{S}$ 为环境的状态空间,$\boldsymbol{A}_i(i=1,2,\cdots,n)$ 表示智能体 i 的动作空间,R_i 表示智能体 i 的回报函数,f 表示联合状态转移函数,γ 表示折扣因子,$\gamma\in[0,1]$。用 $\boldsymbol{A}$ 表示所有智能体的联合动作空间,则 $\boldsymbol{A}=\boldsymbol{A}_1\times\boldsymbol{A}_2\times\cdots\times\boldsymbol{A}_n$,联合状态转移函数 f 可表示为

$$f:\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}\rightarrow[0,1] \tag{1}$$

联合状态转移函数决定了在执行联合动作 $\boldsymbol{a}$($\boldsymbol{a}\in\boldsymbol{A}$)的情况下,由状态 $s(s\in\boldsymbol{S})$ 转移到下一个状态 $s'(s'\in\boldsymbol{S}',\boldsymbol{S}'$表示执行动作后的新状态集合)的概率分布。将状态 – 动作 – 下一状态映射到实数集 $\mathbf{R}$,回报函数 R_i 表示为

$$R_i:\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}\rightarrow\mathbf{R} \tag{2}$$

在多智能体环境中,状态转移是所有智能体共同作用的结果。用 $\boldsymbol{a}_{i,k}$表示智能体 i 在时间步 k 的独立动作,用 $\boldsymbol{a}_k$ 表示在时间步 k 时所有智能体的联合动作,则 $\boldsymbol{a}_k$ 表达式为

$$\boldsymbol{a}_k=[\boldsymbol{a}_{1,k}^{\mathrm{T}},\boldsymbol{a}_{2,k}^{\mathrm{T}},\cdots,\boldsymbol{a}_{n,k}^{\mathrm{T}}]^{\mathrm{T}},\boldsymbol{a}_k\in\boldsymbol{A},\boldsymbol{a}_{i,k}\in\boldsymbol{A}_i \tag{3}$$

智能体 i 的个体策略 π_i 表示为

$$\pi_i:\boldsymbol{S}\times\boldsymbol{A}_i\rightarrow[0,1] \tag{4}$$

由个体策略共同构成联合策略 $\boldsymbol{\pi}$。用 $r_{i,k+1}$表示智能体 i 在时间步 k 结束时获得的回报,该回报取决于联合动作,故总回报取决于联合策略。以 $R_i^{\boldsymbol{\pi}}(s)$表示在状态 s 下智能体 i 遵循联合策略 $\boldsymbol{\pi}$ 时所获得的期望总奖励,其表达式为

$$R_i^{\boldsymbol{\pi}}(s)=E\{\sum_{k=0}^{\infty}\gamma^k r_{i,k+1}\mid s_0=s,f\} \tag{5}$$

式中:E 表示期望;s_0 表示初始状态。引入折扣因子 γ 的目的是减少未来奖励对当前决策的影响,γ^k 用于权衡即时奖励和未来奖励的重要性,随着时间步 k 的增加,γ^k 的值逐渐减小,意味着越远的未来奖励对当前决策的贡献越小。

在多智能体系统中,每个智能体都有其 Q 函数,用于评估在特定状态下执行某个动作后获得的长期回报。智能体 i 的 Q 函数取决于联合动作,用 $Q_i^{\boldsymbol{\pi}}$ 表示,表达式为

$$Q_i^{\boldsymbol{\pi}}:\boldsymbol{S}\times\boldsymbol{A}\rightarrow\mathbf{R} \tag{6}$$

智能体 i 在状态 s 执行动作 $\boldsymbol{a}$ 的 Q 函数表示为 $Q_i^{\boldsymbol{\pi}}(s,\boldsymbol{a})$,其表达式为

$$Q_i^{\boldsymbol{\pi}}(s,\boldsymbol{a})=E\{\sum_{k=0}^{\infty}\gamma^k r_{i,k+1}\mid s_0=s,\boldsymbol{a}_0=\boldsymbol{a},f\} \tag{7}$$

式中 $\boldsymbol{a}_0$ 表示初始动作。

2　MA-POCA 算法

为提高强化学习模型的应用效果,本文采用 MA-POCA 算法建模[14],该算法是针对多智能体强化学习中智能体提前终止场景设计的新型架构,旨在解决传统方法中“死后信用分配”难题并提升算法对动态智能体数量的适应性。MA-POCA 算法基于集中训练 – 分散执行框架,对反事实多智能体策略梯度算法进行改进,核心是引入自注意力机制替代含吸收状态的全连接层,以动态处理活跃智能体信息[15]。MA-POCA 结构如图 2 所示。

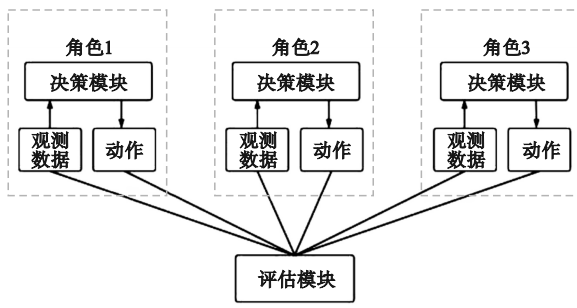

图 2　MA-POCA 结构

Fig. 2　MA-POCA structure

在多智能体协作任务中,智能体可能因“死亡”、资源耗尽等原因提前终止执行任务,导致其无法直接获取后续团队奖励,难以评估自身前期行为对整体目标的贡献,此即“死后信用分配”问题[16]。传统方法通过将终止智能体置于吸收状态使其持续存在于状态空间,从而传递后续奖励信号。但这样会导致两种情况:一是吸收状态作为固定输入会增加神经网络学习复杂度,尤其当智能体数量变化时,全连接层需预设最大数量,导致参数冗余和计算资源浪费;二是吸收状态的数值需与活跃状态严格区分,若取值落入有效观测范围,会引发部分可观测性问题,进一步降低样本效率。

MA-POCA 算法通过实体编码器将各智能体的观测或观测 – 动作对映射至嵌入空间,再通过残差自注意力模块对活跃智能体的嵌入向量进行加权聚合,从而生成集中式价值函数和反事实基线[17]。在价值函数估计中,自注意力机制根据各时间步活跃智能体数量动态处理观测序列,通过递归计算跨时间步的长期回报,将后续奖励反向传播至提前终止的智能体,使其学习到自身行为

与团队未来收益的关联。反事实基线的计算则通过蒙特卡洛采样边缘化单个智能体的动作,利用注意力机制聚合其他智能体的观测 – 动作对,从而评估该智能体对团队奖励的独立贡献,解决信用分配模糊问题。在处理不同数量的智能体时,MA-POCA 采用了两项关键技术:一是基于掩码的注意力机制,在计算注意力权重时,通过将非活跃智能体的键值对权重强制置为负无穷大,确保这些智能体不会影响信息聚合;二是动态批处理策略,将与活跃智能体数量相同的样本分组处理,避免零填充带来计算资源浪费。

上述设计使 MA-POCA 具备两大优势:一是 MA-POCA 在多智能体强化学习领域展现出显著优势,与多智能体深度确定性策略梯度(MADDPG)算法[18]需要固定输入维度和近端策略优化(PPO)算法[19]无法处理团队协作不同,MA-POCA 利用自注意力机制动态处理可变数量的智能体,通过掩码机制仅对活跃智能体进行编码,完全规避了传统方法中吸收状态带来的资源浪费问题;二是注意力机制的动态加权特性可捕获智能体间的复杂依赖关系,提升价值函数和基线估计的准确性,尤其在智能体提前终止或需协作完成长程目标的场景中表现更优。

3 奖励机制

本文提出一种基于时间衰减的分层奖励机制,该机制可通过时序动态调整的奖励函数有效解决战术射击游戏中长期信用分配的难题。采用双层奖励架构设计,其中个体层奖励包括命中奖励 R_{hit}、基础命中奖励 R_{base} 和射击惩罚 R_{shoot}。命中奖励计算采用时间衰减的动态计算模型,表达式为

$$R_{\text{hit}} = R_{\text{base}} \times \alpha(k) \tag{8}$$

式中 $\alpha(k)$ 为时间衰减系数,$\alpha(k) = 2.0 - \beta k / T_{\max}$,其中 β 为时间奖励强度系数(默认 1.5),$T_{\max}$ 为最大步数。基础命中奖励取值为 $R_{\text{base}} = 0.5$,射击动作采用固定惩罚,$R_{\text{shoot}} = -0.1$。

团队层奖励根据胜利情况设计,获胜奖励采用剩余时间比例与剩余人数进行调整,表达式为

$$R_{\text{team}} = \begin{cases} \alpha(k) + R_{\text{num}}, & \text{获胜团队} \\ -1.0, & \text{失败团队} \end{cases} \tag{9}$$

式中:R_{team} 为团队奖励;R_{num} 为队伍存活人数比例,即存活人数与总人数之比。

时间衰减系数 $\alpha(k)$ 的参数设定基于战术射击游戏信用分配问题的理论分析与机制设计目标。β 与奖励值的对应关系如图 3 所示,β 取 1.5 时,既可平衡游戏典型回合时长与智能体学习效率,又可通过合理的奖励衰减节奏避免早期奖励过高导致激进策略或后期奖励过低抑制探索。该取值符合领域内回合制游戏研究结果(β 在 1.2 ~ 1.8 区间可有效引导动态策略调整)[20],兼具理论与实践可行性。经敏感性分析验证,上述参数设置能平衡个体行为激励与团队协作引导,防止单一模式固化,强化时序策略选择意识,保障多智能体协作训练效果。

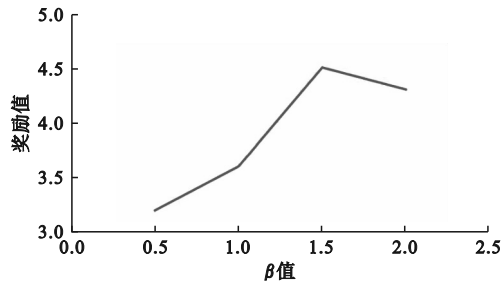

图 3 β 与奖励值关系图

Fig. 3 Relationship between β and reward values

本文设计的基于时间衰减的分层奖励机制优势在于:引入时间衰减系数,动态调整不同阶段行为的奖励权重,使智能体在关键时间节点做出更有价值的决策,同时随着时间推移逐步降低奖励强度,可有效避免早期过度激进或后期过于保守的问题;分层设计的奖励结构将个体表现与团队成果有机结合,既鼓励个人战术执行,又强化团队协作意识,使智能体在追求个人最优与团队最优之间找到平衡点;采用固定射击惩罚与动态命中奖励的搭配设计,在保持行为多样性的同时引导智能体优化战术选择,可避免单一行为模式的固化。此外,该机制的计算复杂度低,仅需维护当前步数状态即可实现奖励计算,既可保证算法效率,又能有效引导长期策略,可为多智能体战术协作训练提供兼具理论合理性和工程实用性的解决方案。基于时间衰减奖励的算法模型架构如图 4 所示。

4 实验与分析

基于 Unity ML-Agents 框架构建的射击环境进行模型训练,场景中包含障碍物、围栏和智能体。智能体状态空间涵盖自身位置、生命值、武器数量、技能冷却、自身所处状态、敌人位置及队友的生命值、相对位置、所处状态及相对速度等信息。通过射线感知等方式获取周围环境信息,如

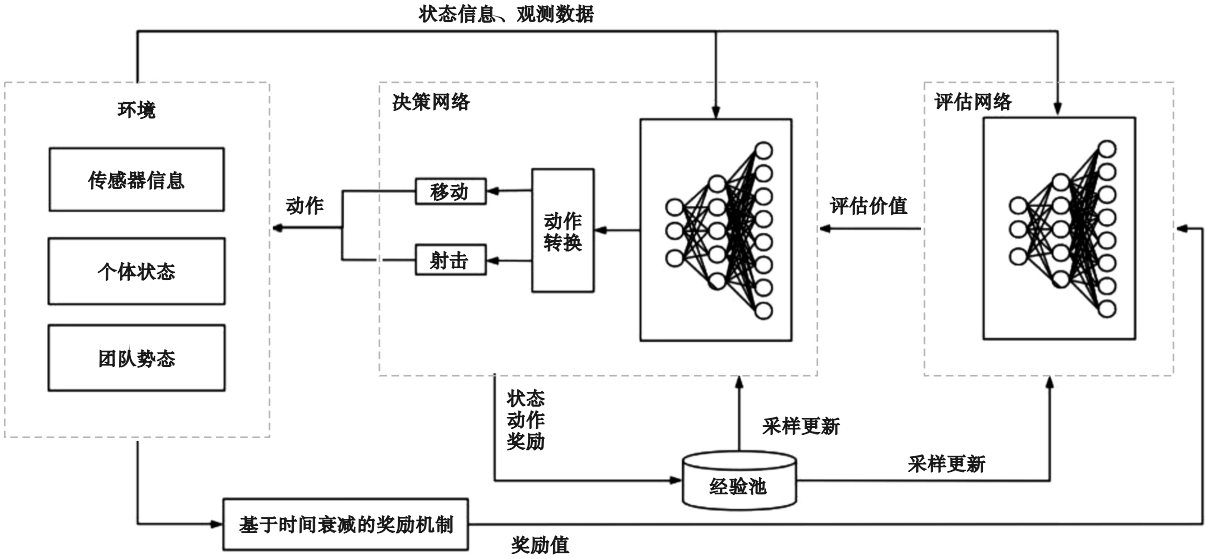

图 4　基于时间衰减奖励的算法模型架构

Fig. 4　Algorithm model architecture based on time decay reward

敌人是否在视野范围内、距离障碍物远近等。射线传感器探测图如图 5 所示。动作空间包括移动、旋转、射击、投掷等连续与离散动作,移动动作可通过连续的速度控制实现,投掷动作为离散操作。游戏场景如图 6 所示。

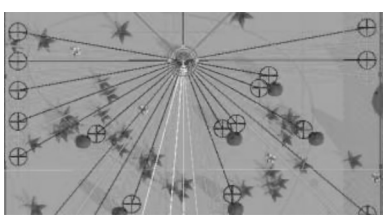

图 5　射线传感器探测图

Fig. 5　Radiation sensor detection map

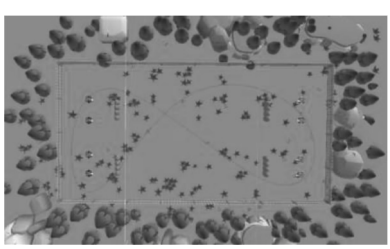

图 6　游戏场景图

Fig. 6　Game scene diagram

基于 Python 的 PyTorch 库实现深度强化学习,网络采用 3 层全连接层,隐藏层维度为 128,折扣因子为 0.99,学习率为 0.000 15,隐藏层神经网络单元个数为 512,神经网络的隐藏层数为 3,视觉编码器使用两层卷积神经网络。

在简单奖励机制($R_{\text{hit}}=0.5$,$R_{\text{shoot}}=-0.1$,团队获胜奖励 $R_{\text{team}}=2.0-k/T_{\max}$,对于失败团队 $R_{\text{team}}=-1.0$)下,使用 MA-POCA 算法和 PPO 算法分别进行模型训练,得到的个体奖励结果如图 7 所示。

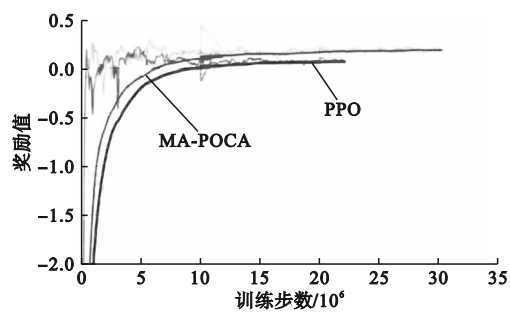

图 7　MA-POCA 与 PPO 模型的个体奖励对比

Fig. 7　Comparison of individual rewards by using MA-POCA and PPO models

由图 7 可见:采用 MA-POCA 算法和 PPO 算法训练得到的模型个体奖励值均随训练步数的增加先升高然后趋于稳定,训练效果均较好;相比 PPO 算法,使用 MA-POCA 算法训练的模型奖励值更高。

图 8 所示为采用 MA-POCA 算法训练的模型组内奖励。可见,组内奖励随着训练步数的增加

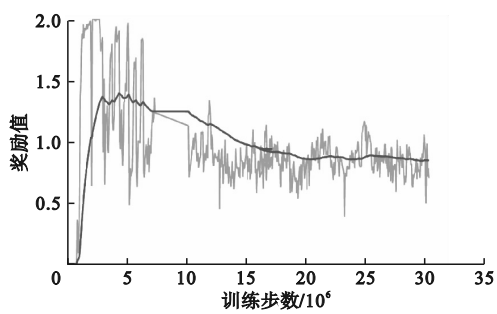

图 8　MA-POCA 模型的组内奖励

Fig. 8　Intra group rewards with MA-POCA model

先升高后降低,先从初始值较快上升至 1.3,然后缓慢降低,最终稳定在 0.8 附近。在上升阶段,智能体学会了基础协作,在稳定波动阶段达到了动态平衡,实现了多智能体训练的预期结果,团队整体策略有效。

在 4v4 的测试场景中使用 MA-POCA 模型对战 PPO 模型,结果显示,MA-POCA 胜率达到 66%。

为进一步验证 MA-POCA 模型的学习效果,在 4v4 测试场景中增加障碍物。采用各阶段模型与训练步数为 30×10^6 的模型进行对打,结果如表 1 所示。可见,随着训练步数的增加,胜率逐渐升高。最终模型与自身副本(最终阶段)对战胜率接近 50%,说明训练收敛,策略达到均衡,可证明模型在正向学习。

表 1 各阶段模型与最终模型对比

Table 1 Comparison of the models of each stage with the final model

阶段模型(训练步数/10^6)	胜率/%
10	14.60
15	20.16
20	35.54
25	42.50
30	47.75

采用本文改进的 MA-POCA 算法进行模型训练,并与原 MA-POCA 算法进行对比,个体奖励及组内奖励对比结果如图 9 和图 10 所示。

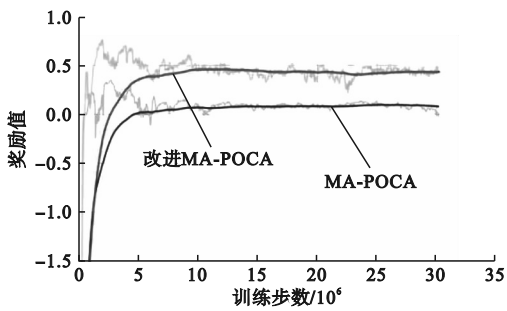

图 9 个体奖励对比

Fig. 9 Comparison of individual rewards

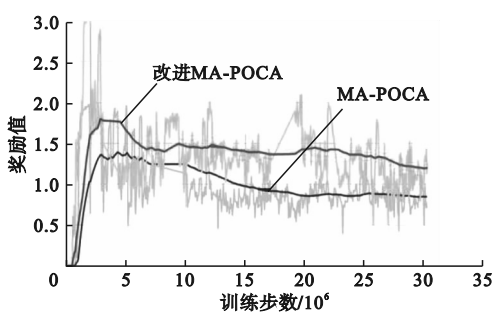

图 10 组内奖励对比

Fig. 10 Comparison of intra group rewards

由图 9 和图 10 可见,不论个体奖励还是组内奖励,本文改进 MA-POCA 模型的奖励值均更高。

在 4v4 的测试场景中使用改进 MA-POCA 模型对战 MA-POCA 模型,结果显示,本文改进 MA-POCA 模型的胜率达到 59%。

图 11 所示为模型的外部价值评估对比图。可见:MA-POCA 模型与本文改进 MA-POCA 模型的外部价值在初期都上升较快,且波动较大,MA-POCA 模型的外部价值最终稳定在 1.5 左右,改进 MA-POCA 模型的外部价值最终稳定在 2.0 左右。综合来看,本文改进的 MA-POCA 模型长期性更优,更稳定。

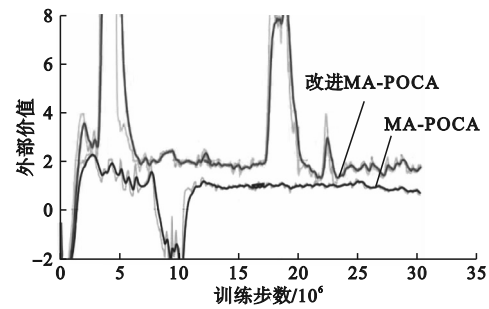

图 11 模型的外部价值评估对比

Fig. 11 Comparison of external value evaluation of models

图 12 所示为模型的 Episode 长度对比。可见:随着训练步数增加,本文改进 MA-POCA 模型的 Episode 长度先快速下降,然后在短暂升高至 125 左右后再次下降,在训练步数为 $10\times10^6\sim20\times10^6$ 之间时稳定在 75 ± 3,训练步数超过 20×10^6 后又上升至 125 左右,然后逐渐回落至 80 左右;MA-POCA 模型的 Episode 长度随着训练步数增加先降至 50,训练步数达到 25×10^6 后异常攀升至 120。综合来看,本文改进模型展现出更优的稳定性。

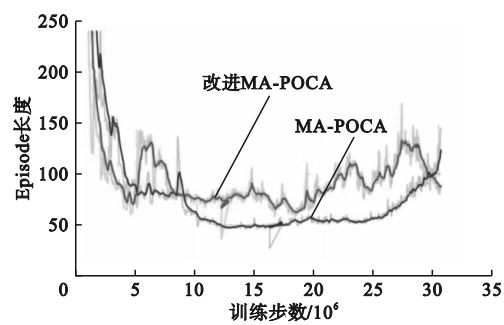

图 12 模型的 Episode 长度对比

Fig. 12 Comparison of Episode length of models

5　结论与展望

提出了一种基于时间衰减的分层奖励机制,改进了 MA-POCA 算法,经改进算法训练后的模型团队奖励收敛更快,智能体学到的合作行为更多、学习用时更少。在动态智能体协作中通过自注意力机制直接进行智能体间交互建模,避免了引入吸收状态造成的样本效率损失,通过反事实基线解决了稀疏奖励下的信用分配歧义问题,明确了智能体牺牲行为的长期价值。

本文研究结果验证了深度强化学习在多智能体协同决策中的有效性,具有重要的理论价值和实践意义,为游戏 AI 开发提供了新的技术路径,其训练框架可拓展应用于虚拟训练、机器人协作等多个领域。但本文改进算法仍存在一些不足:模型约需 200 万步训练才能达到稳定性能,样本效率仍有提升空间;训练周期长,计算成本较高。针对上述问题,未来研究工作将重点在两方面进行改进:其一,引入优先经验回放和课程学习策略,以减少训练步数;其二,探索分布式训练框架,通过多 GPU 并行计算缩短训练时间。此外,未来还会将本文方法扩展应用于异构智能体协作场景中,并探索人机混合团队的训练范式。

参考文献(References):

[1] 庞皓冰,崔林,周建山,等. 基于深度强化学习的空地协同组网与资源优化研究综述[J]. 人工智能,2025,12(1):1-14.

[2] 孙彧,曹雷,陈希亮,等. 多智能体深度强化学习研究综述[J]. 计算机工程与应用,2020,56(5):13-24.
Sun Y,Cao L,Chen X L,et al. Overview of multi-agent deep reinforcement learning[J]. Computer Engineering and Applications,2020,56(5):13-24. (in Chinese)

[3] 李艺春,刘泽娇,洪艺天,等. 基于多智能体强化学习的博弈综述[J]. 自动化学报,2025,51(3):540-558.
Li Y C,Liu Z J,Hong Y T,et al. Multi-agent reinforcement learning based game:a survey[J]. Acta Automatica Sinica,2025,51(3):540-558. (in Chinese)

[4] Mnih V,Kavukcuoglu K,Silver D,et al. Human-level control through deep reinforcement learning[J]. Nature,2015,518(7540):529-533.

[5] Silver D,Huang A,Maddison C J,et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature,2016,529(7587):484-489.

[6] 赵星宇,丁世飞. 深度强化学习研究综述[J]. 计算机科学,2018,45(7):1-6.
Zhao X Y,Ding S F. Research on deep reinforcement learning[J]. Computer Science,2018,45(7):1-6. (in Chinese)

[7] Brown N,Sandholm T. Superhuman AI for multiplayer poker[J]. Science,2019,365(6456):885-890.

[8] Han H G,Zhang Y B,Huang Y T. Collision-free motion-constrained path planning for multiple unmanned delivery vehicles based on heuristic deep reinforcement learning[J]. Neurocomputing,2025,648:130586.

[9] Zou W R. Overview on reinforcement learning of multi-agent game[J]. Journal of Physics:Conference Series,2023,2646(1):012021.

[10] 张艳珠,侯亢钧,陈勇,等. 基于强化学习的改进 RRT* 路径规划[J]. 沈阳理工大学学报,2025,44(4):1-6,12.
Zhang Y Z,Hou K J,Chen Y,et al. Improved RRT* path planning based on reinforcement learning[J]. Journal of Shenyang Ligong University,2025,44(4):1-6,12. (in Chinese)

[11] 许可,吉兰萍,孙文娟,等. 地对空武器-目标分配的多目标决策问题研究[J]. 沈阳理工大学学报,2022,41(5):13-20.
Xu K,Ji L P,Sun W J,et al. Research on multi-target decision-making of ground-to-air weapon-target assignment[J]. Journal of Shenyang Ligong University,2022,41(5):13-20. (in Chinese)

[12] 赵天亮,张小俊,张明路,等. 基于深度强化学习的无人驾驶路径规划研究[J]. 河北工业大学学报,2024,53(4):21-30.
Zhao T L,Zhang X J,Zhang M L,et al. Unmanned driving path planning based on deep reinforcement learning[J]. Journal of Hebei University of Technology,2024,53(4):21-30. (in Chinese)

[13] 孙英博,苗国英,庄亚楠. 基于改进的深度强化学习多智能体协作方法[J]. 传感器与微系统,2023,42(9):25-29.
Sun Y B,Miao G Y,Zhuang Y N. Multi-agent collaboration method based on improved deep reinforcement learning[J]. Transducer and Microsystem Technologies,2023,42(9):25-29. (in Chinese)

[14] Cohen A,Teng E,Berges V P,et al. On the use and misuse of absorbing states in multi-agent reinforcement learning[PP/OL]. arXiv(2021-11-10)[2025-06-10]. https://doi.org/10.48550/arXiv.2111.05992.

[15] 张耐民,蔡秉辰,于浛,等. 基于多智能体强化学习的对抗博弈技术综述[J]. 海军航空大学学报,2024,39(4):395-410.
Zhang N M,Cai B C,Yu H,et al. Review of adversarial game techniques based on multi-agent reinforcement learning[J]. Journal of Naval Aviation University,2024,39(4):395-410. (in Chinese)

[16] 白天,吕璐瑶,李储,等. 基于深度强化学习的游戏智能引导算法[J]. 吉林大学学报(理学版),2025,63(1):91-98.
Bai T,Lü L Y,Li C,et al. Game intelligent guidance algorithm based on deep reinforcement learning[J]. Journal of Jilin University (Science Edition),2025,63(1):91-98. (in Chinese)

[17] 曹毅,郭银辉,李磊,等. 基于深度强化学习的机械臂避障轨迹规划研究[J]. 机械传动,2023,47(12):40-46,96.
Cao Y,Guo Y H,Li L,et al. Deep reinforcement learning-based trajectory planning for manipulator obstacle avoidance[J]. Journal of Mechanical Transmission,2023,47(12):40-46,96. (in Chinese)

[18] Wan K F,Wu D W,Zhai Y W,et al. An improved approach towards multi-agent pursuit-evasion game decision-making using deep reinforcement learning[J]. Entropy,2021,23(11):1433.

[19] 秦湖程,黄炎焱,陈天德,等. 基于 PPO 算法的集群多目标火力规划方法[J]. 系统工程与电子技术,2024,46(11):3764-3773.
Qin H C,Huang Y Y,Chen T D,et al. Clustermulti-target fire planning method based on PPO algorithm[J]. Systems Engineering and Electronics,2024,46(11):3764-3773. (in Chinese)

[20] Silver D,Hubert T,Schrittwieser J,et al. A general reinforcement learning algorithm that masters chess,shogi,and Go through self-play[J]. Science,2018,362(6419):1140-1144.

(责任编辑:宋颖韬)