中国科学: 技术科学 2023 年 第 53 卷 第 3 期: 405 ~ 416
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
论 文

基于深度强化学习的不完美信息群智夺旗博弈

王健瑞, 黄家豪, 唐漾*

华东理工大学信息科学与工程学院, 能源化工过程智能制造教育部重点实验室, 上海 200237
* E-mail: yangtang@ecust.edu.cn

收稿日期: 2021-08-15; 接受日期: 2022-01-27; 网络版发表日期: 2023-02-23
国家自然科学基金基础科学中心项目(批准号: 61988101)、国家杰出青年科学基金项目(批准号: 61925305)和重点国际(地区)合作研究项目(编号: 61720106008)资助

摘要 复杂环境中群智博弈问题是近年来的研究热点之一. 为解决不完美信息条件下多智能体夺旗博弈问题, 本文提出了一种基于多智能体双重决斗深度Q网络(multi-agent dueling double deep Q-network, MAD3QN)以及图注意力网络(graph attention network, GAT)的多智能体夺旗博弈深度强化学习算法(G-MAD3QN). 该算法在实现多智能体在迷宫地图中路径规划的同时, 建模不完美信息条件下多智能体合作与竞争关系, 从而确定夺旗博弈策略. 在实验中, 本文基于二维迷宫环境, 考虑智能体观测信息不完美条件, 将G-MAD3QN算法与多智能体深度Q网络(multi-agent deep Q-network, MADQN)、MAD3QN等多智能体深度强化学习的基线算法进行对比, 从而验证了在二对二夺旗博弈中本文G-MAD3QN算法的有效性.

关键词 夺旗博弈, 不完美信息, 深度强化学习, 图注意力网络

1 引言

博弈是参与者在合作或竞争行为下的策略优化过程[1], 目前广泛应用于政治、经济、军事、计算机科学等领域, 例如国际外交[2]、人机对抗[3]、囚徒困境[4]等. 特别地, 多智能体博弈, 如追逃博弈[5]、触及-避免博弈[6]、夺旗博弈[7]等, 在军事领域具有极大的研究潜力. 以导弹制导问题为例, 将博弈论引入多导弹集群作战时, 亟待研究在动态博弈框架下的多导弹攻防问题, 以提升强对抗环境下集群作战的智能性与抗扰性, 从而实现精准打击与防御. 将导弹拦截问题建模为单阶段夺旗博弈模型, 其中攻击导弹实现对敌方目标的袭击, 而防御导弹则试图摧毁攻击导弹以阻止袭击. 当博弈是低维时, 人们往往尝试以解析的形式求解夺旗博弈的最优策略. Huang等人[7]将夺旗博弈刻画为微分博弈模型, 并通过求解哈密顿-雅可比-埃萨克斯(Hamilton-Jacobi-Isaacs, HJI)方程为参与者设计最优策略并构建获胜区域. 然而, 求解HJI方程需要通过复杂的计算以获得最优策略, 因此Garcia等人[8]提出了夺旗问题的简化. 虽然他们将安全返回区域简化为安全边界, 并提出了夺旗第一阶段和第二阶段的分离优化策略, 但是并未求解博弈的获胜区域. Pachter等人[9]将获胜区域简化为点, 并限制攻击者与防守者的速度相同. 现有文献中[7~9]使用解析手段求解夺旗

引用格式: 王健瑞, 黄家豪, 唐漾. 基于深度强化学习的不完美信息群智夺旗博弈. 中国科学: 技术科学, 2023, 53: 405–416
Wang J R, Huang J H, Tang Y. Swarm intelligence capture-the-flag game with imperfect information based on deep reinforcement learning (in Chinese). Sci Sin Tech, 2023, 53: 405–416, doi: 10.1360/SST-2021-0382

博弈的方法大多都有两个共同点: 一对一夺旗, 地图中无复杂障碍物. 将夺旗博弈推广到复杂场景下多对多博弈的情况时, 由于智能体的最优策略往往取决于其他智能体的选择, 同时复杂地图造成HJI方程维数爆炸等问题, 导致目前无法解析求解多智能体夺旗博弈的最优策略. 为解决这些挑战, 本文考虑在复杂离散迷宫地图条件下二对二夺旗博弈, 通过端到端地训练智能体探索夺旗策略, 避免了求解HJI方程的复杂计算, 鼓励智能体间动态选择偏向合作或竞争的行为.

为解决环境图像输入下离散迷宫空间的路径规划问题, 本文考虑引入深度强化学习方法端到端地训练多智能体探索复杂环境. Minh等人[10]提出解决离散空间路径规划的深度Q网络(deep Q-network, DQN), 智能体成功地从Atari游戏图像中直接学习到控制策略, 并在某些游戏设置下的表现超越了人类专家. Tai等人[11]利用DQN和来自RGB-D传感器的深度图像, 实现了智能体对离散环境的探索与避障. 为了解决DQN中价值函数高估动作值的缺点, Van等人[12]提出了双重深度Q网络(Double DQN)算法, 将选择与评价分离以减少动作值高估. Schaul等人[13]开发了一个优先经验框架, 该框架在DQN中使用优先经验重放, 通过考虑经验的重要性保证了算法的收敛性. Wang等人[14]提出了决斗深度Q网络(Dueling DQN)算法, 它们更改了DQN的底层网络结构, 使得在出现多个相似动作的情况下, 智能体依然能够获得准确的策略评估. 现有文献[15~18]结合上述DQN及其演化算法的优点, 使用双重决斗深度Q网络(Dueling Double DQN, D3QN)以解决环境图像输入下复杂离散空间的探索问题, 在减少动作值高估与网络偏差的同时, 保证了算法收敛与准确估计. 具体而言, Han等人[15]与Ruan等人[16]使用D3QN为智能体在给定地图中规划路径的同时避免碰撞. Chen等人[17]在使用D3QN训练智能体避障导航的基础上, 增加了课程学习策略, 帮助智能体在多障碍环境中适应由易到难的避障任务. Li等人[18]应用D3QN探究基于行为信息的智能体导航与避障问题. 因此, 本文选用深度强化学习D3QN算法在二维迷宫地图中实现路径规划与避障.

由于单智能体深度强化学习在具有高维输入和较大状态空间的场景中已取得显著成功[15~17], 为解决多智能体动态规划与避障, 将深度强化学习方法从单智能体直接拓展到多智能体是一个直观可行的思路[18,19]. 针对每一个智能体, 通过将其他智能体视为环境, 构造相同奖励从而使用同一单智能体深度强化学习网络来迭代训练. 然而, 在博弈的框架下, 多智能体间往往存在合作或竞争的关系, 这意味着合作智能体与竞争智能体拥有不同的奖励函数, 从而无法使用同一深度强化学习网络进行策略搜索[20]. 为解决该问题, 现有文献[21~26]考虑将多智能体重新划分为若干队性质相同的多智能体. 对于同队智能体而言, 由于具有相同的奖励函数、目标等条件, 此时可使用同一单智能体深度强化学习算法训练该队智能体得到最优策略. 比如, Tampuu等人[21]分两个独立的DQN训练智能体实现像人一样玩乒乓游戏. 基于每个智能体的离散动作空间和全局观测, 通过调整奖励函数以实现合作或竞争. 后来, Leibo等人[22]在社会困境的背景下训练独立的DQN. 这项工作表明, 合作或竞争的环境不仅可以影响离散的动作, 还可以改变同队智能体的整体策略. 文献[23~25]在该训练思想的基础上针对强化学习算法进行多种拓展, Bansal等人[23]使用近端策略优化算法在MuJoCo模拟器上分别训练每个智能体, 从而完成一对一竞争任务. 使用密集探索奖励, 鼓励智能体学习基本、非竞争性的行为, 并随着时间的推移减少该类型奖励, 从而给予环境、竞争性奖励更多权重. Jin等人[24]使用深度确定性策略梯度算法训练每个智能体学习合作避障. Yan等人[25]使用D3QN算法单独训练智能体, 利用智能体的局部观察实现集群飞行与避障. 除此之外, Jin等人[26]还做了更多改进, 他们利用其他智能体动作的隐式估计, 无需策略函数指导智能体学习自身策略, 在完全可观的离散环境中实现多智能体合作导航. 综上所述, 针对将多个性质相同的单智能体使用同一强化学习网络独立训练的思想, 现有文献在强化学习方法上仍未引入D3QN算法实现决策; 在应用场景上缺乏多智能体动态合作与竞争行为的关注. 因此, 本文以单智能体强化学习D3QN算法为基础, 在二维迷宫环境中, 采用同一D3QN训练同队合作多智能体; 使用不同D3QN分别训练相互竞争的异队多智能体, 从而同时关注竞争和合作行为, 最终得到MAD3QN框架, 探索夺旗博弈策略.

考虑所有智能体仅可观测到一定范围内的环境信息, 因此在路径规划的过程中存在很多隐藏信息(视野

范围外的地图信息、异队智能体和敌方旗帜的位置), 导致夺旗博弈信息不完美. 不完美信息博弈指参与者具体的策略已知, 而博弈的状态(部分)未知, 此时一些关于博弈状态的信息对参与者是隐藏的[27]. 不完美信息博弈有着广泛的实际应用背景与研究价值. 以军事领域为例, 在噪声、电子干扰、信号去阻断、数据链切换等复杂场景下, 敌我双方多智能体的交互过程, 即为一种不完美信息博弈. 以星际争霸II[28]、王者荣耀[29]等实时策略博弈为例, 由于视图受限, 参与者视野范围外的敌方智能体信息对于该参与者是隐藏的, 使得博弈信息不完美. 在本文中, 夺旗博弈的场景为二维迷宫地图, 距离智能体观测范围外的非己方信息对于该智能体是隐藏的. 因此, 本文夺旗博弈是智能体观测信息受限的情况下, 简化动作空间, 复杂化地图障碍的实时策略博弈.

在处理不完美信息时, 现有文献[28,29]往往通过输入局部图像, 并采用卷积神经网络(convolutional neural networks, CNN)提取图像特征. 然而, CNN仅从智能体每一步状态的角度分析, 无法定量判断下一时刻智能体的决策偏向于合作或竞争, 从而影响最终多智能体夺旗博弈策略的制定[30]. 因此, 本文在使用CNN的基础上, 引入广泛应用于语义分割[31]以及深度估计[32]的注意力机制模块, 提升CNN的泛化能力. 具体而言, 首先将智能体看作节点, 构建注意力节点网络, 再对节点周边的邻居关系进行空间上的卷积, 利用GAT[33]实时分配节点的权重, 基于网络中最相关的节点智能体进行决策. 在每步迭代中不断学习并更新各节点间的注意力权重值, 辅助MAD3QN算法制定夺旗博弈策略, 提升学习效率.

综上所述, 本文的贡献可归纳为如下三点.

(1) 基于MAD3QN与GAT算法, 提出了G-MAD3QN算法, 针对外部竞争、内部合作的多智能体端到端训练夺旗博弈策略, 实现二维迷宫环境中不完美信息条件下多智能体夺旗博弈.

(2) 将单智能体深度强化学习DQN算法拓展为多智能体深度强化学习MAD3QN算法, 有效提升二维迷宫环境多智能体夺旗成功率.

(3) 针对CNN无法量化多智能体博弈过程的问题, 引入GAT定量地确定不完美信息条件下智能体间合作与竞争的关系, 利用注意力权重辅助MAD3QN算法, 高效优化了多智能体夺旗策略.

2 背景知识

2.1 部分可观马尔可夫博弈

部分可观马尔可夫博弈是部分可观马尔可夫决策过程在多智能体博弈中的推广. 在t时刻, 部分可观马尔可夫博弈由N个智能体定义, 其中包含一组局部观测$O_1^t,\cdots,O_N^t$、一组动作$A_1^t,\cdots,A_N^t$和一组状态S和状态转移函数$Tr: S\times A_1^t\times A_2^t\times\cdots\times A_N^t\to S$. 对于每个智能体$i$, 通过局部观测$o_i^t\in O_i^t$, 遵循策略$\pi_i: O_i^t\times A_i^t\to[0,1]$, 得到智能体动作集中选取动作$a_i^t$的动作概率, 并且通过与环境的交互, 转移到下一个状态并获得奖励$r_i: S\times A_i^t\to R$以判断策略的好坏. 初始状态由初始状态分布$\rho: S\to[0,1]$来决定. 每个智能体i都试图最大化累积的折扣奖励$R_i=\sum_{t=0}^{T}\gamma^t r_i^t$, 其中$T$是期望时间范围, $\gamma\in[0,1]$是折扣参数.

2.2 不完美的信息博弈

在不完美信息博弈中, 博弈状态对参与者(部分)未知[34]. 当参与者同时行动时, 某一参与者无法获得其他参与者的部分/全部行动信息. 因此, 若把其他参与者的行动视为某一参与者面对的环境, 那么不完美信息指参与者不知道自身所处的决策环境. 不完美信息博弈模拟了具有隐藏信息的参与者之间的交互, 广泛应用于谈判、网络安全、拍卖等场景中.

2.3 深度强化学习D3QN算法

DQN算法由深度神经网络与Q函数结合而成, 根据t时刻智能体的观测值, 输出最大Q值. 使用损失函数的最小权值训练DQN算法. 具体而言, 首先构建DQN的损失函数以及用于更新目标Q值的贝尔曼方程:

$$\min_\theta\sum_{t=0}^{T}\left[\bar{Q}\left(s_t,a_t\mid\theta\right)-y_t\right]^2, \tag{1}$$

$$y_t=r_t+\gamma\max\bar{Q}_{a_{t+1}}\left(s_{t+1},a_{t+1}\mid\bar{\theta}\right), \tag{2}$$

其中, θ为权重, s_t, a_t, r_t分别表示智能体在t时刻的状态、动作与奖励, $\bar{Q}\left(s_t,a_t\mid\theta\right)$表示DQN算法在$t$时刻预测的最大$Q$值, y_t表示权重为$\bar{\theta}$时DQN的目标Q值.

通过最小化损失函数, 基于DQN算法的目标Q值近似其预测Q值, 如式(2)所示. 值得注意的是, DQN算法利用经验回放机制[35]解决样本之间的相关性问题, 从而避免网络仅根据输入的最新动作进行学习, 保证了算法的收敛性.

由于DQN一定条件下会高估动作值, 因此引入Double DQN思想, 将DQN算法计算最大Q值的过程改为由两个不同的DQN网络执行. 首先, 主DQN用于计算当前状态下最可能的动作. 然后, 将该动作传递给辅DQN, 计算当前训练步骤的目标Q值. 通过将动作选择与动作评估解耦, 从而显著减少动作值高估, 缩短训练时间并提升准确性.

为适应复杂环境中路径规划存在网络偏差的问题, 本文将Double DQN思想与Dueling DQN算法结合, 从而将Q值进一步分解为两个基本概念.

(1) 主、辅DQN的状态-价值函数, 分别为$V(s_t,a_t|\theta,\alpha)$和$V(s_t\mid\theta,\beta)$ (维数为$|\mathcal{A}|$), 以量化智能体在任何给定状态下的优劣. 其中α, β分别是Double DQN中主、辅DQN的参数.

(2) 优势函数$A(s_t,a_t\mid\theta,\alpha)$, 度量智能体选择特定动作的优劣. Dueling DQN通过将优势函数设置为单个动作的优势函数与所有动作的平均优势函数之差来估计状态值和每个动作的优势, 从而计算出最终的Q值:

$$Q(s_t,a_t|\theta,\alpha,\beta)=V(s_t|\theta,\beta)+\left(A(s_t,a_t|\theta,\alpha)-\frac{1}{|\mathcal{A}|}\sum_{a_{t+1}}A(s_t,a_{t+1}\mid\theta,\alpha)\right). \tag{3}$$

由于D3QN能够独立微调状态-价值函数以及优势函数的相关参数, 确保智能体无论是否转移状态, 都能够获得显著的奖励. 同时包含的经验回放机制减少了样本训练方差, 因此可以在保持D3QN算法收敛的同时实现更准确的状态估计, 提升训练效率.

2.4 图注意力网络

图注意力网络(graph attention network, GAT)是在图卷积网络[36]的基础上引入注意力机制, 为所有邻域节点分配一个可学习的注意力系数, 从而使每个对象都能对其所有邻域节点具有不同的感知力.

具体而言, 首先使用共享参数化矩阵$W\in R^F$, 其中F表示特征向量集的特征维度, 目标节点i的特征向量为h_i, 邻域节点$\{j\in N_i\}$的特征向量为h_j. 计算场景图中对象间的注意力权重过程为

$$e_{ij}=\varphi\left(Wh_i,Wh_j\right), \tag{4}$$

式中, e_{ij}为场景图中对象o_j对于对象o_i的贡献程度, φ为注意力系数计算网络. 接着, 使用softmax函数对每个对象的所有邻域节点进行标准化, 从而获得标准化注意力系数a_{ij}, 使得不同节点之间的注意力系数易于比较, 计算过程为

$$a_{ij}=\mathrm{softmax}_j(e_{ij})=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})}, \tag{5}$$

式中, N_i表示目标节点i的所有邻域节点(包括i本身). 计算a_{ij}对应的节点特征的线性组合, 作为每个对象节点的最终输出, 即

$$h_i^{'}=\sum_{j\in N_i}a_{ij}Wh_j, \tag{6}$$

式中, $h_i^{'}$包含场景图中其他所有对象的特征信息, 因此无论对象类别是否相同, 不同对象抽象向量的输出表示不同, 因此可以用于预测不同对象的位置.

3 多智能体夺旗博弈模型结构

3.1 问题描述

本文研究在二维离散迷宫环境下, 多智能体二对二夺旗博弈问题[37]. 如图1所示, 迷宫地图中心对称, 分为红、蓝两半区域. 本文假设红、蓝两队智能体的初始位置分别为地图的左下角和右上角. 观测范围是在迷宫环境约束下, 以智能体为圆心, 半径为5欧氏距离的范围. 动作集合指同队所有智能体动作的集合. 以红队为例, 包含智能体1的动作集{上1, 下1, 左1, 右1, 停1}与智能体2的动作集{上2, 下2, 左2, 右2, 停2}的5×5=25组动作组合. 主要目标是夺取敌方旗帜. 具体而言, 当智能体位于己方地图时, 该阶段目标为保护己方旗帜, 同时击杀入侵的敌方智能体(两方智能体位于同一坐标位置即实现击杀, 当智能体被击杀后, 在固定的初始位置复活); 当智能体位于敌方地图时, 该阶段目标为夺取敌方旗帜, 同时躲避敌方智能体的击杀. 智能体在每一步以相同移动速度从动作集合中选取动

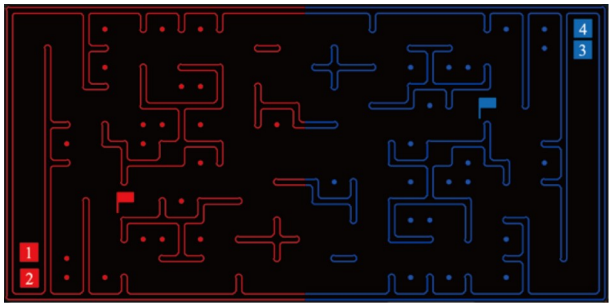

图 1　本文二维迷宫夺旗博弈环境

Figure 1　The two-dimensional maze environment of capture-the-flag game in this paper.

作，从当前时刻的位置坐标转移到下一时刻的位置坐标，当智能体移动到1200步时，夺旗博弈结束. 在二维迷宫地图环境限制下，智能体夺取旗帜后，仅当其成功返回己方领地时，己方分数才会发生变化(成功夺旗积1分)，最终得分高的队伍获胜. 如果最终得分为零，则比赛以平局结束. 本文从头开始使用Nvidia Quadro RTX 8000 GPU训练G-MAD3QN模型至收敛，在测试阶段，G-MAD3QN模型完成一轮测试(包含1200步动作)需花费大约35 s.

3.2　环境不完美信息分析

本文夺旗博弈框架中存在观测信息不完美. 智能体每一步可获得的完美信息为己方智能体与旗帜的位置；不完美信息包含迷宫地图中智能体视野范围外的非己方信息. 以红队智能体1为例，其观测范围受限，如图2中紫色圆圈范围所示. 当敌方旗帜在观测范围内时，智能体以其为目标进行探索；当敌方旗帜不在观测范围内时，智能体则进行随机探索. 智能体成功夺旗后，若观测范围内均为敌方领地，则进行随机探索，直至探索到己方领地边界线；若观测范围内有己方领地，其目标为在观测范围内己方领地边界线上的点集，如图2中绿色区域所示. 此时，智能体在躲避敌方追击的过程中不断优化目标，选择既能安全到达又距离较近的目标点，以促使智能体携旗快速返回得分.

3.3　观测信息处理

假定迷宫地图的全景为FOV，由于智能体仅可观测一定距离视场内的环境信息，本文将智能体的局部观测信息处理为$Z_t^i \in \mathbb{R}^{3\times W_{\text{in}}\times H_{\text{in}}}$，由3个通道组成，分别代表迷宫地图、智能体以及目标，如图3所示. 值得注意的是，$W_{\text{in}}=W_{\text{FOV}}+2$，$H_{\text{in}}=H_{\text{FOV}}+2$，其中$W_{\text{FOV}}$和$H_{\text{FOV}}$是矩形视场的宽度和高度.

3.4　网络体系结构

(1) 基于CNN的感知网络

相较于之前关于智能体夺旗的工作[30]，本文使用残差网络(ResNet)[38]对CNN模块进行升级，实现了一个具有3个堆叠ResNet模块的特征提取器，如图3所示. 与传统CNN的连接结构不同的是，每个模块间都有一个跳跃连接，将前后模块的特征连接在一起. 该类型的残差连接已广泛用于特征提取工作，有利于减少过

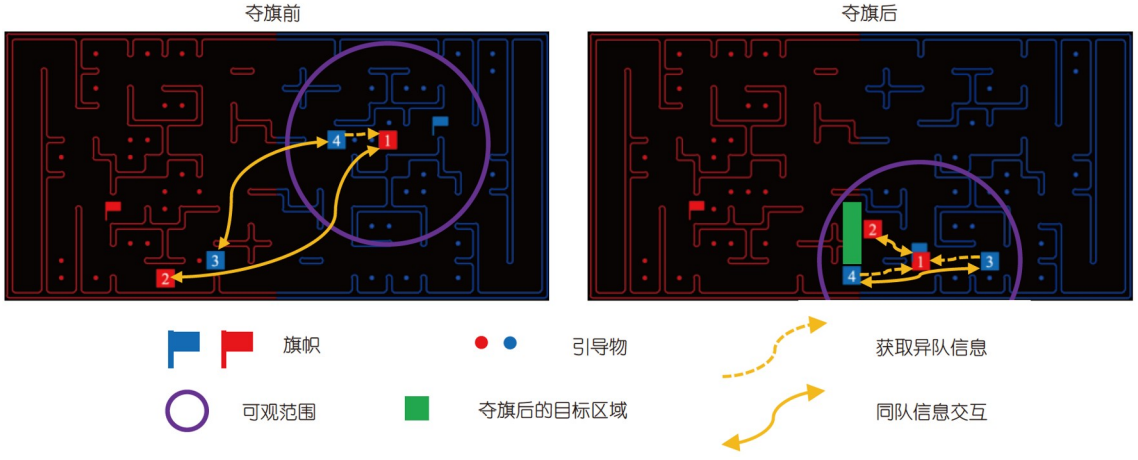

图 2　部分可观环境中多智能体博弈示意图

Figure 2　Schematic diagram of the multi-agent game in partially observable environments.

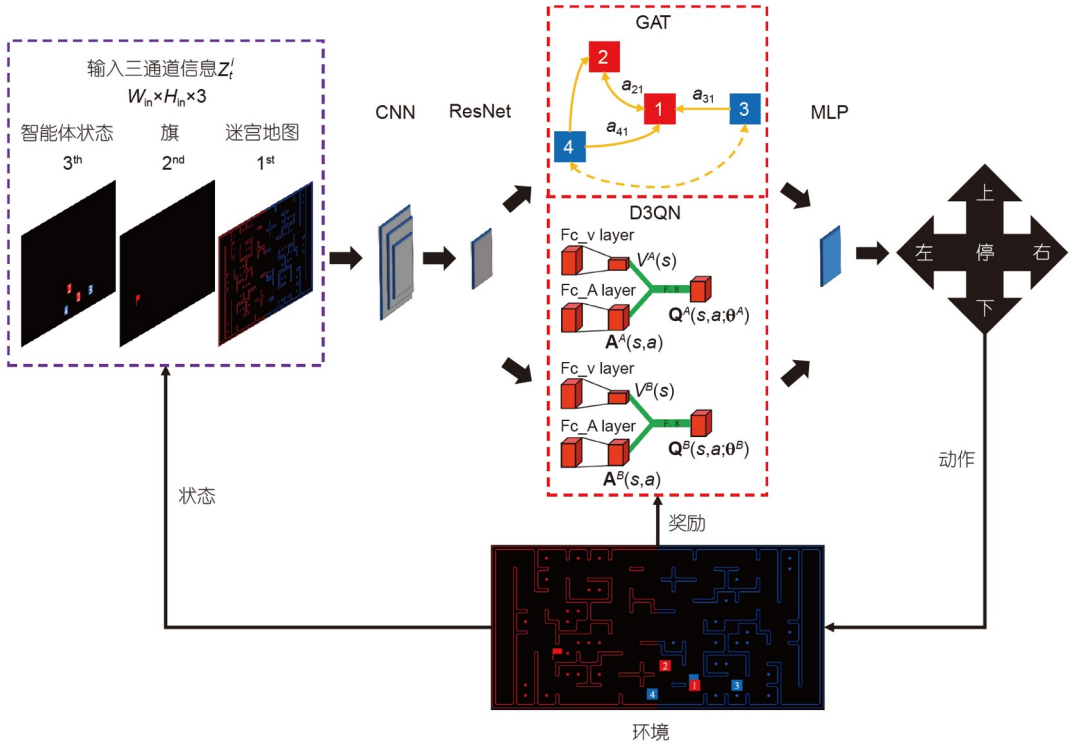

图 3 多智能体夺旗博弈模型结构图

Figure 3 Structure diagram of the multi-agent capture-the-flag game model.

拟合并提高性能[39].

(2) 基于GAT的交互网络

GAT交互网络中, 图注意力层训练一个参数化矩阵, 进行对象向量的特征变换后, 得到共享参数化矩阵$W\in R^F$和注意力系数计算网络$\varphi: R^{3F}\to R$. 如图3所示, 在本文模型中, 设在t时刻智能体i对周围环境的观测向量为o_t^i, 提取特征后传递到GAT模块, 获得智能体j对智能体i的标准化注意力系数a_{ij}, 从而增强了从智能体局部观测中提取的特征. 此外, 本文使用多层感知(multi-layer perceptron, MLP), 降维CNN模块中提取的特征维数, 有效减少了交互网络可共享的特征维度. 具体而言, 在每一步中, 智能体将提取的观测向量发送给队友一次, 同时接收一次队友的观测信息. 智能体根据接收到的队友信息, 融合自身对周围环境的观测信息, 根据相对重要性, 更改注意力权重.

(3) 基于D3QN的决策网络

基于D3QN的决策网络伪代码如表1所示, 分别训练红、蓝两队, 实现二维迷宫环境下的路径规划与避障, 从而得到夺旗策略.

3.5 奖励设计

深度强化学习奖励的设计将直接影响学习结果的质量. 因此, 在训练过程中定义合适的奖励与惩罚至关重要. 本文分为探索、进攻以及防守三个过程定义奖励函数, 具体由表2可得.

探索过程: 针对引导物、智能体移动设置奖励鼓励探索迷宫地图.

进攻过程: 相较于探索过程设置的奖励, 设置较大夺旗奖励以鼓励夺旗行为. 在夺旗后通过边界奖励与距离奖励, 引导智能体合作夺旗, 安全携旗返回己方领地.

防守过程: 同样地, 设置较大惩罚以鼓励护旗行为. 通过调整边界奖励与距离奖励, 鼓励智能体合作围堵.

4 仿真实验与结果分析

4.1 仿真实验环境

本文仿真环境以二维迷宫游戏平台[40]为基础, 如

表 1　D3QN算法伪代码

Table 1　D3QN algorithm pseudo code

D3QN算法

将评价网络的权重初始化为θ, θ'
设置记忆D以存储经验回放
设置重放缓冲区最大规模为N_r, 训练批次规模为N_b, 以及目标网络更替频率为N_{rf}
设置目标距离阀值d_{arrive}、碰撞距离阀值$d_{\text{collision}}$、回合数(episode)为M
for episode=1 to M do
　重置环境
　for t = 1 to 1200 do
　　获取t时刻智能体与目标的距离d_t、环境的部分观测o_t和所有智能体的注意力权重a_{ij}
　　根据a_{ij}选择下一步动作偏向进攻/防守
　　以ε概率选择偏向动作a_t
　　否则, $a_t = \text{argmax}_a Q(o_t, a \mid \theta)$
　　使用新的深度信息更新d_t
　　if $d_t \le d_{\text{arrive}}$ and $d_t > d_{\text{collision}}$ then
　　　if 处于进攻状态 then
　　　　$r(o_t, a_t) = c\left(d_{t-1} - d_t\right) + r_{\text{offensive}} + r_9 + r_{11}$
　　　else
　　　　$r(o_t, a_t) = c\left(d_{t-1} - d_t\right) + r_{\text{defensive}} + r_9 + r_{11}$
　　　end if
　　else if $d_t > d_{\text{arrive}}$ and $d_t > d_{\text{collision}}$ then
　　　$r(o_t, a_t) = r_9 + r_{11}$
　　else
　　　$r(o_t, a_t) = r_{10} + r_{11}$
　　end if
　　获取新的环境部分观测o_{t+1}
　　在记忆D中存储转移组(o_t, a_t, r_t, o_{t+1}), 当$|D| \geqslant N_r$时, 替换旧的元组
　　在记忆D中随机选择一批N_b的转移组(o_k, a_k, r_k, o_{k+1})
　　对每个转移组:
　　$a^{\max}(o_{k+1} \mid \theta) = \text{argmax}_{a'} Q(o_{k+1}, a' \mid \theta)$
　　$y_k = r_k + \gamma Q(o_{k+1}, a^{\max}(o_{k+1} \mid \theta) \mid \theta')$
　　以$(y_k - Q(o, a \mid \theta))^2$损失进行梯度下降更新
　　每N_{rf}步将目标参数θ'替换为θ
　end for
end for

图4所示, 该平台是加州伯克利大学人工智能课程提供的实践游戏环境. 本文针对该平台框架做出了以下修改.

(1) 更改游戏中引导物的性质. 从收集引导物得分变更为收集引导物获得奖励但不得分, 以辅助多智能体高效探索未知迷宫环境.

(2) 更改游戏目标, 引入旗帜概念. 从收集引导物得分变更为夺取敌方旗帜并返回己方领地得分.

(3) 考虑不完美信息观测. 即智能体仅可观测半径为5欧氏距离的范围.

通过以上修改, 使得二维迷宫游戏平台满足了不完美信息条件下多智能体夺旗博弈的条件. 即红蓝两队攻击智能体的目标是在不被捕获的情况下在敌方领地完成夺旗并返回己方领地, 而防守智能体试图在己方领地捕获攻击智能体以阻止其夺旗. 攻击、防守智能体通过同队合作与异队竞争, 利用不完美信息观测, 借助引导物探索迷宫地图, 最终实现夺旗目标.

4.2　实验结果与分析

(1) GAT模块分析

红、蓝两队均经过本文G-MAD3QN算法训练. 博弈开始后, 智能体试探性夺旗, 如图5(a)所示, 此时敌我双方在地图的不同位置实现两两攻防博弈. 以智能体1为例, 根据GAT模块学习到的注意力表可知, 其主要关注防守追击的智能体3; 由于同队智能体间的通信, 智能体1和2间存在相互注意, 但由于它们所处的周围环境不同, a_{12}与a_{21}并不相同.

如图5(b)所示, 从形式分析, 红队智能体合作同时

表 2　本文奖励函数的定义[a)]

Table 2　The definition of the reward function in this paper

鼓励进攻: $r_{\text{offensive}}$	鼓励防守: $r_{\text{defensive}}$	鼓励探索: r_{else}
r_1=−30, 被敌方智能体击杀; r_2=300, 夺取敌方旗帜; $r_3 = \begin{cases} -30, & \text{当智能体在己方时,} \\ \dfrac{100}{d_{\text{boundary}}}, & \text{当智能体在敌方时,} \end{cases}$ 当己方夺旗后, 边界奖励; $r_4 = \begin{cases} -150 / d_{\text{enemy}}, & d_{\text{enemy}} \leqslant 5, \\ 0, & d_{\text{enemy}} > 5, \end{cases}$ 当己方夺旗后, 距离奖励;	r_5=30, 击杀敌方智能体; r_6=−300, 己方旗帜被夺; $r_7 = \begin{cases} \dfrac{30}{d_{\text{boundary}}}, & \text{当智能体在己方时,} \\ -30, & \text{当智能体在敌方时,} \end{cases}$ 当己方旗帜被夺后, 边界奖励; $r_8 = \begin{cases} 150 / d_{\text{enemy}}, & d_{\text{enemy}} \leqslant 5, \\ 0, & d_{\text{enemy}} > 5, \end{cases}$ 当己方旗帜被夺后, 距离奖励;	r_9=10, 获取敌方引导物 r_{10}=−500, 智能体触碰迷宫障碍物 r_{11}=−5, 智能体每步移动

a) d_{enemy}表示智能体与敌方智能体的距离; d_{boundary}表示智能体与红、蓝队地图边界的距离

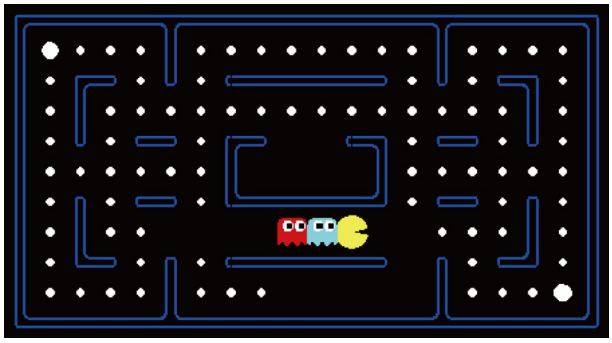

图 4 二维迷宫游戏平台
Figure 4 The platform of the two-dimensional maze game.

与智能体4竞争实现夺旗. 从注意力权重表分析, 由a_{13}=0可知, 智能体3不在智能体1的观测范围内; a_{34}=1则表明由于智能体3的观测范围内无智能体, 智能体3只关注同队智能体4的动向. 此外, 由于a_{21}较a_{12}显著增大且a_{41}远大于a_{42}, 因此智能体2更关注智能体1合作实现夺旗; 智能体4更关注距离旗更近的智能体1.

如图5(c)所示, 当红队成功夺旗后, 智能体1需尽快返回己方领地以得分. 从形式分析, 蓝队智能体合作防守围堵, 红队智能体合作护旗撤离. 从注意力权重表分析, 对于蓝队, 由a_{41}>a_{42}以及a_{43}从0.09提升至0.33, 表明相较于智能体2, 智能体4更关注携旗撤离的智能体1, 同时由于智能体1均位于蓝队智能体的观测范围内, 因此蓝队增强了同队合作; 对于红队, 由a_{24}从0.32增加到0.42以及a_{14}远大于a_{13}, 表明智能体2更加关注对携旗智能体1产生威胁的智能体4. 尽管智能体3在智能体1的观测范围内, 此时智能体3对智能体1返回己方领地已不构成威胁.

夺旗博弈过程中注意力值的变化如图6所示, 以红队智能体1为例, 由于GAT模块求解得到的是标准注意力值, 因此多智能体的注意力值在每一步均满足a_{12}+a_{13}+a_{14}=1. 夺旗开始时两队智能体从初始位置出发, 由于智能体可观距离的限制, 因此从图6中可以看出, 在150步左右智能体1的观测范围内出现敌方智能

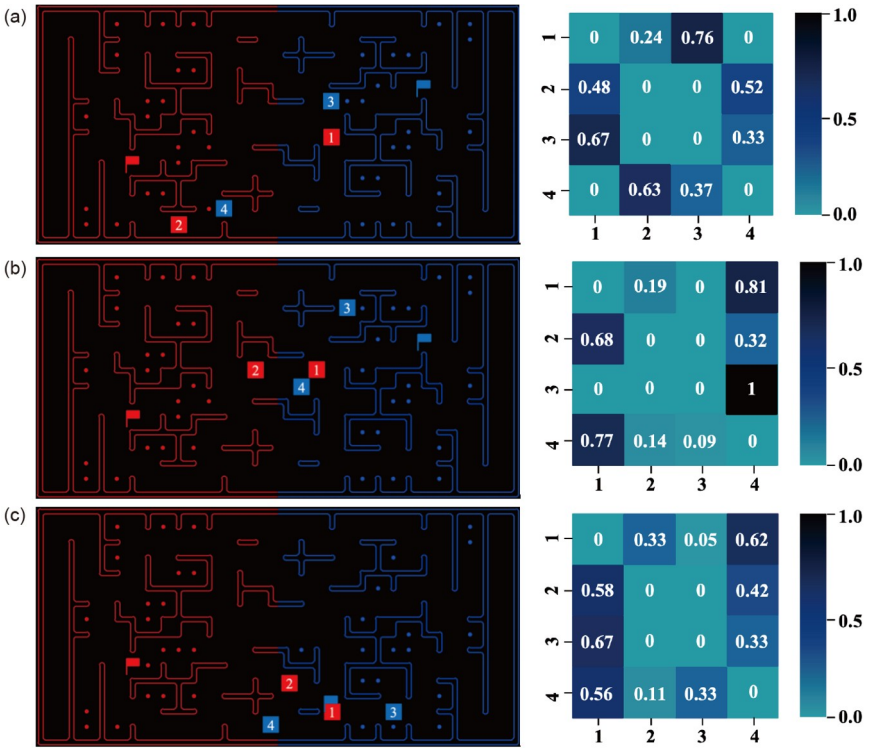

图 5 所有智能体在夺旗博弈不同时期的状态及相应的注意力权重分布. 红蓝两队均使用本文G-MAD3QN算法训练, 表中数字表示在特定时刻, 纵轴智能体对横轴智能体的注意力权重值. 智能体1和2为红队, 智能体3和4为蓝队. (a) 步数=427时; (b) 步数=831时; (c) 步数=916时
Figure 5 The states of all agents in different periods of capture-the-flag game and the distribution of the corresponding attention weight. The red and blue teams are trained using the proposed G-MAD3QN algorithm, and the numbers in the table represent the attention weight value of the ordinate agent to the abscissa agent at a specific moment. Agents 1 and 2 belong to the red team, and agents 3 and 4 belong to the blue team. (a) Number of steps =427; (b) number of steps = 831; (c) number of steps = 916.

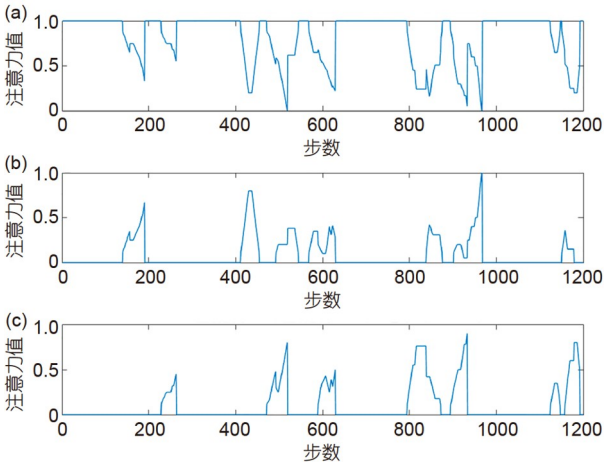

图 6　夺旗博弈过程中注意力值变化曲线图(以智能体1为例, 与图5瞬时状态相对应). (a) 智能体1对智能体2的注意力值曲线图; (b) 智能体1对智能体3的注意力值曲线图; (c) 智能体1对智能体4的注意力值曲线图

Figure 6　The change curve of attention values during the capture-the-flag game (this paper takes agent 1 as an example, corresponding to the instantaneous state in Figure 5). (a) The attention value curve of Agent 1 to Agent 2; (b) the attention value curve of Agent 1 to Agent 3; (c) the attention value curve of Agent 1 to Agent 4.

体. 因此在相遇前, a_{12}=1, a_{13}=a_{14}=0. 对于同队合作的智能体1与智能体2而言, 当a_{12}=1时, 智能体1仅与智能体2合作; 当a_{12}≠1时, 智能体1既与智能体2合作, 又与异队智能体竞争. 类似地, 对于异队竞争的智能体1与智能体3而言, 当a_{13}=1时, 智能体1仅与智能体3竞争; 当a_{13}≠1时, 智能体1既与智能体3竞争, 又与同队智能体合作.

因此, GAT模块不仅能够量化不完美信息下迷宫地图中智能体合作与竞争的态势, 还能够辅助智能体优化夺旗博弈策略.

(2) G-MAD3QN算法分析

为验证本文G-MAD3QN算法性能, 引入基线算法MADQN和MAD3QN, 分别训练红蓝两队进行二对二夺旗博弈, 并与G-MAD3QN算法进行比较, 所有算法均使用CNN感知网络. 表3以红队的角度衡量经过不同算法训练后, 在测试阶段夺旗得分的均值与标准差, 均值越大则表明红队实现夺旗的成功率越高, 反之亦然. 若红队成功夺旗得1分, 平局得0分, 被蓝队夺旗得−1分. 即每局的得分值集合为{1; 0; −1}. 由表3可得, 当使用本文G-MAD3QN算法训练红队时, 与其他基线算法相比得分均值为正, 特别是面对传统MADQN算法时夺旗得分均值超过0.9, 因此本文算法实现夺旗的成功率相对基线算法具有显著的优势. 从GAT交互网络角度分析, 经过本文G-MAD3QN算法训练的模型夺旗得分均值比缺少GAT模块的MAD3QN算法的高, 表明通过GAT交互网络对不完美信息的关注, 提升了夺旗得分; 从D3QN决策网络角度分析, 由于MAD3QN算法训练的模型夺旗得分均值比MADQN算法高, 从而证明了深度强化学习D3QN算法通过有效的环境探索, 有效提升了夺旗的成功率. 特别地, 由于引导物的分布不对称, 因而导致当红蓝两队训练算法交替后, 交替后的得分结果具有一定的波动.

为深入研究GAT交互网络的高效性, 本文增加了消融实验, 以分别验证在有无GAT模块训练后得到的夺旗博弈模型的得分效率. 具体而言, 在二对二夺旗训练中, 本文使用相同的D3QN决策网络与CNN感知网络, 引入GAT交互网络作为变量. 针对红蓝两队分别使用MAD3QN算法、G-MAD3QN算法训练, 测试50次红队取胜时的平均步数与标准差. 如表4所示, 红队使用G-MAD3QN算法实现夺旗的平均步数少于MAD3QN算法. 这意味着当训练框架增加GAT交互网络时, 测试时夺旗取胜所需步数相对未增加GAT交互网络时更少. GAT交互网络由于增加了对不同智能体间的注意力分析, 根据注意力权重值辅助选择目标智能体, 采取偏向合作/竞争的策略, 从而高效实现了夺旗.

表 3　二对二夺旗中红队得分的均值与标准差

Table 3　The mean and standard deviation of the red team's score in the two-to-two capture-the-flag game

红队	蓝队		
	MADQN[21]	MAD3QN[25]	G-MAD3QN
MADQN[21]	0.09±0.14	−0.45±0.07	**−0.94±0.02**
MAD3QN[25]	0.44±0.06	−0.05±0.13	−0.63±0.09
G-MAD3QN	**0.91±0.03**	0.57±0.10	0.02±0.16

表 4　二对二夺旗中红队取胜时的平均步数与标准差

Table 4　The average steps and standard deviation when the red team wins the two-to-two capture-the-flag game

红队	蓝队	
	MAD3QN[25]	G-MAD3QN
MAD3QN[25]	–	1037±15
G-MAD3QN	620±8	–

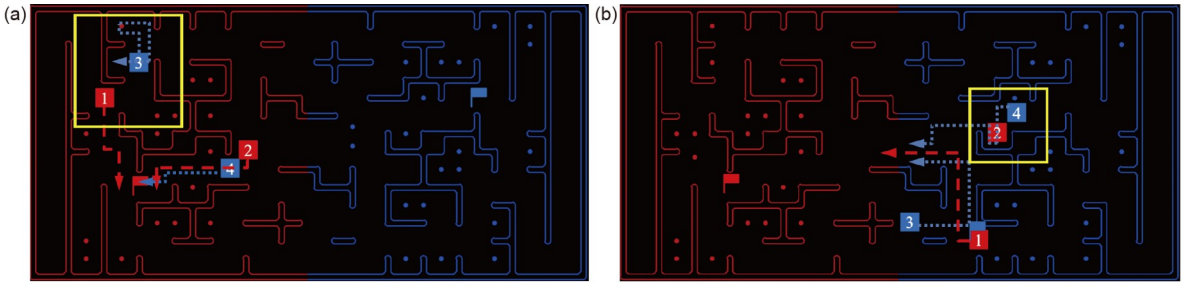

图 7　夺旗过程中特定场景下各智能体的部分运动轨迹. 红队使用本文G-MAD3QN算法训练, 蓝队使用基线算法训练; 红队的轨迹用虚线表示, 蓝队的轨迹用点线表示. (a) 合作围捕; (b) 合作撤退

Figure 7　Part of the trajectory of each agent in a specific scene during the process of capturing the flag. The red team agents are trained by G-MAD3QN algorithm, while the blue team agents are trained by the baseline algorithm; the red team agents' trajectories are represented by long dotted lines, while the blue team agents' trajectories are represented by short dotted lines. (a) Cooperative hunting; (b) cooperative withdrawal.

(3) 夺旗博弈策略分析

为分析G-MAD3QN算法相比基线算法在博弈策略上的提升, 本文使用G-MAD3QN算法训练红队, 使用MADQN基线算法训练蓝队, 在测试过程中, 分别选取红队合作围捕与夺旗后合作撤退两种博弈场景, 各智能体的运动轨迹如图7所示.

1) 红队合作围捕: 红队在防守过程中, 尽管智能体1的观测范围内只有敌方智能体3, 根据强化学习奖励设置, 智能体1将追击智能体3. 但是由于智能体2在追击敌方智能体4的过程中, 智能体4逐渐靠近红队旗帜, 因而智能体1逐渐增加对智能体2的注意力权重, 使得智能体1逐渐放弃追击更近的智能体3, 与智能体2合作围捕相对较远的智能体4. 最终, 红队成功将智能体4拦截, 该过程如图7(a)所示.

2) 红队合作撤退: 红队在夺旗成功后撤退的过程中, 尽管智能体2可躲避智能体4从而安全返回己方区域. 然而, 为保护携旗智能体1, 智能体2逐渐增加对智能体1的注意力权重, 从而忽略了智能体4的追击, 尽管最终被敌方捕获, 但间接为智能体1的撤退争取更多的时间. 由于基线算法训练的智能体4选择追击可观范围内的智能体2, 最终导致蓝队未能实现对智能体1的围捕. 最终, 智能体1顺利返回己方, 红队成功夺旗, 该过程如图7(b)所示.

综上所述, 使用本文G-MAD3QN算法训练的夺旗博弈框架进行夺旗的过程中, 智能体通过快速选择目标, 动态合作与竞争, 以提升夺旗的成功率.

5　讨论与结论

本文以复杂环境下群智的合作与竞争为切入点, 研究二维迷宫环境中多智能体在观测范围受限的不完美条件下实现夺旗博弈, 提出了基于MAD3QN算法与GAT的多智能体夺旗博弈G-MAD3QN算法. 智能体在探索迷宫地图的同时, 完成路径规划与避障, 制定并优化合作与竞争策略, 最终实现夺旗目标. 在实验部分, 本文以二维迷宫游戏平台为载体, 通过与多智能体深度强化学习基线算法MADQN和MAD3QN相比较, 验证了本文提出的G-MAD3QN算法实现夺旗的有效性与高效性. 未来, 为提升夺旗博弈的成功率与效率, 在感知网络方面, 我们将考虑夺旗过程中的时域信息, 引入时域卷积网络, 在提取图像局部特征的同时捕捉时序上的依赖关系; 在交互网络方面, 我们将考虑分层GAT网络, 利用智能体间与队伍间的多层注意力关系实现多队夺旗博弈, 并促进交互策略向不同智能体组成的新任务迁移; 同理, 在决策网络方面, 将考虑引入分层强化学习的思想, 为复杂夺旗博弈场景提供更高效的解决方案; 在夺旗博弈框架方面, 将考虑更复杂的夺旗博弈平台, 比如地图非对称, 攻防多队的智能体数量时变、速度异质, 增加旗帜数量等条件, 以丰富夺旗博弈的场景.

参考文献

1　Du W, Ding S F. Overview on multi-agent reinforcement learning (in Chinese). Comput Sci, 2019, 46: 1–8 [杜威, 丁世飞. 多智能体强化学习综

述. 计算机科学, 2019, 46: 1–8]
2 DeCanio S J, Fremstad A. Game theory and climate diplomacy. Ecol Econom, 2013, 85: 177–187
3 Huang K Q, Xing J L, Zhang J G, et al. Intelligent technologies of human-computer gaming (in Chinese). Sci Sin Inf, 2020, 50: 540–550 [黄凯奇, 兴军亮, 张俊格, 等. 人机对抗智能技术. 中国科学: 信息科学, 2020, 50: 540–550]
4 Liu W Q. Public data evolution games on complex networks and data quality control (in Chinese). Sci Sin Inf, 2016, 46: 1569–1590 [刘文奇. 复杂网络上的公共数据演化博弈与数据质量控制. 中国科学: 信息科学, 2016, 46: 1569–1590]
5 Luo Y Z, Li Z Y, Zhu H. Survey on spacecraft orbital pursuit-evasion differential games (in Chinese). Sci Sin Tech, 2020, 50: 1533–1545 [罗亚中, 李振瑜, 祝海. 航天器轨道追逃微分对策研究综述. 中国科学: 技术科学, 2020, 50: 1533–1545]
6 Selvakumar J, Bakolas E. Feedback strategies for a reach-avoid game with a single evader and multiple pursuers. IEEE Trans Cybern, 2021, 51: 696–707
7 Huang H, Ding J, Zhang W, et al. Automation-assisted capture-the-flag: A differential game approach. IEEE Trans Contr Syst Technol, 2014, 23: 1014–1028
8 Garcia E, Casbeer D W, Pachter M. The capture-the-flag differential game. In: Proceedings of the 2018 IEEE Conference on Decision and Control (CDC). Miami: IEEE, 2018. 4167–4172
9 Pachter M, Casbeer D W, Garcia E. Capture-the-flag: A differential game. In: Proceedings of the 2020 IEEE Conference on Control Technology and Applications (CCTA). Montreal: IEEE, 2020. 606–610
10 Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning. arXiv: 1312.5602
11 Tai L, Liu M. Towards cognitive exploration through deep reinforcement learning for mobile robots. arXiv: 1610.01733
12 Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 2016 AAAI Conference on Artificial Intelligence (AAAI). Phoenix, Arizona, USA: AAAI, 2016
13 Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay. In: Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, 2016. 322–355
14 Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning. In: Proceedings of the 2016 International Conference on Machine Learning (ICML). New York: ACM, 2016. 1995–2003
15 Han S H, Choi H J, Benz P, et al. Sensor-based mobile robot navigation via deep reinforcement learning. In: Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp). Shanghai: IEEE, 2018. 147–154
16 Ruan X, Ren D, Zhu X, et al. Mobile robot navigation based on deep reinforcement learning. In: Proceedings of the 2019 Chinese Control and Decision Conference (CCDC). Nanchang: IEEE, 2019. 6174–6178
17 Chen G, Pan L, Xu P, et al. Robot navigation with map-based deep reinforcement learning. In: Proceedings of the 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC). Nanjing: IEEE, 2020. 1–6
18 Li J, Ran M, Wang H, et al. A behavior-based mobile robot navigation method with deep reinforcement learning. Unman Syst, 2021, 9: 201–209
19 Wang S, Jin X, Mao S, et al. Model-free event-triggered optimal consensus control of multiple Euler-Lagrange systems via reinforcement learning. IEEE Trans Netw Sci Eng, 2021, 8: 246–258
20 Movahedi Z, Bastanfard A. Toward competitive multi-agents in Polo game based on reinforcement learning. Multimed Tools Appl, 2021, 80: 26773–26793
21 Tampuu A, Matiisen T, Kodelja D, et al. Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE, 2017, 12: e0172395
22 Leibo J Z, Zambaldi V, Lanctot M, et al. Multi-agent reinforcement learning in sequential social dilemmas. arXiv: 1702.03037
23 Bansal T, Pachocki J, Sidor S, et al. Emergent complexity via multi-agent competition. arXiv: 1710.03748
24 Jin Y, Wei S, Yuan J, et al. Hierarchical and stable multiagent reinforcement learning for cooperative navigation control. IEEE Trans Neural Netw Learn Syst, 2021, doi: 10.1109/TNNLS.2021.3089834
25 Yan C, Wang C, Xiang X, et al. Deep reinforcement learning of collision-free flocking policies for multiple fixed-wing UAVs using local situation maps. IEEE Trans Industr Inform, 2021, 18: 1260–1270
26 Jin Y, Wei S, Yuan J, et al. Stabilizing multi-agent deep reinforcement learning by implicitly estimating other agents' behaviors. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2020. 3547–3551
27 Brown N, Sandholm T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science, 2018, 359: 418–424

28 Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575: 350–354
29 Ye D, Chen G, Zhang W, et al. Towards playing full MOBA games with deep reinforcement learning. In: Proceedings of the 2020 Advances in Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2020. 33
30 Herrera S R. Applying deep reinforcement learning to Berkeley's capture the flag game. Uniandes, 2019
31 Sun Q Y, Zhao C Q, Tang Y, et al. A survey on unsupervised domain adaptation in computer vision tasks (in Chinese). Sci Sin Tech, 2022, 52: 26–54 [孙琦钰, 赵超强, 唐漾, 等. 基于无监督域自适应的计算机视觉任务研究进展. 中国科学: 技术科学, 2022, 52: 26–54]
32 Zhao C Q, Sun Q Y, Zhang C Z, et al. Monocular depth estimation based on deep learning: An overview. Sci China Tech Sci, 2020, 63: 1612–1627
33 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. In: Proceedings of the 2017 International Conference on Learning Representations (ICLR). Palais des Congrès Neptune, Toulon, 2017. 1–12
34 Osborne M J, Rubinstein A. A Course in Game Theory. Cambridge, MA: MIT Press, 1994
35 Liu Q, Zhai J W, Zhang Z C, et al. A survey on deep reinforcement learning (in Chinese). J Comput, 2018, 41: 1–27 [刘全, 翟建伟, 章宗长, 等. 深度强化学习综述. 计算机学报, 2018, 41: 1–27]
36 Scarselli F, Gori M, Ah Chung Tsoi M, et al. The graph neural network model. IEEE Trans Neural Netw, 2008, 20: 61–80
37 Lipovetzky N, Sardina S. Pacman capture the flag in AI courses. IEEE Trans Games, 2018, 11: 296–299
38 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778
39 Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the 2017 AAAI Conference on Artificial Intelligence (AAAI). San Francisco: AAAI, 2017
40 DeNero J, Klein D. Teaching introductory artificial intelligence with Pac-Man. In: Proceedings of the First AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI). Atlanta: AAAI, 2010

Swarm intelligence capture-the-flag game with imperfect information based on deep reinforcement learning

WANG JianRui, HUANG JiaHao & TANG Yang

Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

One of the major research areas has been the problem of swarm intelligence games in complex environments. This study offers G-MAD3QN, a multi-agent deep reinforcement learning system based on Multi-agent Dueling Double Deep Q-Network (MAD3QN) and Graph Attention Network (GAT), to handle the challenges of multi-agent capture-the-flag games under imperfect information settings. The algorithm realizes the path planning in the labyrinth map while also modeling the cooperation and competition relationships of multi-agents under imperfect information conditions at the same time so as to determine the strategy of the capture the flag game. In the experiment, we consider the imperfect observation information of the agents based on the two-dimensional maze environment. Moreover, in the two-on-two capture-the-flag game, we compared the G-MAD3QN algorithm to baseline multi-agent deep reinforcement learning algorithms, such as Multi-agent Deep Q-Network (MADQN) and MAD3QN, to verify the proposed algorithm's effectiveness.

capture-the-flag game, imperfect information, deep reinforcement learning, graph attention network

doi: 10.1360/SST-2021-0382