多智能体深度强化学习研究进展

丁世飞[1),2)] 杜 威[1)] 张 健[1),2)] 郭丽丽[1),2)] 丁 玲[3)]

[1)](中国矿业大学计算机科学与技术学院 江苏 徐州 221116)

[2)](矿山数字化教育部工程研究中心(中国矿业大学) 江苏 徐州 221116)

[3)](天津大学智能与计算学部 天津 300350)

摘 要 深度强化学习(Deep Reinforcement Learning, DRL)在近年受到广泛的关注,并在各种领域取得显著的成功.由于现实环境通常包括多个与环境交互的智能体,多智能体深度强化学习(Multi-Agent Deep Reinforcement Learning, MADRL)获得蓬勃的发展,在各种复杂的序列决策任务上取得优异的表现.本文对多智能体深度强化学习的工作进展进行综述,主要内容分为三个部分.首先,我们回顾了几种常见的多智能体强化学习问题表示及其对应的合作、竞争和混合任务.其次,我们对目前的MADRL方法进行了全新的多维度的分类,并对不同类别的方法展开进一步介绍.其中,我们重点综述值函数分解方法,基于通信的MADRL方法以及基于图神经网络的MADRL方法.最后,我们研究了MADRL方法在现实场景中的主要应用.希望本文能够为即将进入这一快速发展领域的新研究人员和希望获得全方位了解并根据最新进展确定新方向的现有领域专家提供帮助.

关键词 多智能体深度强化学习;基于值函数;基于策略;通信学习;图神经网络

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2024.01547

Research Progress of Multi-Agent Deep Reinforcement Learning

DING Shi-Fei[1),2)] DU Wei[1)] ZHANG Jian[1),2)] GUO Li-Li[1),2)] DING Ling[3)]

[1)](*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116)

[2)](*Mine Digitization Engineering Research Center of the Ministry of Education (China University of Mining and Technology), Xuzhou, Jiangsu* 221116)

[3)](*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

Abstract Reinforcement learning is a traditional machine learning method to solve complex decision-making problems. With the advent of the era of artificial intelligence, deep learning has achieved remarkable success thanks to the vast amount of data and the increase in computing power brought by hardware development. Deep reinforcement learning (DRL) has been widely paid attention in recent years and achieved remarkable success in various fields. Because the real environment usually includes multiple agents interacting with the environment, the multi-agent deep reinforcement learning (MADRL) has gained vigorous development and achieved excellent performance in a variety of complex sequential decision tasks. This paper summarizes the research progress of multi-agent deep reinforcement learning, which is divided into three parts. First, we review several common multi-agent reinforcement learning problem representations such as Markov games and partially observable Markov games and their corresponding cooperative,

收稿日期:2023-07-09;在线发布日期:2024-04-18. 本课题得到国家自然科学基金项目(No. 62276265, No. 61976216, No. 62206297)资助. **丁世飞**,博士,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为人工智能、模式识别、机器学习、数据挖掘. E-mail: dingsf@cumt.edu.cn. **杜 威**(通信作者),博士研究生,中国计算机学会(CCF)学生会员,主要研究领域为深度学习、强化学习. E-mail: 1394471165@qq.com. **张 健**(通信作者),博士,讲师,中国计算机学会(CCF)会员,主要研究领域为机器学习、模式识别. Email: Zhangjian10231209@cumt.edu.cn. **郭丽丽**,博士,讲师,中国计算机学会(CCF)会员,主要研究领域为深度学习、多模态情感计算. **丁玲**,博士研究生,中国计算机学会(CCF)学生会员,主要研究领域为图机器学习、深度聚类.

competitive, and mixed cooperative-competitive tasks. Second, we make a new multi-dimensional classification of the current MADRL method and further introduce the methods of different categories. Concretely, we divide MADRL into value-based function methods and policy-based methods according to different ways of solving optimal policies. Besides, we divide MADRL into cooperative tasks and general tasks (cooperative, competitive, or mixed task) according to applicable task types. In addition, we introduce a new dimension, that is, whether a communication mechanism is established between agents, dividing the MADRL into communication and non-communication methods. Based on the above three dimensions, the popular MADRL methods are divided into eight categories. Among them, we focus on the value function decomposition method, communication-based MADRL method, and graph neural network based MADRL method. Value function decomposition methods can be divided into simple factorization, IGM principle based, and others. Communication structures are divided into fully connected, star, tree, neighbor, and layered types. In addition, we study the main applications of MADRL methods in real-world scenarios such as autonomous driving, traffic signal control, and recommendation systems. The classification in this paper is based on several common types of MADRL problem representation and model-free MADRL methods, so there are many unfocused but promising directions, which we briefly analyze in section 5, including extensive game problems, model-based MADRL methods, and safe and robust MADRL. Finally, we give a summary of this paper. With the rapid development of deep learning methods, the MARL field is undergoing rapid change, and many previously unsolvable problems are gradually becoming easier to handle with MADRL methods. MADRL is a developing field, that attracts more interest from scholars, but also faces many challenges such as non-stationarity, dimensional curse, and credit assignment. Overall, DRL can improve the intelligence and efficiency of systems in various fields by learning optimal decision strategies, bringing tremendous impact and change to human society. In this paper, we provide a broad overview of the latest work in the emerging field of multi-agent deep reinforcement learning, starting from extended game theory, model-based MADRL, and secure and robust MADRL. We expect this paper will be helpful to new researchers entering this rapidly developing field and to existing field experts who want to gain a comprehensive understanding and determine new directions based on the latest advances.

Keywords multi-agent deep reinforcement learning; value-based; policy-based; communication learning; graph neural network.

1 引 言

随着人工智能和大数据时代的到来,受益于海量的数据以及硬件发展带来的算力提升,深度学习(Deep Learning, DL)[1]在诸多领域取得了引人注目的成功,如计算机视觉[2-4]、语音识别[5-7]、自然语言处理[8-10]等.深度学习模型的神经网络的训练,通常依赖大量的有标签数据为网络中巨量的参数提供稠密的监督信号.这类学习范式是机器学习中一类常用的学习范式,通常被称为监督学习.监督学习假设深度神经网络做出的分类或者回归的决策,与使用的数据是相互独立的,即深度神经网络的输出不会对数据集中样本点的分布造成影响.我们假设固定的训练集构成了一个场景,无状态的场景表示无论哪个时刻从该场景中采集样本数据,所得到的数据分布都是相同的.因此监督学习的训练集可以看作是一个无状态的场景.如果训练集是一个有状态的场景,即模型的决策可以改变环境的状态,那么监督学习在上一个数据集上训练完成的模型,在下一轮训练时面对改变后的数据集,可能无法做出正确的预测.现实世界中有许多任务都属于有状态的场

景,如推荐算法会改变用户未来的行为,进而影响未来的数据分布.扫地机器人一个原本安全的决策,由于环境的随机性,可能会撞倒花瓶,这也会改变未来的数据分布.在有状态的场景下进行决策通常被建模为序列决策问题,该问题能够由马尔可夫决策过程(Markov Decision Process, MDP)[11]进行建模,而且能够借助强化学习(Reinforcement Learning, RL)方法来求解最优策略.

强化学习是一种经典的机器学习方法,其中一个智能体(agent)或多个智能体与环境(environment)不断交互来实现其长期累积回报最大化.智能体不被告知应该选择和执行什么动作(action),需要通过不断的试错和学习来获取它们的最优行为.其中,智能体通过获得对理想动作的奖励(reward)和对不理想动作的负奖励来学习.同时由于奖励可能会延迟,智能体要在探索有可能产生更高的回报和利用当前奖励最高的状态两者之间做出必要的权衡.强化学习的最终目标是使智能体在与环境反复地互动中学得一个最优策略(policy),该最优策略可以最大化智能体整个互动过程中的期望累积奖励.然而,传统 RL 方法存在维度诅咒的缺点:随着状态空间和动作空间维度的增加,算法的效率会降低.近年来,随着深度学习的显著成功,强化学习和深度学习结合形成了诸多深度强化学习(Deep Reinforcement Learning, DRL)[12]方法,使人工智能在众多现实的复杂序列决策任务取得超越人类水平的性能表现.尽管如此,前期的工作主要关注单个智能体学习解决序列决策任务的情景,即单智能体强化学习.然而,在大量现实任务中,环境往往包括多个与环境互动并同时学习的智能体,这类问题通常被称作多智能体序列决策问题,如自动驾驶[13]和推荐系统[14-15]等.为解决这些任务,多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)应运而生.

多智能体强化学习是机器学习领域的重要理论分支,其交叉融合强化学习、博弈论、控制论、社会心理学等学科的方法,适用于解决各种复杂多智能体序列决策问题.MARL 结合深度学习形成的多智能体深度强化学习方法(Multi-Agent Deep Reinforcement Learning, MADRL)已被广泛运用于解决各类现实问题如交通信号控制、自动驾驶、智慧医疗、推荐系统等,具有广阔的应用前景.MADRL[16-18]不同于单智能体 DRL 最重要的地方在于,环境的动态是由环境中的所有因素以及所有智能体的共同行为决定的.每个智能体都面临着环境的不平稳性问题:智能体的最优策略随着其它智能体策略的变化而改变.在多智能体系统中,维度诅咒问题也会变得更严重,因为环境中每个增加的智能体都会使状态-动作空间的维度增加.此外还有局部可观测性限制,信用分配等问题[19-21].同时多智能体深度强化学习涉及合作、竞争或混合合作竞争等不同的任务场景.在面对不同的任务场景时,通常有不同的的解决方案.

由于 MADRL 方法种类繁多且不断涌现,之前的综述工作[22-24]或基于学习框架或基于问题与挑战对现有的 MADRL 进行分类与综述,或重点关注其中的某一个分支如基于通信的方法或基于值函数分解的方法.总体而言,我们认为目前缺乏一个比较系统的多维度的分类方式,将主流的 MADRL 工作方法尽可能的涵盖和归类.基于这个原因,在本文中,我们从多个维度对 MADRL 方法进行系统的分类和综述.在第 2 节中,我们回顾了单智能体强化学习的相关背景知识.在第 3 节中,我们综述了多智能体强化学习方法,展示提出的多个分类维度,并分析了每个类别的最近工作.同时,重点总结了值函数分解方法,基于通信的方法,以及基于图神经网络的方法.在第 4 节中,我们综述了 MADRL 在各个领域的应用.最后,我们讨论 MADRL 一些未重点介绍但很有前景的方向.

2　单智能体强化学习

在第 2 节我们回顾单智能体强化学习的基础知识,以方便后续在第 3 节引出多智能体强化学习的问题表示、前沿工作的系统分类.

2.1　马尔可夫决策过程

大多数单智能体强化学习问题可以被定义为马尔可夫决策过程(Markov Decision Process, MDP):一种定义了智能体与其环境交互的序列决策模型.形式上,它可以定义为一个多元元组$<S,A,P,R,\gamma>$,其中 S 是状态的集合,A 是动作的集合,P 表示转移概率函数,R 表示奖励函数,$\gamma\in[0,1]$是未来奖励的折扣因子.智能体以离散时间步与环境相互作用,在每个时间步 t,智能体处于状态 $s_t\in S$ 选择动作 $a_t\in A$.在每个时间步 $t+1$,智能体获得奖励值 $r_{t+1}\in R$,同时环境转移进入一个新状态 s_{t+1}.具体来说,状态转移概率函数通常被定义为 $P(s',r\mid s,a)=Pr\{S_t=s',R_t=r\mid S_{t-1}=s,A_{t-1}=a\}$.

MDP 进一步假设智能体对状态是完全可观测的,并且环境是稳定的:转移概率函数和奖励随时间

保持不变[20]. 智能体不具备完全可观测状态的设置称为部分可观测马尔可夫决策过程(Partially Observable Markov Decision Process, POMDP). 策略 π 是一种从环境状态到选择每个动作的概率的映射,它可以是确定性的,也可以是随机的. 智能体的目标是学习一种使其性能最大化的策略,通常定义为预期收益,计算为在轨迹内 $\tau=(s_0,a_0,s_1,a_1,\cdots,s_t,a_t)$ 预期奖励的折扣总和:

$$\mathbb{E}E_\tau\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{1}$$

折扣因子 $\gamma\in[0,1]$ 描述了奖励如何被赋值. 一个接近 0 的 γ 表示智能体更关注眼前的奖励,而一个接近 1 的 γ 表示智能体更关注未来的奖励. 使上述函数最大化的策略称为最优策略,记为 π. 大多数 MDP 求解方法可以分为三类:基于值函数的、基于策略的和基于模型的方法,接下来对前两类展开介绍.

2.2 基于值函数的方法

基于值函数的方法学习价值函数,并从最优价值函数推导出最优策略. 有两种价值函数,状态-价值函数和动作-价值函数. 状态-价值函数描述智能体处于状态 s 然后遵循策略 π 获得的期望收益[20],表示为

$$V_\pi(s)=\mathbb{E}_{s_0=s,\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{2}$$

动作-价值函数(有时称为 Q 函数)描述智能体在状态 s 中执行动作 a 获得的期望收益,表示为

$$Q_\pi(s,a)=\mathbb{E}_{s_0=s,a_0=a,\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{3}$$

最优策略 π^* 最大化状态-价值函数,使得对于所有的状态 $s\in S$ 和所有的策略 π,$V_{\pi^*}(s)\geqslant V_\pi(s)$.

如果我们有最优状态-价值函数,那么最优策略可以通过选择对状态 s 给出最大动作的动作来获取. 即最优策略由 $\pi^*=\max_\pi V_\pi(s)=\max_\pi Q_\pi(s,a)$.

2.3 基于策略的方法

与基于值的方法相比,基于策略的方法直接搜索最优策略,输出表示为动作的概率分布. 最优策略是通过梯度上升对目标策略的参数 θ 进行优化得到的. 策略网络的权重迭代更新,使得产生更高收益的状态-动作对更有可能被选择. 目标策略是所有完整轨迹的预期收益,定义如下:

$$J(\theta)=\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \tag{4}$$

基于策略的方法在连续和随机的环境中表现更好,学习每个动作的特定概率以及学习适当的探索水平. 然而基于策略的方法通常采样效率不是很高,因为策略梯度的新估计是独立于过去的估计来学习的. 同时,由于奖励稀疏以及只尝试了有限的状态和动作集,基于策略的方法的梯度估计量的方差较大. 演员评论家方法(Actor-Critic, AC)结合基于策略和基于值函数的方法来解决这些限制,同时演员评论家方法在保持学习过程中的稳定性的同时保留了理想的收敛特性. 演员评论家方法由学习策略的演员家网络和学习评估状态-行为对的价值函数的评论家网络组成. 评论家网络对状态-价值 $V(s;w)$ 或动作-价值函数 $Q(a\mid s;w)$ 的参数 w 进行近似和更新,演员网络根据评论家网络指导的方向更新策略参数 $\pi_\theta(a\mid s)$.

2.4 单智能体深度强化学习

单智能体深度强化学习的开创性工作是 Mnih 等[25]在 2015 年提出的深度 Q 网络(Deep Q-Network, DQN),用于解决离散动作空间下的强化学习问题. DQN 的核心思想是使用深度神经网络来近似 Q 值函数,从而实现对复杂环境的学习和决策. 在 DQN 中,使用一个深度神经网络来估计 Q 值函数,但存在一个问题,就是该网络可能会高估某些动作的价值,导致训练不稳定性和收敛困难. 双重深度 Q 网络(Double Deep Q-Network, DDQN)[26]通过引入两个 Q 网络来解决这个问题. 具体地,DDQN 使用一个 Q 网络来选择动作,另一个 Q 网络来评估所选择的动作的价值,从而减轻了高估动作价值的问题. 通过这种方式,DDQN 能够更稳定地学习到 Q 值函数,并且在许多基准任务上取得了比 DQN 更好的性能. 此外,对决深度 Q 网络(Dueling Deep Q-Network, Dueling Q)[27]模型把 Q-网络分成两部分,一个部分是状态-价值函数部分,这部分仅与状态有关,而与具体的动作无关. 另一部分是与状态和动作同时相关的优势函数. 状态-价值函数部分的输出和优势函数部分的输出的线性组合构成最终对决网络模型的输出. 智能体通过这种网络结构,可以在策略学习阶段更高效地选择动作. 上述 DQN 变体都能够在某个方向改进 DQN 性能,Hessel 等[28]提出 Rainbow 算法,整合了各种 DQN 变体方法并证明了这些方法很大程度上是互补的.

在处理复杂连续动作空间问题时,相比 DQN 等基于值函数的强化学习方法,基于策略的强化学习方法通常表现更佳. 确定性策略梯度(Deterministic Policy Gradient, DPG)[29]方法是一种基于策略的

强化学习方法,其主要思想是通过直接学习策略函数,而不是学习值函数,来实现策略的优化. DPG 方法直接学习一个确定性策略,即给定状态,直接输出对应的最优动作,而不是学习一个概率分布. DPG 方法使用策略梯度方法来更新策略,通过对策略函数的参数求梯度并沿着梯度方向更新参数,以使得长期累积回报最大化. 基于 DPG 方法,深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)[30] 方法被提出. DDPG 是一种基于演员评论家框架的深度强化学习方法,它结合了 DQN 的优点,并在连续动作空间任务中实现了稳定的收敛性. 与 DQN 相比,DDPG 通常能够以更少的时间步学习到最优策略. 现实世界的诸多场景涉及多个与环境交互的智能体,使用单智能体 DRL 来处理这些场景无法取得令人满意的效果. 因此研究者们开始关注多智能体强化学习的研究. 本文重点关注 MADRL 的最新研究工作,因此在本章节只简要回顾几种经典的单智能体深度强化学习方法,有许多学者详细综述了单智能体 DRL 的工作进展,感兴趣的研究人员可以参考 Wang 等[12] 的工作.

3　多智能体强化学习

3.1　多智能体问题表示

在多智能体强化学习中一组智能体在环境中相互作用以学习如何实现其目标. 虽然 MDP 已被证明有助于在单智能体序列决策问题中建模最优决策,但多智能体环境需要不同的问题表示. 在多智能体环境中,状态动态和期望收益在所有智能体共同行动时发生变化,违反了 MDP 的核心平稳性假设. 在多智能体设置中,问题表示取决于多智能体任务的类型,可以是合作的、竞争的或混合合作竞争的,如表 1 所示. 问题表示还取决于智能体对环境状态是完全可观测的还是局部可观测的. 根据这两个维度,图 1 展示了多智能体强化学习中主要问题表示. 当智能体对状态具有完全可观测性时, 问题通常用马尔可夫博弈(Markov games, MG)来表示. 当智能体对状态是局部可观测时,问题一般用局部可观测马尔可夫博弈(Partially Observable Markov Games, POMG)来表示. 一种特殊的类型是合作任务类型下的局部可观测马尔可夫博弈,在这种博弈中,智能体合作以获得最大的共同奖励,这种问题可表示为局部可观测马尔可夫决策过程(Partially Observable Markov Decision Process, POMDP). 图 1 展示了多智能体强化学习中主要问题表示的可视化描述(其中马尔可夫决策过程是单智能体强化学习的问题表示). 为简单起见,所有图都显示了两个智能体之间的交互,但可以扩展到更多的智能体.

在 POMDP 问题表示中,更特殊的情境是智能体协作但分散执行操作,则由分散的局部可观测马尔可夫决策过程(decentralised partially observable MDP, Dec-POMDP)表示,这也是 MARL 中最常用的问题表示形式,其可以用元组$<I,S,A,P,R,O>$表示. 其中 I 表示索引从 1 到 n 的有限智能体集合,S 表示有限状态集合,O 是联合局部观测集合,其中 $o_i \in O$ 是智能体 i 的局部观测. A 表示联合动作的有限集. 在每个时间步,智能体 i 根据局部观测 o_i 执行动作 a_i. $a=(a_1,\cdots,a_n)\in A$ 表示联合动作. 环境的状态根据状态转移函数 $P:S\times A\times S\to[0,1)$ 改变. $R:S\times A\to R$ 表示奖励函数. 智能体 i 的最终目标是最大化它的累积折扣奖励值,其中 $\gamma\in[0,1]$是折扣因子. 在 Dec-POMDP,合作智能体要学得一个最优策略 $\pi(\tau,a)$ 来最大化全局价值 $Q_{tot}^{\pi}(\tau,a)=E_{\{s,a\}}\left[\sum_{t=0}^{\infty}\gamma^t R(s,a)\right]$,其中 τ 表示联合观测历史.

表 1　任务类型划分

类型	奖励设置	任务描述
合作任务	全局奖励或局部奖励	一个智能体团队可以得到一个相等的共享全局奖励,它不考虑每个智能体的贡献. 智能体还可以获得局部奖励,额外设计使奖励取决于队友的集体表现,或与邻居共享奖励以鼓励合作.
竞争任务	冲突的奖励	当智能体需要与其他智能体竞争有限的资源或同一对象时,为它们分配对抗性的学习目标(冲突的奖励),该目标试图最大化自己的累积奖励同时最小化竞争对手的奖励.
混合任务	利己主义奖励	对于利己主义智能体组成的多智能体系统,可以将不依赖于其他智能体的个体奖励分配给每个智能体. 因此在学习过程中,合作和竞争行为可能共存.

3.2　多智能体深度强化学习方法

在本小节中,我们试图建立一个系统的多维度的 MADRL 分类方式,将目前的 MADRL 工作方法进行归类和总结. 如前文所述,强化学习方法可根据求解最优策略的方式不同分为基于值函数,基于策略的方法,可根据适用的任务类型分为合作任务及

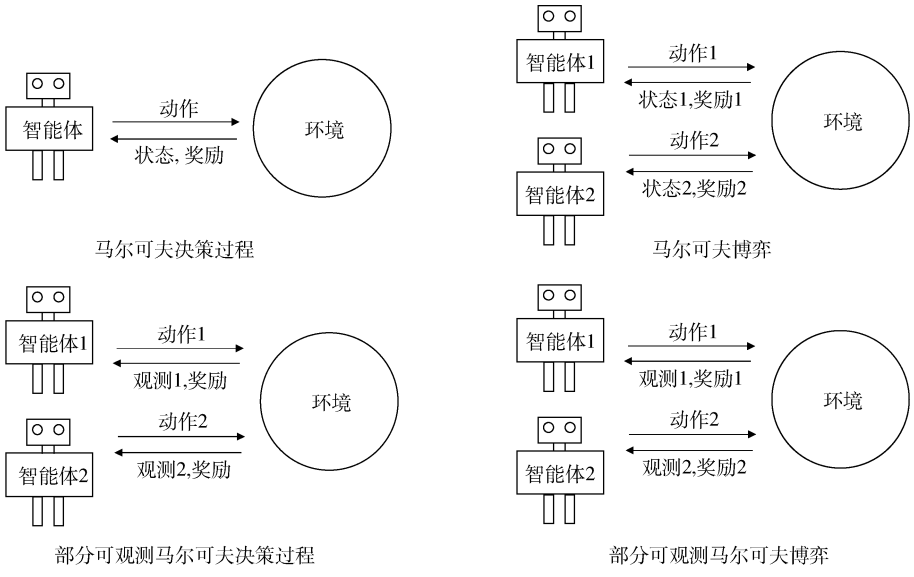

图 1 多智能体强化学习中的主要问题表示

一般任务(合作,竞争或混合任务).此外,我们根据智能体之间是否建立通信机制,将 MADRL 分为有通信和无通信的方法.基于以上三个维度,本小节将前沿的 MADRL 方法分为八个类别,如表 2 所示.接下来我们面向每个类别对目前的主流方法进行研究和分析.MADRL 方法通常存在几个共同的问题:(1)可扩展性:随着智能体数量的增加,动作空间指数增长,造成计算复杂度高;(2)非平稳性:每个智能体的行为通常会引发环境的改变,使环境变得不稳定,进而影响其他智能体的动作选择和策略选择;(3)奖励问题:由于一般任务中不同智能体的任务和学习目标通常不同,同时智能体间的互相作用会致使目标奖励的确定变得困难,进而严重影响方法的收敛.合作任务中智能体目标一致但存在信用分配问题,即怎样为每个智能体正确地分配奖励信号以更好地协调,最大化总体收益.

表 2 方法分类

	基于值函数		基于策略或 AC 框架	
	无通信	有通信	无通信	有通信
合作任务 (Dec-POMDP)	VDN[31][AAMAS 2018] QMIX[32][JMLR 2020] WQMIX[33][NIPS 2020] QPLEX[34][ICLR 2020] QTRAN[35][ICML 2019] Q-DPP[36][ICML 2020] UneVEn[37][ICML 2021] QPD[38][ICML 2020] REFI[39][ICML 2021] PAC[40][NIPS 2022] ResQ[41][NIPS 2022] IGM-DA[42][NIPS 2022]	DIAL[49][NIPS 2016] RIAL[49][NIPS 2016] VBC[50][NIPS 2019] TMC[51][NIPS 2020] MA-HA[52][TNNLS 2022] NDQ[53][ICLR 2019] MAIC[54][AAAI 2022] THGC[55][APIN 2021] MASIA[56][NIPS 2022] SOG[57][NIPS 2022]	COMA[61][AAAI 2018] BAD[62][ICML 2019] SAD[63][ARXIV 2019] GB[64][ICML 2022] LINDA[65][ARXIV 2021] CoachReg[66][NIPS 2020] LICA[67][NIPS 2020] FACMAC[68][NIPS 2021] DOP[69][ICLR 2020] DAE[70][ICML 2022]	CommNet[76][NIPS 2016] BicNet[77][ARXIV 2017] ATOC[78][NIPS 2018] SchedNet[79][ARXIV 2019] IMAC[80][ICML 2020] GA2Net[81][AAAI 2020] I2C[82][NIPS 2020] IS[83][ICLR 2021] Flowcomm[84][AAMAS 2021] HetGAT[85][AAMAS 2022]
一般任务 (POMG)	MFQ[43][ICML 2018] MTMFQ[44][ARXIV 2020] MFVFD[45][IJCAI 2021] MDQ[46][PR 2023] DMFG[47][AAAI 2022] GAMFQ[48][ARXIV 2023]	DGN[58][ICLR 2018] MARGIN[59][TNNLS 2022] LSC[60][AAMAS 2022]	MADDPG[71][NIPS 2017] MA3DPG[72][AAAI 2019] MAAC[73][ICML 2019] EOI[74][ICML 2021] MACPF[75][ARXIV 2022]	TarMAC[86][ICML 2019] IC3Net[87][ARXIV 2018] MAGIC[88][AAMAS 2021]

3.2.1　基于值函数的 MADRL 方法

为解决以上的问题,基于值函数的方法大致可分成 2 类:值函数分解方法和平均场方法.值函数分解方法利用各种约束或限制将全局价值函数分解为多个智能体的个体价值函数的组合,同时确保在个体价值函数最优时,全局价值函数也实现最优.平均场方法假设个体智能体间的互相作用能够表示成个体智能体与总体平均效应间的互相作用,即能够使用状态-动作分布来替代联合状态-动作,进而降低联合状态-动作的维度.一般情况下,值函数分解方法适用于合作任务,平均场方法适用于一般任务.接下来我们对相关工作进行详细的综述.

(1)合作任务+无通信

值函数分解方法将全局价值函数分解为个体价值函数,很大程度上降低了动作空间的维度,解决了可拓展性问题.同时每个智能体仅根据个体价值函数选择动作,解决了智能体之间互相影响导致的非平稳性问题以及信用分配问题.值函数分解方法以其明显的优势已经成为 MADRL 的一个主流研究方向.值函数分解方法一般包括两类值函数:全局价值函数和个体价值函数,所有智能体共享全局价值函数.个体价值函数的输入是智能体的本地局部观测,输出是个体的 Q 值来引导智能体选择和执行动作.因为值函数分解方法中全局价值函数被所有智能体共享,所以该方法一般只适合处理合作任务.本小节主要关注智能体之间无通信机制的值函数分解方法.Sunehag 等[31]提出了第一个值函数分解方法:值分解网络(Value Decomposition Network, VDN),将全局价值函数分解为局部个体价值函数的线性累加.然而 VDN 假设了严格的累加限制,并忽略了在训练阶段可以访问的任何额外信息.为解决 VDN 的局限性,Rashid 等[32]提出了 QMIX,使用一个单调混合网络结构对全局价值进行估计,该结构实现对个体价值函数的单调非线性聚合.然而,QMIX 无法表示具备非单调性特性的全局价值函数,于是在应对非单调类型任务时,QMIX 可能陷入局部次优而无法收敛到最佳策略.

为解决这个问题,加权 QMIX(Weighted QMIX, WQMIX)[33]引入了权重网络来对各个智能体的 Q 值进行加权融合,以促进各个智能体之间的协作学习.Wang 等[34]提出双对决多智能体 Q-学习方法(duplex dueling multi-agent Q-learning, QPLEX)来分解价值函数,利用优势函数约束完成对价值函数的高效学习.价值函数分解方法之间的主要差异在于混合网络,目前的价值函数分解方法其混合网络的表征复杂性不断增加.Son 等[35]提出基于变换的多智能体 Q-学习(multi-agent Q-learning with Transformation, QTRAN),利用更通用的方式分解价值函数,摆脱了单调性或累加性的结构约束.

Yang 等[36]提出一种基于行列式点过程的 Q-学习方法(Q-learning with Determinantal Point Process, Q-DPP),使用一种概率化的集合建模方式,不但考虑了使智能体奖励最大化的动作,而且考虑了智能体行为的多样性.通过学习无结构约束的值函数分解,Q-DPP 解决了上述几种方法的局限性.Gupta 等[37]提出一种通用价值探索方法(Universal Value Exploration, UneVEn),它通过对通用后继特征的线性分解同时学习一组相关任务,利用已经解决的相关任务的策略来改进所有智能体的联合探索过程,帮助它们更好地协调动作.Yang 等[38]提出一种 Q 值路径分解方法(Q-value Path Decomposition, QPD),将系统的全局 Q 值分解为单个智能体的 Q 值.与上述限制个体 Q 值与全局 Q 值表示关系的研究不同,QPD 将积分梯度归因技术应用到 MADRL 中,直接沿轨迹路径分解全局 Q 值,为智能体分配信用.

Iqbal 等[39]提出随机实体分解(Randomized Entity-wise Factorization for Imagined Learning, REFIL)方法,正则化价值函数以共享由实体子组组成的因子,从而促进复杂合作多智能体任务内部和之间的泛化和知识转移.Zhou 等[40]证明了智能体对其自身行为的排序可能会对可表示的函数类施加并发约束(跨不同状态),从而在训练期间导致显著的估计误差.为解决了这一限制,他们提出 PAC 方法,利用最优联合行动选择的反事实预测作为辅助信息为价值函数分解提供帮助.

现有的值函数分解方法受到表征能力、样本效率和近似误差的限制,为了解决这些问题,Shen 等[41]提出一种值函数分解方法 ResQ,通过残差函数找到任何值函数的最优联合策略.该方法既能满足个体全局最大值(Individual Global Max, IGM),又能满足分布 IGM 原则,没有表示限制.Hong 等[42]指出了 IGM 分解是一种有损分解,有损分解产生的误差可能在训练过程中积累,进而可能会严重降低值函数分解方法的性能.为了解决上述问题,他们提出了 IGM-DA,采用模仿学习策略将有损分解与贝尔曼迭代分离,从而避免误差积累.

(2)一般任务+无通信

平均场方法是基于值函数的方法中另一个主要

的方法,将多智能体之间的互动简化为单个智能体与其它智能体的平均效应之间的互动.平均场方法由于不局限于学习全局价值函数,可以适用于合作、竞争或混合合作竞争任务.Yang 等[43]提出第一个平均场方法:平均场 Q-学习(Mean Field Q-learning,MFQ),利用邻居智能体的平均动作获取动作经验分布来近似联合动作.然而,MFQ 只适用于所有智能体都属于同一种类型的场景.Subramanian 等[44]提出多类型平均场 Q-学习方法(Multi Type Mean Field Q-learning, MTMFQ),使用 k-means 方法将平均场方法扩展到包括多种类型智能体的场景.然而,随着智能体类型数量的增加,MTMFQ 的性能有所下降.Zhang 等[45]提出基于平均场的值函数分解方法(Mean Field Value Function Decomposition,MFVFD),利用平均场理论将个体价值函数分解为局部价值函数和平均场价值函数.Du 等[46]提出多智能体对决 Q-学习 (Multi agent Dueling Q learning,MDQ) 方法,引入平均场理论来评估个体价值函数受到总体的平均效应的影响.同时采用对决网络架构来系统地区分动作层级和状态层级的影响,从而实现价值函数学习的高效和稳定. Subramanian 等[47]放宽了智能体是不可区分的假设,并提出了一个新的平均场方法,称为去中心化平均场博弈(Decentralized Mean Field Games, DMFG)方法,其中每个智能体可以与其他智能体完全不同.所有智能体都基于它们的局部观测以分散的方式学习独立的策略.大多数平均场工作采用加权平均场或概率分布更新邻居智能体的平均动作,没有充分考虑邻居智能体的特征信息,容易导致局部最优.Yang 等[48]提出了一种基于图注意力的部分可观测平均场 Q 学习(Graph-Attention Mean Field Q-learning,GAMFQ)来弥补这一缺陷.GAMFQ 使用图注意力和平均场模块来描述智能体如何在每个时间步受到其他智能体动作的影响.

(3)合作任务+有通信

有通信的方法指在智能体间构建通信学习模式,在训练过程中,个体智能体能够利用其它智能体传递的通信信息进行训练学习.上述值函数分解方法侧重于完全分解的方式,即个体智能体只依靠各自的局部观测计算个体价值函数.在现实场景中,多智能体任务一般需要智能体之间的学习通信,以有效地协调多个智能体的动作完成合作任务.在缺少通信机制的方法下,智能体受到局部可观测性和随机性的限制,会加剧分散执行过程中智能体的状态和动作对其他智能体的不确定性,进而导致严重的智能体动作协作失调.学者们发现基于值函数的方法可以在训练过程中,引入通信学习机制来提高智能体之间的动作协作.

最早在基于值函数的 MADRL 引入通信学习的工作是 Foerster 等[49]提出的强化智能体间学习(Reinforced Inter-Agent Learning, RIAL)和可微分智能体间学习(Differentiable Inter-Agent Learning, DIAL).在 RIAL 和 DIAL 中,每个智能体学习共享二进制或实值信息,适用于有限的通信.DIAL 和 RIAL 智能体之间使用完全连接的结构,在具有少量智能体的完全合作环境中进行评估,并将如何更有效地进行通信的问题作为一个开放的问题.Zhang 等[50]提出基于方差的控制(Variance Based Control, VBC)方法,通过阈值来限制传输信息的方差,以过滤方差较大的信息,从而实现较低的通信成本.在训练阶段,通过限制智能体之间交换消息的方差,可以有效地去除消息中的噪声成分,同时保留有用的部分,并被智能体利用以获得更好的性能.Zhang 等[51]提出临时信息控制(Temporal Message Control, TMC)方法,不允许在一个时间步内产生类似信息的智能体广播它们的信息.然后,那些与之前发送的信息有很大不同的信息将被共享给所有智能体.此外,TMC 使用信息缓冲区来存储接收到的信息,以补偿丢失的信息.Du 等[52]提出一种基于异质图注意力网络的通信机制(Multi-Agent reinforcement learning with Heterogeneous graph Attention network, MA-HA),该方法充分考虑了智能体层次和关系层次的重要性,通过分层聚合相邻智能体的潜在特征信息,生成每个智能体的融合特征表示.

Wang 等[53]提出一种学习近似可分解 Q 函数(Nearly Decomposable Q-functions, NDQ)的新框架,其中智能体在大多数时间步内自主行动,但偶尔会向其他智能体发送信息以有效协调动作.该框架通过引入两个信息论正则化器,将价值函数分解学习和通信学习融合.这些正则化器最大限度地提高了智能体的动作选择和通信信息之间的互信息,同时使智能体之间的信息熵最小化.Yuan 等[54]提出一种多智能体激励通信框架 (Multi-Agent Incentive Communication, MAIC),该框架允许每个智能体通过学习产生激励信息,并直接影响其他智能体的价值函数,从而实现有效的显式协调.此外,MAIC 引入了一种新的正则化来利用交互稀疏性并提高通信效率.

Jiang 等[55]提出一种基于类型的分层通信(Type-based Hierarchical Group Communication, THGC)模型,首先利用先验领域知识或预定义规则对群体智能体进行分类,通过知识共享保持群体的认知一致性.随后,引入了一种群体通信和价值分解方法,以确保各个群体之间的合作.Guan 等[56]提出基于自监督信息聚合的多智能体通信(Multi-Agent communication via Self-supervised Information Aggregation, MASIA),智能体可以将接收到的信息聚合成具有高相关性的紧凑表示,从而增强局部策略.Shao 等[57]提出一种自发分组机制,称为自组织组(Self-Organized Group, SOG),其特点是指挥选举和信息汇总机制.在指挥选举过程中,每一定时间步选出一定数量的指挥者来临时构建组,每个组具有指挥者-追随者共识,其中追随者被约束仅与他们的指挥者通信.在信息汇总过程中,每个指挥者汇总并分发接收到的信息给所有附属组成员,以保持统一调度.

(4)一般任务+有通信

智能体之间的通信机制一般用于协调智能体的动作以完成合作任务,还有一些基于值函数的工作引入了通信来解决更一般的任务.Jiang 等[58]提出图卷积强化学习(graph convolutional reinforcement learning, DGN),其中图卷积通过卷积核来捕获智能体之间的相互作用.利用逐渐增加的接受域的卷积层产生的潜在特征来学习特征,并通过时间关系正则化来进一步提高一致性.Ding 等[59]提出基于图信息最大化的多智能体强化学习(Multi-Agent Reinforcement learning with Graphical mutual Information maximization, MARGIN)方法,以最大化相邻智能体的输入特征信息与输出隐藏特征表示之间的相关.该方法将传统的互信息优化思想从图域扩展到多智能体系统,从智能体特征信息和智能体拓扑关系两个方面对互信息进行优化.Sheng 等[60]提出结构化通信学习(Learning Structured Communication, LSC)方法,通过使用更灵活和高效的通信拓扑,允许自适应智能体分组,以形成不同的分层结构.给定每个形成的拓扑结构,学习一个分层图神经网络,以实现有效的信息生成和在组间和组内通信之间传播.此外,该工作总结了通信学习方法的一般通信结构,如图 2 所示.全连接型结构中智能体需要与所有其他智能体通信,因此当智能体数量较大时,需要较高的带宽.星型结构中所有智能体都需要将信息传输到虚拟中心智能体,这导致了一个很大的通信瓶颈.邻居型结构中智能体与邻居智能体同时通信以降低通信成本.树型结构中智能体只与邻居通信,但是必须允许组之间按顺序进行通信,导致时间复杂度很高.分层结构中智能体被分到不同的组,每个组都有一个高级智能体,实现组内、组间通信.

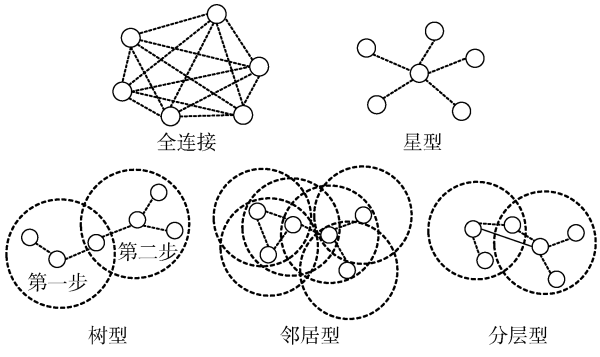

图 2　通信结构类型[60]

3.2.2　基于策略的 MADRL 方法

在多智能体深度强化学习中,另一类方法是基于策略的方法,这类方法能够直接对策略的参数进行梯度更新.基于值函数和直接策略优化的方法各有优势,演员评论家(Actor-Critic, AC)方法就是融合以上两类方法优势的强化框架.这种方法包含两个主要部分,一个是演员家网络(actor network),它表示策略网络,主要通过策略梯度来进行优化求解.另一个是评论家网络(critic network),它是一个值函数.AC 框架是策略梯度方法和 Q-学习的结合,也被广泛应用于多智能体强化学习方法中.在本文中,我们将基于 AC 框架的方法归类到基于策略的方法中.同样的,下面我们根据智能体之间是否有通信,以及任务类型对基于策略的 MADRL 方法进行分类和综述.

(1)合作任务+无通信

在解决一些完全合作任务时,常见的基于策略的方法之一是推理其他智能体的行为.这通过建立模型来实现,智能体可以预测模型智能体可能感兴趣的属性,如行为、目标和信念.这个模型通常是一个函数,它利用观察到的交互历史的部分作为输入,并输出关于所建模智能体的相关属性的预测.Foerster 等[61]设计了反事实多智能体策略梯度(Counterfactual Multi-Agent policy gradients, COMA)方法,旨在解决在多智能体环境中的合作与竞争问题.COMA 方法的核心思想是引入对抗性训练,使得每个智能体都可以从其他智能体的行为中学习.它通过训练一个对抗性网络来估计其他智能体的行为策略,并利用这些估计来计算每个智能体的价值函数.这样,

每个智能体可以根据其他智能体的行为来调整自己的策略,以获得更好的回报. Foerster 等[62]提出贝叶斯动作解码器(Bayesian Action Decoder, BAD),通过探索局部确定性策略的空间而不是动作,并基于贝叶斯信念这个知识从分布中采样这些策略来实现这一解决方案. 虽然它有效地解决了这个问题,但它是以牺牲简单性和通用性为代价的. Hu 等[63]提出简化动作解码器(Simplified Action Decoder, SAD),是另一种解决探索与利用冲突的方法. 在 SAD 中,智能体在每个时间步中采取两个动作:一个是典型的环境动作,它通常由环境执行,并作为环境状态的一部分被其他智能体观测到. 另一个是一个贪婪动作,不被智能体执行,但在下一步作为其他智能体的补充输入. Muglich 等[64]提出一个跨多个策略学习广义信念(General Belief, GB)的方法,其中模型学习对每个策略的专门约定进行推理.

Cao 等[65]提出多智能体局部信息分解(Local Information Decomposition for Awareness of teammates, LINDA)方法,通过该框架,智能体学习分解局部信息并为每个队友建立意识. 该方法将感知建模为随机变量,并通过最大化感知与相应智能体实际轨迹之间的互信息来进行表征学习,以确保感知表征的信息性. Roy 等[66]鼓励使用协调策略来简化在合作多智能体任务中发现成功策略的过程,并提出了 CoachReg 方法来促进 MADRL 算法的协调. Zhou 等[67]提出一种多智能体演员评论家方法 LICA,旨在隐式地解决完全合作条件下的信用分配问题. Peng 等[68]提出因子多智能体集中策略梯度(Factored Multi-Agent Centralised policy gradients, FACMAC),是一种在离散和连续动作空间中进行合作多智能体强化学习的新方法. Wang 等[69]提出(multi-agent decomposed policy gradient method, DOP),DOP 将值分解的思想从价值函数分解方法引入到多智能体 AC 框架中. 基于这一思想,DOP 支持有效的非策略学习,可以适用于离散和连续行动空间中的任务. Li 等[70]提出差异优势估计(Different Advantage Estimation, DAE),一种指数加权优势估计器,用于解决多智能体策略梯度方法中明确的多智能体信用分配问题.

(2)一般任务+无通信

在应对混合竞争合作任务时,一个主流的解决方案是基于 AC 框架的 MADRL 方法. Lowe 等[71]首先将 AC 框架拓展到多智能体系统,提出多智能体深度确定性策略梯度(Multi Agent Deep Deterministic Policy Gradient, MADDPG)方法. 该方法允许智能体在训练过程中使用其他信息来增强训练. 训练阶段完成后,在执行阶段只使用局部演员家网络,指导智能体以分散的方式选择和执行动作. MADDPG 首次将集中训练与分散执行(Centralized Training and Decentralized Execution, CTDE)范式引入到 MADRL. 基于 CTDE 框架,一系列基于 AC 框架的 MADRL 方法被提出. 在 MADDPG 基础上,Li 等[72]设计了 minimax DDPG 方法. 该方法旨在解决多智能体在连续动作场景中训练的鲁棒性问题,以确保训练后的智能体在对手策略变化时仍能具备泛化能力. Shariq 等[73]提出多智能体演员家-注意力-评论家网络(Multi-Actor-Attention-Critic network, MAAC),该方法在多智能体设置中使用集中计算的评论家网络来训练分散的策略,这些评论家共享一个注意力机制,该机制在每个时间步为每个智能体选择相关信息. Jiang 等[74]提出一种关注个性的出现(Emergence of Individuality, EOI)的多智能体强化学习方法. EOI 学习了一个概率分类器,该分类器根据智能体的观测结果预测其概率分布,并给每个智能体一个被分类器正确预测的内在奖励. 内在奖励鼓励智能体访问他们自己熟悉的观测,并且通过这些观测来学习分类器使内在奖励信号更强,从而使智能体更容易识别这些奖励信号. Wang 等[75]提出多智能体条件策略分解(Multi-Agent Conditional Policy Factorization, MACPF),显式地将智能体之间的依赖关系纳入集中训练,能够从一个策略推导出另一个联合策略,该联合策略实现了相同的最优性.

(3)合作任务+有通信

Sukhbaatar 等[76]提出通信网络(Communication Network, CommNet),CommNet 学习了一个共享神经网络,该神经网络用来编码智能体的局部观测. 每个智能体的决策依赖于来自其他智能体的观测结果和通信信息的平均向量. 原则上,智能体可以使用共享神经网络的副本以分散的方式在环境中执行,同时需要与所有智能体进行即时通信. Zhu 等[77]提出多智能体双向协调网络(Bidirectionally Coordinated Network, BiCNet),其通过双向长短期记忆网络层连接每个智能体的策略网络和价值函数. 因此,智能体能够捕获具有长期依赖关系的其他记忆状态,并相应地交换信息.

Jiang 等[78]提出注意力通信模型(Attentional Communication, ATOC),智能体只能与可观测范围内的某些智能体进行通信. 只有邻近的智能体才

能参与到一个通信组中,这是由概率门机制决定的.在通信组中,双向长短期记忆网络层用于自动组合来自每个智能体的信息并将其发送回每个成员. Kim 等[79]提出通信调度网络(communication Scheduling Network, SchedNet),考虑了有限带宽和共享通信通道约束下的通信开销和争用问题,只选择有限数量的智能体将信息发送到通信通道. SchedNet 通过学习每个智能体的部分观测到的信息的重要性,来决定哪些智能体能够广播它们编码过的信息. Wang 等[80]提出信息丰富的多智能体通信模型(Informative Multi-Agent Communication, IMAC),定义了一个调度器,它聚合来自所有智能体的编码信息,并向每个智能体发送单独的消息.

Liu 等[81]提出基于图注意力网络的通信机制(Graph Attention communication mechanism, GA2Net),学习全局信息处理器,根据智能体的权重集成信息. Ding 等[82]发现现有的工作侧重于广播通信,通常会导致信息冗余,甚至会影响学习过程,提出个体推断通信(Individually Inferred Communication, I2C),该模型可以简单而有效地帮助智能体学习通信的先验,通过评估其他智能体对一个智能体自身策略的影响,学会双边地决定是否与其他每个智能体进行通信. Kim 等[83]提出一种通信模型称为意图分享(Intention Sharing, IS)模型,以增强智能体之间的协调能力.每个智能体通过对环境动力学和其他智能体的动作建模来生成想象轨迹.想象轨迹是基于学习到的环境动力学模型和其他智能体的模型,模拟每个智能体的未来轨迹,代表每个智能体的未来行动计划.每个智能体压缩捕获其未来行动计划的想象轨迹,通过应用注意机制来根据从其他智能体接收的信息学习想象轨迹中组件的相对重要性,从而生成其意图信息用于通信. Du 等[84]提出基于规范化流编码的图通信模型(communication mechanism based on graph encoded by a normalizing flow, Flowcomm),采用一种规范化流程来编码智能体交互之间的相关性.动态通信拓扑是通过最大化智能体奖励来直接学习的.通过学习到的动态通信拓扑,实现了对评论家网络和图推理策略的集中训练,以及对局部观测和接收到的信息的分散执行. Seraj 等[85]提出异质策略网络(Heterogeneous policy Networks, HetNet)来学习高效和多样化的通信模型,以协调合作的异质团队. HetNet 不仅促进了每个现有智能体类的异质协作策略的学习,而且还支持端到端训练以学习高效的二值化信息.

(4)一般任务+有通信

Das 等[86]提出目标多智能体通信(Targeted Multi-Agent Communication, TarMAC)模型,其中智能体在部分可观测的环境中执行一般任务时,既可以学习发送什么信息,也可以学习向谁发送信息.同时, TarMAC 还通过多轮通信方法增强了通信学习能力. Singh 等[87]提出个性化控制连续通信模型(Individualized Controlled Continuous Communication model, IC3Net),从 CommNet 扩展而来,也使用门控机制,同时确定地将信息发送给所有智能体或根本不发送给智能体.此外,IC3Net 对每个智能体采用个性化的奖励,而不是像 CommNet 那样采用全局共享的奖励,从而在一般任务中表现出更多样化的行为. Niu 等[88]提出多智能体图注意力通信(Multi-Agent Graph-attention Communication, MAGIC)模型,通过调度模块来帮助解决何时通信以及向谁发送信息的问题,同时使用带有动态图的图注意力网络的信息处理器来处理通信信号.调度模块由一个图注意力编码器和一个可微分注意力机制组成,可微分注意力机制向信息处理器输出动态的、可微分的图,从而使调度模块和信息处理模块能够端到端进行训练.

3.2.3　其他分类方式

本章最后介绍了一些其他分类方式,重点分类值函数分解方法,基于通信的 MADRL 方法以及基于图神经网络的 MADRL 方法.首先对价值函数分解方法进行了分类,分为简单因子分解型、基于 IGM 原则型,和其他类型,如表 3 所示.

简单因子分解型方法采用直接分解联合动作价值函数为个体动作价值函数的方式,例如简单相加或单调性约束等,而不额外引入其他分解机制.简单因子分解型方法比较容易实现,但其受到很强的限制,导致其表达能力有限.因此简单因子分解型方法更适合智能体之间关系简单的情景.个体全局最大值(Individual Global Max, IGM)原则是用于实现值函数分解的一种常用原则. IGM 原则降低了限制条件,并可以确保算法的收敛性.然而,由于需要调整的参数较多,IGM 原则型方法的实现难度较高,更适用于较少智能体数量的场景.此外,还有其他类型的方法,通常关注智能体之间关系较复杂的多智能体环境或大规模多智能体场景.

然后对所有通信方法按照通信结构进行了分类,分为全连接型、星型、树型、邻居型和分层型,这些类型在前文中已经进行了介绍,分类结果如表 4

所示.此外,表 5 根据图结构、图神经网络结构、信息利用方式、智能体类型,对基于图神经网络的通信学习方法进行了对比.其中图结构根据智能体构成的图的方式,分为完全图(所有智能体全连接),学得的图(通过学习构建图),以及基于环境的图(如根据智能体观测范围或空间位置构建的图).图神经网络结构包括图卷积网络(Graph Convolutional Network, GCN),图注意力网络(Graph Attention network, GAT),以及异质图注意力网络(Heterogenous Graph Attention network, HGAT).

表 3 值函数分解方法分类

类型	方法	方法简介
简单因子分解型	VDN[31][AAMAS 2018]	使用值分解网络将全局价值函数分解为局部个体价值函数的线性累加
	QMIX[32][JMLR 2020]	使用单调混合网络结构将全局价值函数分解为个体价值函数的单调非线性聚合
	WQMIX[33][NIPS 2020]	更新网络过程中对联合动作值的平方误差进行加权操作,以获得最优联合策略
基于 IGM 原则型	QPLEX[34][ICLR 2020]	将 IGM 原则对动作价值函数的一致性约束转化为对优势函数的一致性约束
	QTRAN[35][ICML 2019]	采用更通用的方式分解价值函数,摆脱了单调性或累加性的结构约束
	Q-DPP[36][ICML 2020]	概率化的集合建模方法,考虑了智能体行为的多样性,学习没有结构约束的值函数分解
	UneVEn[37][ICML 2021]	同时学习一组相关任务,并对通用后继特征进行线性分解改进所有智能体的联合探索过程
	ResQ[41][NIPS 2022]	通过残差函数找到任何值函数的最优联合策略既,能满足 IGM 原则,又能满足分布 IGM 原则
	IGM-DA[42][NIPS 2022]	证明 IGM 是有损分解,采用模仿学习策略将有损分解与贝尔曼迭代分离,从而避免误差积累
其他类型	QPD[38][ICML 2020]	使用积分梯度归因技术,直接沿轨迹路径分解全局价值,为智能体分配信用
	REFIL[39][ICML 2021]	个体动作价值函数分成组内效用和组外效用,并利用因子分解的损失作为辅助目标完成训练
	PAC[40][NIPS 2022]	利用最优联合行动选择的反事实预测作为辅助信息为价值函数分解提供帮助

表 4 根据通信结构划分

类型	特点	方法
全连接型	智能体需要与所有其他智能体通信,因此当智能体数量较大时,需要较高的带宽.	DIAL[49], RIAL[49], VBC[50], TMC[51], NDQ[53], MAIC[54], THGC[55], MASIA[56], BicNet[77], SchedNet[79], TarMAC[86]
星型	所有智能体都需要将消息传输到虚拟中心智能体,这导致了一个很大的通信瓶颈.星型结构确保所有智能体都可以访问消息,但一旦大量的信息同时出现,提取有价值的信息将变得非常困难.	SOG[57], CommNet[76], IMAC[80], I2C[82], IS[83], GA2Net[83], IC3Net[87], MAGIC[88]
树型	智能体只与邻居通信,但是必须允许组之间按顺序进行通信,导致时间复杂度很高.	ATOC[78]
邻居型	智能体与邻居智能体同时通信以降低通信成本.基于图卷积神经网络的通信结构很多使用这种通信方式,树结构和邻域结构限制了与邻域的通信从而能够提高信息的理解能力.	MA-HA[52], DGN[58], MARGIN[59], FlowComm[84], HetGAT[85]
分层型	智能体被分到不同的组,每个组都有一个高级智能体,通过组内、组间通信,实现全局通信.	LSC[60]

表 5 基于图神经网络的通信学习方法简介

方法	图结构	图神经网络结构	信息利用	智能体类型
CommNet[76][NIPS 2016]	完全图	GCN	基于策略	同质
IC3Net[87][ARXIV 2018]	完全图	GCN	基于策略	同质
DGN[58][ICLR 2018]	基于环境	GCN	基于价值	同质
TarMAC[86][ICML 2019]	完全图	GAT	基于价值和策略	同质
GA2Net[81][AAAI 2020]	学得的图	GAT	基于价值和策略	同质
MAGIC[88][AAMAS 2021]	学得的图	GAT	基于价值和策略	同质
LSC[60][AAMAS 2022]	学得的图	GCN	基于价值	同质
MA-HA[52][TNNLS 2022]	基于环境	HGAT	基于价值	异质
MARGIN[59][TNNLS 2022]	基于环境	GAT	基于价值	同质
HetGAT[85][AAMAS 2022]	基于环境	HGAT	基于策略	异质

4　多智能体深度强化学习的应用

多智能体深度强化学习领域在过去几年取得了长足的进步,在各种游戏中取得超越人类水平的表现[89-90],多智能体深度强化学习技术还广泛地应用于各种现实任务中.表 6 概述了 MADRL 在这几种应用场景中的代表方法.

表 6　MADRL 在主要应用场景中的方法介绍

应用场景	文献	方法	工作概述
交通信号控制	Calvo 等[91]	IDQN	采用对决网络结构,双 Q 学习,可处理异构场景
	Chen 等[92]	QMIX	在 QMIX 引入通信模块实现了有效的全局协调
	Chu 等[93]	A2C	完全可扩展和分散,提高每个智能体的局部可观测性
	Devailly 等[94]	IG-RL	分散式学习,可扩展性强,适应任何交通网络
	Huang 等[95]	DSTAN	图卷积网络捕获空间依赖性,注意力机制学习交叉口重要性
	Yang 等[96]	IHG-MA	异质图编码每个节点的异质特征和图结构的异质信息
	Yang 等[97]	IHG-MDGI	异质图注意力网络融合异质特征,互信息优化价值函数
自动驾驶	Zhou 等[98]	A2C	综合燃油效率、驾驶舒适性和自动驾驶安全性的多目标奖励函数
	Han 等[99]	Shapley value	重新分配系统的总奖励,以激励自动驾驶汽车之间的稳定合作
	Palanisamy 等[100]	MACAD	提供了一套可扩展的互联自动驾驶仿真环境
	Zhou 等[101]	SMARTS	开发了一个专用的仿真平台-可扩展多智能体强化学习培训平台
推荐系统	An 等[102]	SAC	信号网络向所有模块发送额外信号,引入熵正则化协调智能体的探索
	Zhang 等[103]	Master	综合考虑各种长期时空因素,智能推荐公共无障碍充电站
	Zhang 等[104]	Value-based	动态合作推荐,通过几个不同的学者相似度测量来优化选择
	Gui 等[105]	AC-based	设计了一个反向动作机制为两个策略梯度智能体获得差异化奖励
	Wang 等[106]	MFHRL	多智能体平均场分层强化学习,为电动出租车提供收费和调动推荐

4.1　交通信号控制

近年来,为优化路口的交通流量解决交通拥堵问题,各种方法不断涌现.其中最有前途的技术是深度强化学习,因为它能够在没有任何人为干预的情况下学习复杂问题的动态.然而,由于交叉路口的数量导致状态和动作空间呈指数增长而产生的维数诅咒,最新的研究利用多智能体深度强化学习方法来解决交通信号控制问题.Calvo 等[91]指出同质路口的设定并不是一个真实世界的场景,因为一个城市通常有不同的十字路口布局,并使用独立深度 Q 网络(Independent DQN, IDQN)来训练异质的多智能体来处理维度的诅咒和协作的需要.IDQN 采用对决网络结构,分别计算值函数和优势函数,提高训练速度,从而提高深度 Q-Network 的性能和稳定性.同时使用双 Q 学习通过防止过于乐观的值估计来选择更好的 Q 值.Chen 等[92]提出了一种基于 QMIX 的通信方法(Communication-QMIX, CQMIX),首先将 QMIX 作为平衡大规模和有效优化的基础,利用其集中训练和分散执行的机制,然后引入通信模块实现了有效的全局协调.Chu 等[93]提出了一种完全可扩展和分散的 MARL 算法,多智能体优势演员评论家(multi-agent A2C, A2C).特别地,该方法通过提高每个智能体的局部可观测性和降低学习难度使学习过程更加稳定.

Devailly 等[94]引入了基于图卷积网络的归纳图强化学习(Inductive Graph Reinforcement Learning, IG-RL),它可以适应任何交通网络的结构,以学习交通信号控制器及其周围环境的详细表示.通过分散式方法使学习可转移自适应交通信号控制策略成为可能.在对任意一组交通网络进行训练后,该模型可以推广到新的交通网络和交通分布,具有很强的可扩展性.Huang 等[95]提出一种基于深度时空-通道注意力神经网络的多智能体强化学习方法(MARL-deep spatiotemporal attentive neural network, MARL-DSTAN)来确定大规模交通网络中的交通信号配时.该模型利用图卷积网络捕获整个交通网络的空间依赖性,并通过注意力机制基于交叉口的重要性对信息进行整合.同时,为了积累更多有价值的样本,提高学习效率,在探索阶段引入了循环神经网络来约束动作搜索空间,而不是完全随机探索.Yang 等[96]提出一种用于多路口交通信号控制的算法,称为归纳异质图多智能演员评论家方法(Inductive Heterogeneous Graph Multi-Agent actor-critic, IHG-MA).与基于同质图神经网络的算法不同,IHG-MA 算法不仅编码每个节点的异质特征,而且编码异质的图结构信息.Yang 等[97]提出一种基于归纳异质图注意力的

多智能体深度图信息方法(Inductive Heterogeneous graph Attention-based Multi-agent Deep Graph Infomax, IHA-MDGI),IHA-MDGI 方法通过提出的归纳异质图注意力方法进行特征融合,并通过提出的多智能体深度图信息框架进行训练.在 MDGI 框架中,设计了一个互信息损失函数,该函数与演员评论家损失函数结合进行联合训练.互信息损失函数关注的是交通网络异质信息表示与 Q 值之间互信息的最大化,可以使 Q 值包含更多关于异质信息的信息.

4.2 自动驾驶

在 21 世纪以来,自动驾驶领域引起了学者的广泛关注,因为自动驾驶技术能够提供许多潜在的收益,比如将司机从疲惫的驾驶中解脱出来,缓解交通拥堵等.近年来,MADRL 在自动驾驶中的应用得到了广泛的探索和研究,并已经取得了令人鼓舞的成功.Zhou 等[98]将混合交通高速公路环境中多自动驾驶汽车的变道决策制定为一个 MADRL 问题,其中每个自动驾驶汽车根据相邻自动驾驶汽车和自动驾驶汽车的运动做出变道决策.具体而言,提出了一种具有局部奖励设计以及参数共享模式的多智能体优势演员评论家方法.特别地,设计了综合燃油效率、驾驶舒适性和自动驾驶安全性的多目标奖励函数.Han 等[99]提出有效地重新分配系统的总奖励,以激励自动驾驶汽车之间的稳定合作.在提出的可转移效用博弈中,正式定义并量化了如何将系统的总奖励重新分配给每个智能体,从而使多智能体之间基于通信的合作增加系统的总奖励.

Palanisamy 等[100]提出使用局部可观测马尔可夫博弈制定具有现实假设的互联自动驾驶问题,根据任务的性质、智能体的性质和环境的性质提供了多智能体学习环境的分类,以对各种自动驾驶问题进行分类.同时,他们提供了一个多智能体连接的自动驾驶智能体学习平台,以进一步研究这一方向.Zhou 等[101]开发了一个专用的仿真平台-可扩展多智能体强化学习培训(Scalable Multi-Agent RL Training School, SMARTS).SMARTS 支持训练、积累和使用道路使用者的各种行为模型,以深入和广泛地研究多智能体的相互作用.自第一次内部发布以来,SMARTS 已经成功支持了三次国际自动驾驶比赛,其中数千个提交的智能体模型已被自动评估并排名.

4.3 推荐系统

推荐系统已经成为我们日常生活中不可分割的一部分,它帮助我们找到想买的东西,找到社交网络上的朋友,找到想看的电影.传统上,推荐问题被认为是一个分类或预测问题,但现在学者普遍认为将其表述为顺序决策问题可以更好地反映用户-系统交互.因此,可以将其表述为马尔可夫决策过程,并通过强化学习方法求解.为解决在线购物平台各推荐模块可能存在竞争导致全局策略次优的问题,An 等[102]提出了一种新的多智能体合作强化学习方法,该方法限制了不同模块之间不能通信.首先,受博弈论中相关均衡解概念的启发,该方法设计了一个信号网络,通过对不同模块产生信号(向量)来促进各模块之间的合作.其次,提出了信号网络的熵正则化版本协调智能体探索全局最优策略.Zhang 等[103]针对电动汽车司机往往找不到合适的充电地点的问题,提出一个多智能体时空强化学习框架(Multi-agent Spatia-temporal Reinforcement Learning, MSRL),用于综合考虑各种长期时空因素,智能推荐公共无障碍充电站.具体而言,通过将每个充电站视为一个独立的智能体,将该问题表述为一个多目标多智能体强化学习任务,并开发了一个多智能体演员-评论家框架,关注智能体之间协调推荐.

在一个跨学科的环境中,科学合作正变得越来越重要,帮助学者正确选择潜在合作者是取得科学成功的关键.Zhang 等[104]提出了一种新的动态合作推荐方法,将多智能体强化学习技术应用于合作者网络分析.合作者的选择是通过几个不同的学者相似度测量来优化的.与以往的研究不同,该方法具有学术竞争的特点,即不同的学者在每次迭代中都会竞争潜在的合作者.Gui 等[105]提出使用一种新颖的合作多智能体强化学习方法来解决类似推特的社交网络中的推荐问题,它比以前的方法包含了更多的历史推文.该方法可以有效地选择一小部分历史推文,并从用户和被提及用户中协同提取相关的指标推文.为解决所有智能体的奖励信号相同的问题,该方法设计了一个反向动作机制,为两个策略梯度智能体获得差异化奖励.

Wang 等[106]设计了一个针对电动出租车司机的联合收费调动推荐系统(joint Charging and Relocation recommendation system, CARE),从电动出租车司机的角度出发,将其决策制定为一个多智能体强化学习问题,其中每个电动出租车司机的目标是最大化自己的累积奖励.该方法提出了一个新的多智能体平均场分层强化学习框架(Multi-agent mean Field Hierarchical Reinforcement Learning, MFHRL),有助于为使用 CARE 的电动出租车司机提供有远见的收费和调动建议.

5　总结与展望

多智能体强化学习作为机器学习的一个重要分支,有着悠久的历史,但传统方法一般适用于状态和动作空间维度较低的场景.随着深度学习方法的蓬勃发展,该领域正在经历着快速的变革,许多以前无法解决的问题逐步变得易于用多智能体深度强化学习方法处理.尽管如此,MADRL 还是一个年轻的领域,在引起学者更多的兴趣的同时,也面临着非平稳性、维度诅咒和信用分配等诸多挑战.

在本文中,我们提供了在新兴的多智能体深度强化学习领域的最新工作的广泛概述.我们首先分析了多智能体深度强化学习常用的问题表示以及面临的困难和挑战.然后我们将最近的研究按照 3 个不同的维度,即(基于价值或基于策略,任务类型,是否通信)分为八个不同的子主题,并详细研究和分析了每个主题的工作.在此部分,我们还重点总结了通信学习的 MADRL 方法和基于图神经网络的 MADRL 方法.

此外,我们系统地研究了 MADRL 在现实场景中的应用,并综述了最新的研究进展.最后,我们总结本文未重点关注但是值得研究的一些方向.有了这篇文章,我们希望通过提供对最新方法的更全面的概述,为感兴趣的读者提供必要的工具,以系统了解 MADRL 中当前的挑战和主流工作.在可预见的未来,我们预计会有大量的新工作涌现,因此我们希望鼓励学者们在这个有趣而年轻的研究领域进一步研究与发展.

本文试图从全面广泛的视角涵盖主流方向的方法,于是对于 MADRL 中各个子方向的方法,综述的不够细致.目前有一些关于某一类方向的综述工作,读者可以参考以获取更细致的分析和总结.如基于通信的 MADRL 方法可参考文献[107]和[108],基于值函数分解的方法可参考文献[109].本文的分类方式基于几种常用的 MADRL 问题表示类型以及无模型的 MADRL 方法,所以有许多未关注的方向,下面对这些方向进行简要的分析.

5.1　扩展式博弈问题

前文总结了几种常用的 MADRL 问题表示形式,都是建立在智能体是同时进行决策的基础上.而当智能体按顺序轮流采取决策时,一般被建模为扩展式博弈(extensive-form game)[20].扩展式博弈以博弈树的形式指定智能体之间的顺序交互.博弈树显示了智能体在每个时间点的移动顺序和可能的行动.当智能体具有不完全信息或对全局状态的部分视图时,可以将其形式化为决策节点被划分为信息集的不完全信息扩展式博弈.当博弈到达信息集时,轮到它的智能体无法区分信息集中的节点,也无法分辨到达了树中的哪个节点.学者们研究试图通过近似纳什均衡来解决扩展式博弈问题,主要在扑克领域[90,110].

5.2　基于模型的 MADRL 方法

前文介绍的 MADRL 方法都是无模型的,从某种意义上说,智能体不需要知道环境是如何工作的,它可以通过纯粹与环境交互来学习如何做出最佳行为.在单智能体 DRL 领域,基于模型的方法已经被广泛研究,其中学习智能体将首先建立一个明确的状态空间"模型",以了解环境如何在状态转移动力学和奖励函数方面工作,然后从"模型"中学习.基于模型的算法的好处在于,它们通常需要的环境数据样本要少得多.近 20 年前,最早的基于模型的 MARL 方法被提出,如著名的 R-MAX 算法[111].令人惊讶的是,基于模型的线程的开发从此停止了.鉴于基于模型的方法在单智能体 DRL 任务上展示的令人印象深刻的结果[112,113],基于模型的 MADRL 方法值得更多的关注.

5.3　安全和鲁棒 MADRL

尽管 DRL 为最佳决策制定提供了一个通用框架,但当强化学习模型真正部署到现实环境中时,它必须包含某些类型的约束.学者们认为首先考虑具有鲁棒性和安全性约束的 MADRL 是至关重要的,一个直接的例子就是自动驾驶.鲁棒性通常指方法可以在与训练环境不同的设置中泛化并保持鲁棒性能的属性[114].安全性是指即使在训练期间,方法也只能在预定义的安全区域内以最小的违规次数运行.事实上,安全和鲁棒强化学习仍处于发展理论框架的早期阶段,以涵盖单智能体设置中的稳健或安全约束.在多智能体设置中,问题只会变得更具挑战性,因为解决方案现在需要考虑智能体之间的相互作用,特别是那些有利益冲突的智能体[115].除此之外,还应该考虑对环境动态不确定性的鲁棒性[116].目前已有部分工作在关注安全和鲁棒性在 MADRL 设置中的实现[117,118],这也是一个有趣的值得探索的方向.

参　考　文　献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature,

2015, 521(7553): 436-444

[2] Zhang S, Gong Y, Wang J. The development of deep convolutional neural networks and its application on computer vision. Chinese Journal of Computers, 2019, 42(3): 453-482 (in Chinese)
(张顺,龚怡宏,王进军.深度卷积神经网络的发展及其在计算机视觉领域的应用.计算机学报, 2019, 42(3): 453-482)

[3] Jiang J, Li Z, Liu X. Deep learning based monocular depth estimation: a survey. Chinese Journal of Computers, 2022, 45(6): 1276-1307(in Chinese)
(江俊君,李震宇,刘贤明.基于深度学习的单目深度估计方法综述.计算机学报, 2022, 45(6): 1276-1307)

[4] Li Y, Gao Y, Yan J, et al. Image inpainting methods based on deep neural networks: A review. Chinese Journal of Computers, 2021, 44(11): 2295-2316(in Chinese)
(李月龙,高云,闫家良,等.基于深度神经网络的图像缺损修复方法综述.计算机学报, 2021, 44(11): 2295-2316)

[5] Malik M, Malik M K, Mehmood K, et al. Automatic speech recognition: a survey. Multimedia Tools and Applications, 2021, 80: 9411-9457

[6] Lee W, Seong J J, Ozlu B, et al. Biosignal sensors and deep learning-based speech recognition: A review. Sensors, 2021, 21(4): 1399

[7] Minaee S, Abdolrashidi A, Su H, et al. Biometrics recognition using deep learning: A survey. Artificial Intelligence Review, 2023, 56(8): 8647-8695

[8] Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 604-624

[9] Pandey B, Pandey D K, Mishra B P, et al. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. Journal of King Saud University-Computer and Information Sciences, 2022, 34(8): 5083-5099

[10] Chai J, Li A. Deep learning in natural language processing: A state-of-the-art survey // Proceedings of the 11st International Conference on Machine Learning and Cybernetics. Kobe, Japan, 2019: 1-6

[11] Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv preprint arXiv:2011.00583, 2020.

[12] Wang X, Wang S, Liang X, et al. Deep reinforcement learning: a survey. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5064-5078

[13] Sallab A E L, Abdou M, Perot E, et al. Deep reinforcement learning framework for autonomous driving. arXiv preprint arXiv:1704.02532, 2017.

[14] Munemasa I, Tomomatsu Y, Hayashi K, et al. Deep reinforcement learning for recommender systems // Proceedings of 2018 International Conference on Information and Communications Technology (ICOIACT). Yogyakarta, Indonesia, 2018: 226-233

[15] Afsar M M, Crump T, Far B. Reinforcement learning based recommender systems: A survey. ACM Computing Surveys, 2022, 55(7): 1-38

[16] Zhang K, Yang Z, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of Reinforcement Learning and Control, 2021, 325: 321-384

[17] Buşoniu L, Babuška R, De Schutter B. Multi-agent reinforcement learning: An overview. Innovations in Multi-agent Systems and Applications-1, 2010, 310: 183-221

[18] Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction // Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, UK, 2019: 1-13

[19] Yang Q, Steinfeld A, Rosé C, et al. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design // Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu, USA, 2020: 1-13

[20] Wong A, Bäck T, Kononova A V, et al. Deep multiagent reinforcement learning: Challenges and directions. Artificial Intelligence Review, 2022, 56: 5023-5056

[21] Du W, Ding S. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. Artificial Intelligence Review, 2021, 54: 3215-3238

[22] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 2022, 55: 895-943

[23] Hernandez-Leal P, Kartal B, Taylor M E. A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems, 2019, 33(6): 750-797.

[24] Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839.

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

[26] Hado V H, Guez A, Silver D. Deep reinforcement learning with double Q-learning // Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, Arizona, 2016, 2094-2100

[27] Wang Z Y, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning // Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016, 48: 1995-2003

[28] Hessel M, Modayil M, Hasselt V H, et al. Rainbow: combining improvements in deep reinforcement learning // Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Orleans, USA, 2018: 3215-3222

[29] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning // Proceedings of the 4th In-

ternational Conference on Learning Representations. San Juan Puerto Rico, 2016: 1-14

[30] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms// Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014: 387-395

[31] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward// Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Stockholm, Sweden, 2018: 2085-2087

[32] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning. The Journal of Machine Learning Research, 2020, 21(1): 7234-7284

[33] Rashid T, Farquhar G, Peng B, et al. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning // Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 10199-10210

[34] Wang J, Ren Z, Liu T, et al. QPLEX: Duplex dueling multi-agent Q-learning// Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-27

[35] Son K, Kim D, Kang W J, 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, USA, 2019, 97: 5887-5896

[36] Yang Y, Wen Y, Wang J, et al. Multi-agent determinantal Q-learning// Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 10757-10766

[37] Gupta T, Mahajan A, Peng B, et al. Uneven: Universal value exploration for multi-agent reinforcement learning // Proceedings of the 38th International Conference on Machine Learning. Virtual Event, 2021: 3930-3941.

[38] Yang Y, Hao J, Chen G, et al. Q-value path decomposition for deep multiagent reinforcement learning // Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 10706-10715.

[39] Iqbal S, De Witt C A S, Peng B, et al. Randomized entity-wise factorization for multi-agent reinforcement learning // Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 2021: 4596-4606.

[40] Zhou H, Lan T, Aggarwal V. PAC: Assisted value factorization with counterfactual predictions in multi-agent reinforcement learning // Proceedings of the 35th Conference on Neural Information Processing Systems. New Orleans, USA, 2022.

[41] Shen S, Qiu M, Liu J, et al. ResQ: A residual Q function-based approach for multi-agent reinforcement learning value factorization// Proceedings of the 35th Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 5471-5483.

[42] Hong Y, Jin Y, Tang Y. Rethinking individual global max in cooperative multi-agent reinforcement learning // Proceedings of the 35th Conference on Neural Information Processing Systems. New Orleans, USA, 2022.

[43] Yang Y, Luo R, Li M, et al. Mean field multi-agent reinforcement learning // Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 5571-5580

[44] Subramanian S G, Poupart P, Taylor M E, et al. Multi type mean field reinforcement learning. arXiv preprint arXiv: 2002.02513, 2020.

[45] Zhang T, Ye Q, Bian J, et al. MFVFD: A multi-agent Q-learning approach to cooperative and non-cooperative tasks// Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada, 2021: 500-506

[46] Subramanian S G, Taylor M E, Crowley M, et al. Decentralized mean field games// Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 9439-9447

[47] Ding S, Du W, Ding L, et al. Multi-agent dueling Q-learning with mean field and value decomposition. Pattern Recognition, 2023, 139: 109436

[48] Yang M, Liu G, Zhou Z. Partially observable mean field multi-agent reinforcement learning based on graph-attention. arXiv preprint arXiv:2304.12653, 2023

[49] Foerster J, Assael I A, De Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning// Proceedings of the 29th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2145-2153

[50] Zhang S Q, Zhang Q, Lin J. Efficient communication in multi-agent reinforcement learning via variance based control // Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 3235-3244

[51] Zhang S Q, Zhang Q, Lin J. Succinct and robust multi-agent communication with temporal message control // Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 17271-17282.

[52] Du W, Ding S, Zhang C, et al. Multiagent reinforcement learning with heterogeneous graph attention network. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10): 6851-6860

[53] Wang T, Wang J, Zheng C, et al. Learning nearly decomposable value functions via communication minimization // Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019.

[54] Yuan L, Wang J, Zhang F, et al. Multi-agent incentive communication via decentralized teammate modeling// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 9466-9474

[55] Jiang H, Shi D, Xue C, et al. Multi-agent deep reinforce-

ment learning with type-based hierarchical group communication. Applied Intelligence, 2021, 51: 5793-5808.

[56] Guan C, Chen F, Yuan L, et al. Efficient multi-agent communication via self-supervised information aggregation // Proceedings of the 35th Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 1020-1033

[57] Shao J, Lou Z, Zhang H, et al. Self-organized group for cooperative multi-agent reinforcement learning // Proceedings of the 35th Conference on Neural Information Processing Systems. New Orleans, USA, 2022: 5711-5723

[58] Jiang J, Dun C, Huang T, et al. Graph convolutional reinforcement learning. arXiv preprint arXiv:1810.09202, 2018

[59] Ding S, Du W, Ding L, et al. Multiagent reinforcement learning with graphical mutual information maximization. IEEE Transactions on Neural Networks and Learning Systems, 2023, doi: 10.1109/TNNLS.2023.3243557.

[60] Sheng J, Wang X, Jin B, et al. Learning structured communication for multi-agent reinforcement learning // Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. London, UK, 2023: 436-438

[61] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients // Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2974-2982

[62] Foerster J, Song F, Hughes E, et al. Bayesian action decoder for deep multi-agent reinforcement learning // Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 1942-1951

[63] Hu H, Foerster J N. Simplified action decoder for deep multi-agent reinforcement learning. arXiv preprint arXiv: 1912.02288, 2019

[64] Muglich D, Zintgraf L M, De Witt C A S, et al. Generalized beliefs for cooperative AI // Proceedings of the International Conference on Machine Learning. Maryland, USA, 2022: 16062-16082.

[65] Cao J, Yuan L, Wang J, et al. Linda: Multi-agent local information decomposition for awareness of teammates. Science China Information Sciences, 2023, 66(8): 182101

[66] Roy J, Barde P, Harvey F, et al. Promoting coordination through policy regularization in multi-agent deep reinforcement learning // Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 15774-15785

[67] Zhou M, Liu Z, Sui P, et al. Learning implicit credit assignment for cooperative multi-agent reinforcement learning // Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 11853-11864

[68] Peng B, Rashid T, Schroeder de Witt C, et al. Facmac: Factored multi-agent centralised policy gradients // Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 12208-12221

[69] Wang Y, Han B, Wang T, et al. DOP: Off-policy multi-agent decomposed policy gradients // Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-21

[70] Li Y, Xie G, Lu Z. Difference advantage estimation for multi-agent policy gradients // Proceedings of the International Conference on Machine Learning, Baltimore, USA, 2022: 13066-13085

[71] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments // Proceedings of the 31st International Conference on Neural Information Processing Systems. California, USA, 2017: 6382-6393

[72] Li S, Wu Y, Cui X, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient // Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 4213-4220

[73] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning // Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 2961-2970

[74] Jiang J, Lu Z. The emergence of individuality // Proceedings of the 38th International Conference on Machine Learning. Virtual Event, 2021: 4992-5001

[75] Wang J, Ye D, Lu Z. More centralized training, still decentralized execution: multi-agent conditional policy factorization. arXiv preprint arXiv:2209.12681, 2022.

[76] Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation // Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2252-2260

[77] Peng P, Wen Y, Yang Y, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. arXiv preprint arXiv:1703.10069, 2017

[78] Jiang J, Lu Z. Learning attentional communication for multi-agent cooperation // Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 7265-7275

[79] Kim D, Moon S, Hostallero D, et al. Learning to schedule communication in multi-agent reinforcement learning. arXiv preprint arXiv:1902.01554, 2019.

[80] Wang R, He X, Yu R, et al. Learning efficient multi-agent communication: An information bottleneck approach // Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 9908-9918

[81] Liu Y, Wang W, Hu Y, et al. Multi-agent game abstraction via graph attention neural network // Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 7211-7218

[82] Ding Z, Huang T, Lu Z. Learning individually inferred communication for multi-agent cooperation // Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 22069-22079

[83] Kim W, Park J, Sung Y. Communication in multi-agent reinforcement learning: Intention sharing // Proceedings of the 9th International Conference on Learning Representations. Virtual Event, 2021: 1-15

[84] Du Y, Liu B, Moens V, et al. Learning correlated communication topology in multi-agent reinforcement learning // Proceedings of the 20th International Conferences on Autonomous Agents and Multiagent Systems. London, UK, 2021: 456-464

[85] Seraj E, Wang Z, Paleja R, et al. Learning efficient diverse communication for cooperative heterogeneous teaming // Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. London, UK, 2022: 1173-1182

[86] Das A, Gervet T, Romoff J, et al. Tarmac: Targeted multi-agent communication // Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 1538-1546

[87] Singh A, Jain T, Sukhbaatar S. Learning when to communicate at scale in multiagent cooperative and competitive tasks. arXiv preprint arXiv:1812.09755, 2018

[88] Niu Y, Paleja R R, Gombolay M C. Multi-agent graph-attention communication and teaming // Proceedings of the 20th International Conferences on Autonomous Agents and Multiagent Systems. London, UK, 2021: 964-973

[89] Jaderberg M, Czarnecki W M, Dunning I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science, 2019, 364(6443): 859-865

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

[91] Calvo J A, Dusparic I. Heterogeneous multi-agent deep reinforcement learning for traffic lights control // Proceedings of the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. Dublin, Ireland, 2018: 2-13

[92] Chen X, Xiong G, Lv Y, et al. A collaborative communication-Qmix approach for large-scale networked traffic signal control // Proceedings of the 24th IEEE International Intelligent Transportation Systems Conference. Indianapolis, USA, 2021: 3450-3455

[93] Chu T, Wang J, Codecà L, et al. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(3): 1086-1095

[94] Devailly F X, Larocque D, Charlin L. IG-RL: Inductive graph reinforcement learning for massive-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 7496-7507

[95] Huang H, Hu Z, Lu Z, et al. Network-scale traffic signal control via multiagent reinforcement learning with deep spatiotemporal attentive network. IEEE Transactions on Cybernetics, 2023, 53(1): 262-274

[96] Yang S, Yang B, Kang Z, et al. IHG-MA: Inductive heterogeneous graph multi-agent reinforcement learning for multi-intersection traffic signal control. Neural Networks, 2021, 139: 265-277

[97] Yang S, Yang B. An inductive heterogeneous graph attention-based multi-agent deep graph infomax algorithm for adaptive traffic signal control. Information Fusion, 2022, 88: 249-262

[98] Zhou W, Chen D, Yan J, et al. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. Autonomous Intelligent Systems, 2022, 2(1): 5

[99] Han S, Wang H, Su S, et al. Stable and efficient shapley value-based reward reallocation for multi-agent reinforcement learning of autonomous vehicles // Proceedings of the 39th IEEE Conference on Robotics and Automation. Philadelphia, USA, 2022: 8765-8771

[100] Palanisamy P. Multi-agent connected autonomous driving using deep reinforcement learning // Proceedings of the 30th International Joint Conference on Neural Networks. Glasgow, UK, 2020: 1-7

[101] Zhou M, Luo J, Villella J, et al. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving. arXiv preprint arXiv:2010.09776, 2020

[102] He X, An B, Li Y, et al. Learning to collaborate in multi-module recommendation via multi-agent reinforcement learning without communication // Proceedings of the 14th ACM Conference on Recommender Systems. Virtual Event, 2020: 210-219

[103] Zhang W, Liu H, Wang F, et al. Intelligent electric vehicle charging recommendation based on multi-agent reinforcement learning // Proceedings of the 30th Web Conference. Ljubljana, Slovenia, 2021: 1856-1867

[104] Zhang Y, Zhang C, Liu X. Dynamic scholarly collaborator recommendation via competitive multi-agent reinforcement learning // Proceedings of the 11th ACM Conference on Recommender Systems. Como, Italy, 2017: 331-335

[105] Gui T, Liu P, Zhang Q, et al. Mention recommendation in Twitter with cooperative multi-agent reinforcement learning // Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris, France, 2019: 535-544

[106] Wang E, Ding R, Yang Z, et al. Joint charging and relocation recommendation for e-taxi drivers via multi-agent mean field hierarchical reinforcement learning. IEEE Transactions on Mobile Computing, 2020, 21(4): 1274-1290

[107] Zhu C, Dastani M, Wang S. A survey of multi-agent reinforcement learning with communication. arXiv preprint arXiv:2203.08975, 2022

[108] Wang H, Yu Y, Jiang Y. Review of the progress of communication-based multi-agent reinforcement learning. SCIENTIA SINICA Informationis, 2022, 52(5): 742-764 (in Chinese)

(王涵，俞扬，姜远. 基于通信的多智能体强化学习进展综述. 中国科学:信息科学，2022，52(05)：742-764)

[109] Xiong L，Cao L，Lai J. Overview of multi-agent deep reinforcement learning based on value factorization. Computer Science，2022，49(9)：172-182(in Chinese)
(熊丽琴，曹雷，赖俊，等. 基于值分解的多智能体深度强化学习综述. 计算机科学，2022，49(9)：172-182)

[110] Brown N，Sandholm T. Superhuman AI for heads-up no-limit poker：Libratus beats top professionals. Science，2018，359(6374)：418-424

[111] Brafman R I，Tennenholtz M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. Journal of Machine Learning Research，2002，3(10)：213-231

[112] Hafner D，Lillicrap T，Ba J，et al. Dream to control：Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603，2019

[113] Schrittwieser J，Antonoglou I，Hubert T，et al. Mastering atari，go，chess and shogi by planning with a learned model. Nature，2020，588：604-609

[114] Abdullah M A，Ren H，Ammar H B，et al. Wasserstein robust reinforcement learning. arXiv preprint arXiv：1907.13196，2019

[115] Guo J，Chen Y，Hao Y，et al. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans，USA，2022：114-121

[116] Zhang K，Sun T，Tao Y，et al. Robust multi-agent reinforcement learning with model uncertainty // Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver，Canada，2020：10571-10583

[117] ElSayed-Aly I，Bharadwaj S，Amato C，et al. Safe multi-agent reinforcement learning via shielding // Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. Virtual Event，UK，2021：483-491

[118] Lu S，Zhang K，Chen T，et al. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning // Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event，2021：8767-8775

DING Shi-Fei，Ph. D.，professor. His research interests include artificial intelligence，pattern recognition，machine learning，data mining.

DU Wei，Ph. D. candidate. His research interests include deep learning，reinforcement learning.

ZHANG Jian，Ph. D.，lecturer. His research interests include machine，pattern recognition.

GUO Li-Li，Ph. D.，lecturer. Her research interests include deep learning，multimodal emotional computing.

DING Ling，Ph. D. candidate. Her research interests include graph machine learning，clustering analysis.

Background

Multi-agent reinforcement learning，as an important branch of machine learning，has a long history，but traditional methods are generally suitable for scenes with low dimensions of state and action space. With the rapid development of deep learning methods，the field is undergoing rapid change，and many previously unsolvable problems are gradually becoming easier to handle with multi-agent deep reinforcement learning (MADRL) methods. Nevertheless，MADRL is still a young field and faces many challenges while attracting more interest from scholars.

There are several common problems with MADRL methods：1) Scalability：With the increase of the number of agents，the action space increases exponentially，making it difficult to obtain optimal policy. 2) Non-stationary：the behavior of each agent usually causes changes in the environment，making the environment unstable，and then affecting the action choice and policy choice of other agents. 3)Reward problem：Because the tasks and learning goals of different agents are usually different in general tasks，and the interaction between agents will make it difficult to determine the goal reward，which will seriously affect the convergence of the method. In the cooperative task，the agents have the same goal，but there is a credit assignment problem，that is，how to assign reward signals correctly for each agent to better coordinate and maximize the overall return.

In this paper，we provide a broad overview of the latest work in the emerging field of multi-agent deep reinforcement learning. We first analyze the common problem representation and the difficulties and challenges faced by multi-agent deep reinforcement learning. We then divide recent research into eight different sub-topics along 3 different dimensions，namely (value-based or policy-based，type of task，whether to communicate)，and examine and analyze the work of each topic in detail. In this part，we also summarize the MADRL method of communication learning

and the MADRL method based on graph neural network. In addition, we systematically study the application of MADRL in real world scenarios, and review the latest research progress. Finally, we summarize some directions that are not focused in this paper but are worth studying. With this article, we hope to provide interested readers with the necessary tools to systematically understand current challenges and mainstream work in MADRL by providing a more comprehensive overview of the latest approaches.

This work is supported by the National Natural Science Foundation of China under Grant No. 62276265, No. 61976216, No. 62206297.