多智能体深度强化学习研究综述

孙 彧[1,2],曹 雷[1],陈希亮[1],徐志雄[1],赖 俊[1]
1.陆军工程大学 指挥控制工程学院,南京 210007
2.中国人民解放军31102部队

摘 要:多智能体深度强化学习是机器学习领域的一个新兴的研究热点和应用方向,涵盖众多算法、规则、框架,并广泛应用于自动驾驶、能源分配、编队控制、航迹规划、路由规划、社会难题等现实领域,具有极高的研究价值和意义。对多智能体深度强化学习的基本理论、发展历程进行简要的概念介绍;按照无关联型、通信规则型、互相合作型和建模学习型4种分类方式阐述了现有的经典算法;对多智能体深度强化学习算法的实际应用进行了综述,并简单罗列了多智能体深度强化学习的现有测试平台;总结了多智能体深度强化学习在理论、算法和应用方面面临的挑战和未来的发展方向。

关键词:强化学习;深度学习;多智能体系统;多智能体深度强化学习
文献标志码:A **中图分类号:**TP181 doi:10.3778/j.issn.1002-8331.1912-0100

孙彧,曹雷,陈希亮,等.多智能体深度强化学习研究综述.计算机工程与应用,2020,56(5):13-24.
SUN Yu, CAO Lei, CHEN Xiliang, et al. Overview of multi-agent deep reinforcement learning. Computer Engineering and Applications, 2020, 56(5):13-24.

Overview of Multi-Agent Deep Reinforcement Learning

SUN Yu[1,2], CAO Lei[1], CHEN Xiliang[1], XU Zhixiong[1], LAI Jun[1]
1.College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China
2.Unit 31102 of PLA, China

Abstract: Multi-agent deep reinforcement learning is an emerging research hotspot and application direction in the field of machine learning and artificial intelligence. It covers many algorithms, rules, and frameworks, and is widely used in autonomous driving, energy allocation, formation control, trajectory planning,routing planning and social dilemma, it has extremely high research value and significance. The paper first briefly introduces the basic theory and development history of multi-agent deep reinforcement learning, then elaborates the existing classic algorithms according to four classification: non-association type, communication rule based type, mutual cooperation type and modeling learning type, then summarizes the practical application of multi-agent deep reinforcement learning and briefly lists the existing test platforms. The paper finally summarizes the challenges and future directions in theory, algorithms and applications of multi-agent deep reinforcement learning.

Key words: reinforcement learning; deep learning; multi-agent system; multi-agent deep reinforcement learning

1 引言

多智能体系统(Multi-Agent System,MAS)[1]是在同一个环境中由多个交互智能体组成的系统,该系统常用于解决独立智能体以及单层系统难以解决的问题,其中的智能可以由方法、函数、过程,算法或强化学习来实现[2]。多智能体系统因其较强的实用性和扩展性,在机

基金项目:国家自然科学基金(No.61806221);国防科技重点实验室基金(No.6142101180304);国防科技创新特区163计划资助项目;装备发展部"十三五"全军共用信息系统装备预研项目。
作者简介:孙彧(1993—),男,硕士研究生,研究领域为深度强化学习,智能化指挥控制;曹雷(1965—),通信作者,男,教授,研究领域为指挥信息系统工程,决策理论与方法,E-mail:caolei.nj@foxmail.com;陈希亮(1985—),男,博士,副教授,研究领域为深度强化学习,指挥信息系统工程。
收稿日期:2019-12-05 **修回日期:**2020-01-03 **文章编号:**1002-8331(2020)05-0013-12
CNKI**网络出版:**2020-02-14, http://kns.cnki.net/kcms/detail/11.2127.TP.20200214.1008.002.html

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

器人合作、分布式控制[3]、资源管理、协同决策支持系统、自主化作战系统、数据挖掘等领域都得到了广泛的应用。

强化学习(Reinforcement Learning,RL)[4]是机器学习的一个重要分支,其本质是描述和解决智能体在与环境的交互过程中学习策略以最大化回报或实现特定目标的问题。与监督学习不同,强化学习并不告诉智能体如何产生正确的动作,它只对动作的好坏做出评价并根据反馈信号修正动作选择和策略,所以强化学习的回报函数所需的信息量更少,也更容易设计,适合解决较为复杂的决策问题。近来,随着深度学习(Deep Learning,DL)[5]技术的兴起及其在诸多领域取得辉煌的成就,融合深度神经网络和RL的深度强化学习(Deep Reinforcement Learning,DRL)[6]成为各方研究的热点,并在计算机视觉、机器人控制、大型即时战略游戏等领域取得了较大的突破。

DRL的巨大成功促使研究人员将目光转向多智能体领域,他们大胆地尝试将DRL方法融入到MAS中,意图完成多智能体环境中的众多复杂任务,这就催生了多智能体深度强化学习(Multi-agent Deep Reinforcement Learning,MDRL)[7],经过数年的发展创新,MDRL诞生了众多算法、规则、框架,并已广泛应用于各类现实领域。从单到多、从简单到复杂、从低维到高维的发展脉络表明,MDRL正逐渐成为机器学习乃至人工智能领域最火热的研究和应用方向,具有极高的研究价值和意义。

2　多智能体深度强化学习基本理论

2.1　单智能体强化学习

单智能体强化学习(Single Agent Reinforcement Learning,SARL)中智能体与环境的交互遵循马尔可夫决策过程(Markov Decision Process,MDP)[8]。图1表示单智能体强化学习的基本框架。

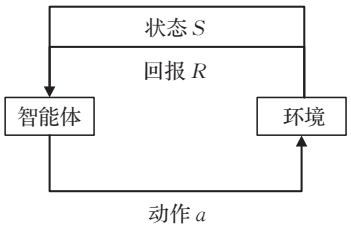

图1　单智能体强化学习基本框架

MDP一般由多元组$\langle S,A,R,f,\gamma\rangle$表示,其中$S$和$A$分别代表智能体的状态和动作空间,智能体的状态转移函数可表示为:

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

它决定了在给定动作$a\in A$的情况下,由状态$s\in S$转移到下一个状态$s'\in S$的概率分布,回报函数为:

$$R{:}S\times A\times S\to \mathrm{R} \tag{2}$$

其定义了智能体通过动作a从状态s转移到状态s'所得到的环境瞬时回报。从开始时刻t到T时刻交互结束时,环境的总回报可表示为:

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \tag{3}$$

其中$\gamma\in[0,1]$为折扣系数,它用于平衡智能体的瞬时回报和长期回报对总回报的影响。智能体的学习策略可表示为状态到动作的映射$\pi{:}S\to A$,MDP的求解目标是找到期望回报值最大的最优策略π^*,一般用最优状态动作值函数(Q函数)形式化表征期望回报:

$$Q^*(s,a)=\max_{\pi}E\left[R_t|s_t=s,a_t=a,\pi\right] \tag{4}$$

其遵循最优贝尔曼方程(Bellman Equation):

$$Q^*(s,a)=E_{s'\sim S}\left[r+\gamma\max_{a'}Q(s',a')|s,a\right] \tag{5}$$

几乎所有强化学习的方法都采用迭代贝尔曼方程[9]的形式求解Q函数,随着迭代次数不断增加,Q函数最终得以收敛,进而得到最优策略:

$$\pi^*=\arg\max Q^*(s,a) \tag{6}$$

Q学习(Q-Learning)[10]是最经典的RL算法,它使用表格存储智能体的Q值,其Q表的更新方式如下所示:

$$Q(s,a)\leftarrow Q(s,a)+\alpha\left[r+\gamma\max_{a'}Q(s',a')-Q(s,a)\right] \tag{7}$$

算法通过不断迭代更新Q函数的方式求得最优解。

与上述基于值函数(Value Based,VB)的RL方法不同,基于策略梯度(Policy Gradient,PG)[11]的方法用参数化的策略θ代替Q函数,并利用梯度下降的方法逼近求解最优策略,该类方法可以用来求解连续动作空间的问题,其代表性算法有REINFORCE[12]、PG[11]、DPG[13]等。

2.2　深度强化学习

传统RL方法有较多局限性,如学习速率慢、泛化性差、需要手动对状态特征进行建模、无法应对高维空间等。为了解决此类问题,研究人员利用深度神经网络对Q函数和策略进行近似,这就是深度强化学习方法,DRL不仅让智能体能够面对高维的状态空间,而且解决了状态特征难以建模的问题,下面简要介绍DRL及其典型算法。

2.2.1　基于值函数的方法

深度Q网络(Deep Q-Network,DQN)结合了深度神经网络和传统RL算法Q-Learning的优点,它使用神经网络对值函数进行近似,与Q学习等传统RL算法不同,DQN放弃了以表格形式记录智能体Q值的方式,而采用经验库(Experience Replay Buffer)[14]将环境探索得到的数据以记忆单元$\langle s,a,r,s'\rangle$的形式储存起来,然后利用随机小样本采样的方法更新和训练神经网络参数。另外DQN还引入双网络结构(Fixed Q-targets),即同时使用Q网络和目标网络训练模型,其中Q网络参

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

数 θ 随训练过程实时更新，而目标网络的参数 θ^- 是每经过一定次数迭代后 Q 网络参数的复制值，DQN 在每轮迭代 i 中的目标为最小化 Q 网络及其目标网络之间的损失函数。

$$L_i(\theta_i)=E_{s,a,r,s'}\left[\left(r+\gamma\max_{a'}Q(s',a';\theta_i^-)-Q(s,a;\theta_i)\right)^2\right] \tag{8}$$

在经验库机制和双网络结构的共同作用下，DQN 有效解决了数据高相关性的问题，提升了神经网络更新效率和算法收敛效果，在实际应用中，DQN 能够在多种策略游戏中战胜高水平人类玩家。研究人员围绕 DQN 在多个方面也进行了改进和拓展，如文献[15]采用双函数近似解决了过估计问题；文献[16]利用优势函数(Advantage Function)将 Q 函数进行分解和整合，提升了动作输出的确定性；文献[17]使用循环神经网络(Recurrent Neural Network，RNN)和长短时记忆单元(Long Short Temporal Memory，LSTM)代替传统的神经网络，强化了算法应对不同环境的鲁棒性；文献[18]则优化了 DQN 的经验库机制，提高了算法训练的效率和效果。

2.2.2　基于策略梯度的方法

与以 DQN 为代表的 VB 方法相比，PG 方法具有能够胜任连续且高维的动作空间的优点。其代表算法为深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)[19]。DDPG 基于演员评论家(Actor-Critic，AC)框架[20]；在输入方面，其通过在 Actor 网络引入随机噪声的方式产生探索策略；在动作输出方面采用神经网络来拟合策略函数，并直接输出动作以应对连续动作空间；在参数更新方面，与 DQN 中直接参数复制的方法不同，该算法采用缓慢更新参数的方法提升稳定性；DDPG 还引入了批正则化(Batch Normalization)方法保证其对多种任务的泛化能力。除了 DDPG 外，AC 框架与 PG 方法相融合衍生出多种 DRL 算法，如使用多 CPU 线程进行分布式学习的异步优势演员评论家(Asynchronous Advantage Actor-Critic，A3C)算法[21]；增强策略梯度稳定性的信赖域策略优化(Trust Region Policy Optimization，TRPO)[22]和近端策略优化(Proximal Policy Optimization，PPO)算法[23]等。

DRL 的成功表明，RL 和神经网络的融合在单智能体领域已较为普遍，并产生了大量成熟的算法，这为 MDRL 的突破指明了方向并提供了开阔的思路。

2.3　多智能体强化学习

与单智能体 RL 不同，多智能体强化学习(Multi-Agent Reinforcement Learning，MARL)遵循随机博弈(Stochastic Game，SG)[24]过程。图 2 描述了多智能体强化学习的基本框架。

SG 可由多元组 $\langle S,A_1,A_2,\cdots,A_n,R_1,R_2,\cdots,R_n,f,\gamma\rangle$ 表示，其中 n 为环境中智能体的数量，S 为环境的

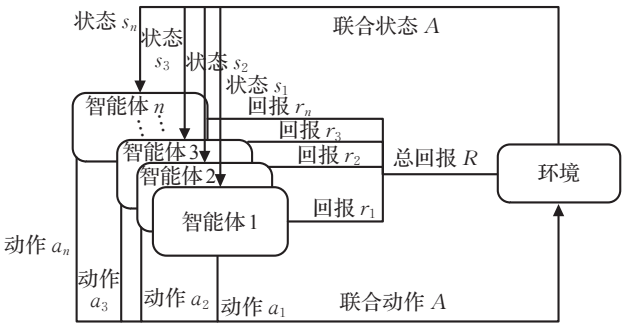

图2　多智能体强化学习基本框架

状态空间，$A_i(i=1,2,\cdots,n)$ 为每个智能体的动作空间，$A=A_1\times A_2\times\cdots\times A_n$ 为所有智能体的联合动作空间，联合状态转移函数可表示为：

$$f:S\times A\times S\to[0,1] \tag{9}$$

它决定了在执行联合动作 $a\in A$ 的情况下，由状态 $s\in S$ 转移到下一个状态 $s\in S'$ 的概率分布，每个智能体的回报函数可表示为：

$$R_i:S\times A\times S\to\mathrm{R},\quad i=1,2,\cdots,n \tag{10}$$

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：

$$a_k=\left[a_{1,k}^{\mathrm{T}},a_{2,k}^{\mathrm{T}},\cdots,a_{n,k}^{\mathrm{T}}\right]^{\mathrm{T}},\quad a_k\in A,\ a_{i,k}\in A_i \tag{11}$$

每个智能体的个体策略为：

$$\pi_i:S\times A_i\to[0,1] \tag{12}$$

它们共同构成联合策略 π。由于智能体的回报 $r_{i,k+1}$ 取决于联合动作，所以总回报取决于联合策略：

$$R_i^{\pi}(x)=E\left\{\sum_{k=0}^{\infty}\gamma^k r_{i,k+1}\middle|s_0=s,f\right\} \tag{13}$$

每个智能体的 Q 函数则取决于联合动作 $Q_i^{\pi}:S\times A\to\mathrm{R}$，求解方式为：

$$Q_i^{\pi}(s,a)=E\left\{\sum_{k=0}^{\infty}\gamma^k r_{i,k+1}\middle|s_0=s,a_0=a,f\right\} \tag{14}$$

MARL 的算法根据其回报函数的不同可以分为完全合作型(Fully Cooperative)[25]、完全竞争型(Fully Competitive)[25]和混合型(Mixed)[25]三种任务类型，完全合作型算法中智能体的回报函数是相同的，即 $R_1=R_2=\cdots=R_n$，表示所有智能体都在为实现共同的目标而努力，其代表算法有团队 Q 学习(Team Q-learning)[26]、分布式 Q 学习(Distributed Q-learning)[27]等；完全竞争型算法中智能体的回报函数是相反的，环境通常存在两个完全敌对的智能体，它们遵循 SG 原则，即 $R_1=-R_2$，智能体的目标是最大化自身的回报，同时尽可能最小化对方回报，其代表算法为 Minimax-Q[28]；混合型任务中智能体的回报函数并无确定性正负关系，该模型适合自利型(Self-interested)智能体，一般来说此类任务的求解大都与博弈论中均衡解的概念相关，即当环境中的一个状态存在多个均衡时，智能体需要一致选择同一个均

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

衡。该类算法主要面向静态任务,比较典型的有纳什 Q 学习(Nash Q-learning)[29]、相关 Q 学习(Correlated Q-learning)[30]、朋友或敌人 Q 学习(Friend or Foe Q-learning)[31]等。表1对多智能体强化学习的算法进行了简要汇总。

表1　多智能体强化学习算法汇总

分类	特点	算法名称
完全合作型	$R_1=R_2=\cdots=R_n$	Team Q-learning、Distributed Q-learning等
完全竞争型	$R_1=-R_2(n=2)$	Minimax-Q
混合型	回报值互不相关	Nash Q-learning、Correlated Q-learning、Friend or Foe Q-learning等

总的来看,传统MARL方法有很多优点,如合作型智能体间可以互相配合完成高复杂度的任务;多个智能体可以通过并行计算提升算法的效率;竞争型智能体间也可以通过博弈互相学习对手的策略,这都是SARL所不具备的。当然MARL也有较多缺陷,如RL固有的探索利用矛盾(Explore and Exploit)和维度灾难(Curse of Dimensionality);多智能体环境非平稳性(None-stationary)问题;多智能体信度分配(Multiagent Credit Assignment)[32]问题;最优均衡解问题;学习目标选择问题等。

3　多智能体深度强化学习及其经典方法

由于传统MARL方法存在诸多缺点和局限,其只适用于解决小型环境中的简单确定性问题,研究如何将深度神经网络和传统MARL相融合的MDRL方法具有很大的现实意义和迫切性。本章将分类介绍主流的MDRL方法并对每类方法的优缺点进行比较。按照智能体之间的通联方式,大致将当前的MDRL方法分为:无关联型、通信规则型、互相协作型和建模学习型4大类。

3.1　无关联型

此类方法并不从算法创新本身入手,而是将单智能体DRL算法直接扩展到多智能体环境中,每个智能体独立地与环境进行交互并自发地形成行为策略,互相之间不存在通信关联,其最初多用于测试单智能体DRL方法在多智能体环境中的适应性。

Tampuu[33]、Leibo[34]、Peysakhovich[35]等人最早将DQN算法分别应用到Atari乒乓球游戏等多种简单博弈场景中,他们在算法中引入了自博弈(Self-play)[36]机制和两套不同的回报函数以保证算法收敛,实验表明,DQN算法在这些简单多智能体场景中能够保证智能体之间的合作和竞争行为;Bansal等人[37]将PPO算法应用到竞争型多智能体模拟环境MuJoCo中,他们引入了探索回报(Exploration Rewards)[38]和对手采样(Opponent Sampling)[39]两种技术保证智能体形成自发性对抗策略,探索回报引导智能体在训练的前期学习到非对抗性的策略,以增加学习策略的维度;对手采样则引导智能体同时对新旧两种对手智能体进行采样,以增加学习策略的广度;Raghu等人[40]则尝试使用DQN、A3C、PPO等多种单智能体DRL算法解决了双人零和博弈问题,实验结果表明算法可以根据博弈问题的难易程度形成不同的行为策略;Gupta等人[41]将DQN、TRPO、DDPG等算法与循环神经网络相结合,应用到多智能体环境中,为了提升算法在多智能体环境中的可扩展性,他们引入了参数共享和课程学习机制,算法在多种场景中都取得了不错的效果。由于无关联型方法属于早期对多智能体学习环境的勇敢尝试,国内研究团队相对来说较为滞后,理论和实验贡献较为有限。表2总结分析了无关联型方法。

表2　无关联型方法总结分析

研究者	完成工作	性能特点
Tampuu/Leibo/Peysakhovich	将DQN等算法应用到简单游戏场景中	完成合作、竞争、混合型多种任务
Bansal等	将PPO算法应用到MuJoCo环境中	引入探索回报和对手采样机制,产生较优策略
Raghu等	使用DQN/A3C/PPO等算法解决零和博弈问题	可以形成多种不同策略
Gupta等	将DQN/TRPO/DDPG等算法应用到复杂多智能体环境中	引入参数共享和课程学习机制,结合RNN,具有较强可扩展性

无关联型方法较易实现,算法无需在智能体之间构建通信规则,每个智能体独立与环境交互并完成训练过程,该类方法能够有效地规避维度灾难带来的影响,且在可扩展性方面有先天性的优势。但它的局限性也十分明显,由于智能体之间互不通联,每个智能体将其他智能体看作环境的一部分,从个体的角度上看,环境是处在不断变化中的,这种环境非平稳性严重影响了学习策略的稳定和收敛,另外该类方法的学习效率和速率都十分低下。

3.2　通信规则型

此类方法在智能[illegible]建立显式的通信机制(如通信方式、通信时间、通信对象等),并在学习过程中逐渐确定和完善该通信机制,训练结束后,每个智能体需要根据其他智能体所传递的信息进行行为决策,此类方法多应用于完全合作型任务和非完全观测环境(详见4.2节)。

强化互学习(Reinforced Inter-Agent Learning,RIAL)[42]和差分互学习(Differentiable Inter-Agent Learning,DIAL)[42]是比较有代表性的通信规则型算法,它们遵循集中训练分散执行框架,都使用中心化的 Q 网络在智能体之间进行信息传递,该网络的输出不仅包含 Q 值,还包括在智能体之间交互的信息,其中RIAL

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

使用双网络结构分别输出动作和离散信息以降低动作空间的维度，而DIAL则建立了专门的通信通道实现信息端到端的双向传递，相比RIAL，DIAL在通信效率上更具优势。

RIAL和DIAL算法只能传递离散化的信息，这就限制了智能体之间通信的信息量和实时度。为了解决这一问题，Sukhbaatar等人提出了通信网(CommNet)算法[43]，该算法在智能体之间构建了一个具备传输连续信息能力的通信通道，它确保环境中任何一个智能体都可以实时传递信息，该通信机制具有两个显著特点：(1)每个时间步都允许所有的智能体自由通信；(2)采用广播的方式进行信息传递，智能体可以根据需求选择接受信息的范围。这样每个智能体都可以根据需要选择和了解环境的全局信息。实验表明，CommNet在合作型非完全观测(详见4.2节)环境中的表现优于多种无通信算法和基线算法。

国内对于通信规则型的MDRL研究也取得了不小的进展，其中最著名的有阿里巴巴团队提出的多智能体双向协同网络(Bidirectionally-Coordinated Nets，BiCNet)[44]，该方法旨在完成即时策略类游戏星际争霸2中的微观管理任务，即实现对低级别、短时间交战环境中己方的单位控制。算法基于AC框架和双向循环神经网络(Bidirectional Recurrent Neural Network，Bi-RNN)，前者使得每个智能体在独立做出行动决策的同时又能与其他智能体共享信息，后者不仅可以保证智能体之间连续互相通信，还可以存储本地信息。该方法的核心思路是将复杂的交战过程简化为双人零和博弈问题，由以下元组表示：

$$\left(S,\{A_i\}_{i=1}^{N},\{B_i\}_{i=1}^{M},T,\{R_i\}_{i=1}^{N+M}\right) \tag{15}$$

其中，S为所有智能体共享的全局状态，M、N和A、B分别为敌对双方智能体的数量和动作空间，全局状态转移概率为：

$$T{:}S\times A^N\times B^M\rightarrow S \tag{16}$$

第i个智能体收到的环境回报为：

$$R_i{:}S\times A^N\times B^M\rightarrow \mathrm{R} \tag{17}$$

其中一方的全局回报函数为：

$$r(s,a,b)=\frac{1}{M}\sum_{j=N+1}^{N+M}\Delta R_j^t(s,a,b)-\frac{1}{N}\sum_{i=1}^{N}\Delta R_i^t(s,a,b) \tag{18}$$

对于敌我双方智能体来说，学习目标分别为最大化和最小化这一全局期望累计回报，二者遵循Minimax原则，最优Q值可表示为：

$$Q^*(s,a,b)=r(s,a,b)+\lambda\max_{\theta}\min_{\phi}Q^*\left(s',a_\theta(s'),b_\phi(s')\right) \tag{19}$$

算法假设敌方策略不变，SG过程可被简化为MDP过程进行求解：

$$Q^*(s,a)=r(s,a)+\lambda\max_{\theta}Q^*\left(s',a_\theta(s')\right) \tag{20}$$

经过充分训练，BiCNet算法可以让游戏中的单位成功实现如进攻、撤退、掩护、集火攻击、异构单位配合等多种智能协作策略。

近来，通信规则型MDRL方法的研究成果主要侧重于改进智能体之间的通信模型以提升通信效率，如北京大学多智能体团队[45]提出了一个基于注意力机制(ATOC Architecture)的通信模型，让智能体具备自主选择通信对象的能力；Kim等人[46]将通信领域的介质访问控制(Medium Access Control)方法引入到MDRL中，提出了规划通信(Schedule Communication)模型，优化了信息的传输模式，让智能体具备全时段通信能力。表3总结了通信规则型方法。

总的来说，通信规则型方法优势在于算法在智能体之间建立的显式的信道可以使得智能体学习到更好的集体策略，但其缺点主要是由于信道的建立所需参数较多，算法的设计架构一般较为复杂。

3.3　互相协作型

此类方法并不直接在多智能体间建立显式的通信规则，而是使用传统MARL中的一些理论使智能体学习到合作型策略。

值函数分解网(Value Decomposition Networks，VDN)[47]及其改进型QMIX[48]和QTRAN[49]等将环境的全局回报按照每个智能体对环境做出的贡献进行拆分，具体是根据每个智能体对环境的联合回报的贡献大小将全局Q函数分解为与智能体一一对应的本地Q函数，经过分解后每个Q函数只和智能体自身的历史状态和动作有关，上述三种算法的区别在于Q函数分解的方式不同，VDN才采用简单的线性方式进行分解，而

表3　通信规则型方法总结分析

算法名称	通信方式	优点	缺点
RIAL/DIAL	使用中心化的Q网络和通信通道在智能体之间进行信息传递	通信模型十分简化，实现了智能体间基本的通信，后者还能实现双向通信	只能传递离散化的信息，无法实现复杂信息传递
CommNet	在智能体之间构建了一个具备传输连续信息能力的通信通道	可在智能体之间进行连续信息通信，智能体可以自由选择通信时间和范围	集中式训练模式，复杂度较高，容易维度爆炸，可扩展性差
BiCNet	使用AC框架和双向循环神经网络共享信息实现通信	实现了异构智能体间的即时通信，可对星际争霸2中多个单位进行复杂配合的控制	专门为特定应用设计，方法泛化能力有限
ATOC/Schedule Communication等	采用注意力机制通信和规划通信模型	对信息传递的模式进行了优化，智能体实现了全时段通信能力	模型复杂度较高，针对特定场景和应用，不具泛化性

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

QMIX和QTRAN则采用非线性的矩阵分解方式，另外，QTRAN在具有更加复杂的Q函数网络结构。该值函数分解思想有效地提升了多智能体环境中的学习效果。

多智能体深度确定性策略梯度(Multi-Agent Deep Deterministic Policy Gradient，MADDPG)[50]是一种基于AC框架的算法，且遵循集中训练分散执行原则。算法中每个智能体都存在一个中心化的Critic接收其他智能体的信息(如动作和观测等)，即：$Q_i^{\mu}(o_1,a_1,o_2,a_2,\cdots,o_N,a_N)$，同时每个智能体的Actor网络只根据自己的部分观测执行策略$a_i=\mu_{\theta_i}(o_i)$，每个智能体Critic网络的梯度遵循：

$$\nabla_{\theta_i}J(\boldsymbol{\mu}_i)=E_{x,a\sim D}\left[\nabla_{\theta_i}\boldsymbol{\mu}_i(a_i|o_i)\nabla_{a_i}Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_N)\big|_{a_i=\boldsymbol{\mu}_i(o_i)}\right] \tag{21}$$

算法通过不断优化损失函数得到最优策略：

$$L(\theta_i)=E_{x,a,r,x'}\left[\left(Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_N)-y\right)^2\right],\quad y=r_i+\gamma Q_i^{\mu'}(\boldsymbol{x}',a'_1,a'_2,\cdots,a'_N)\big|_{a'_j=\boldsymbol{\mu}'_j(o_j)} \tag{22}$$

该算法无需建立显示的通信规则，同时适用合作型、竞争型、混合型等多种环境，能够很好地解决多智能体环境非平稳问题。

反事实多智能体策略梯度(Counterfactual Multi-Agent Policy Gradients，COMA)[51]是另一种基于AC框架的合作型算法。该算法采用完全集中的学习方式，主要解决多智能体信度分配问题，也就是如何在只能得到全局回报的合作型环境中给每个智能体分配回报值，该算法的解决方式是假设一个反事实基线(Counterfactual Baseline)，即在其他智能体的动作保持不变的情况下去掉其中一个智能体的动作，然后计算当前Q值和反事实Q值的差值得到优势函数，并进一步得出每个智能体的回报，COMA不受环境的非平稳性带来的影响，但其可扩展性相对较差。

Pham等人将参数共享(Parameter Sharing，PS)[52]框架与多种DRL算法结合应用于多智能体环境。PS框架的核心思想是利用一个全局的神经网络收集所有智能体的各类参数进行训练。但在执行阶段仍然保持各个智能体的独立，相应的算法有PS-DQN、PS-DDPG、PS-TRPO等。

国内的多智能体协作型算法研究也有不小的进展，天津大学的郝建业等人提出了加权双深度Q网络(Weighted Double Deep Q-Network，WDDQN)算法，该方法将双Q网络结构和宽大回报(Lenient Reward)理论加入到经典算法DQN中[53]，前者主要解决深度强化学习算法固有的过估计问题，后者则侧重于提升合作型多智能体环境随机策略更新能力，此外作者还改变了DQN中的经验库抽取机制以提升样本学习质量。实验结果显示该方法在平均回报和收敛速率上都超过了多种基线算法。表4总结了互相协作型方法。

互相协作型方法虽然不需要复杂的通信建模过程，但由于在训练过程中融入了传统多智能体算法的规则(如值函数分解、参数共享、纳什均衡等)，兼具易实现性和高效性，且此类方法应对不同学习场景的通用性也很强，其缺点是适用环境较为单一(无法应对完全对抗型环境)。

3.4　建模学习型

在此类方法中，智能体主要通过为其他智能体建模的方式分析并预测行为，深度循环对手网络(Deep Recurrent Opponent Network，DRON)[17]是早期比较有代表性的建模学习型算法。它的核心思想是建立两个独立的神经网络，一个用来评估Q值，另一个用来学习对手智能体的策略，该算法还使用多个专家网络分别表征对手智能体的所有策略以提升学习能力。与DRON根据对手智能体特征进行建模的方式不同，深度策略推理Q网络(Deep Policy Inference Q-Network，DPIQN)[54]则完全依靠其他智能体的原始观测进行建模，该算法通过一些附属任务(Auxiliary Task)学习对方智能体的策略，附属任务完成的情况直接影响算法的损失函数，这样就将学习智能体的Q函数和对方智能

表4　互相协作型方法总结分析

算法名称	主要机制	优点	缺点
VDN/QMIX/QTRAN等	采用值函数分解思想，按照智能体对环境的联合回报的贡献大小分解全局Q函数	很好地解决了多智能体信度分配问题	目前不存在一种有效的分解机制对多智能体环境具有普适性，分解方法较为单一，如线性分解、矩阵分解等方式
MADDPG	采用AC框架和集中训练分散执行模式，每个智能体拥有中心化的Critic接受全局信息	无需建立现实通信规则，很好地解决了环境非平稳性问题，算法容易收敛至全局最优解	可扩展性差，不支持大量智能体训练，训练周期较长
COMA	采取完全集中的训练方式，算法中引入反事实基线概念，面向合作型任务	解决多智能体环境中的非平稳性问题，不受环境改变的影响	可扩展性较差，集中式训练器容易维度爆炸，对参数要求较高
PS-DQN/PS-DDPG/PS-TRPO等	将参数共享机制融入DRL算法	提升算法的训练效率和收敛速度	不适用异构型智能体的合作，只支持智能体完成较为简单的任务
WDDQN	将双Q网络结构和宽大回报理论引入DQN	解决了多智能体环境下的过估计问题，提升了随机策略更新能力	算法对数据和参数要求较高，对环境适应性不强

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

体的策略特征联系起来,并降低了环境的非平稳性对智能体学习过程的影响,该算法还引入自适应训练流程让智能体在学习对手策略和最大化 Q 值之间保持平衡,这表明DPIQN可同时适用于敌方和己方智能体。自预测建模(Self Other Modeling,SOM)[55]算法使用智能体自身的策略预测对方智能体的行为,它也有两个网络,只不过另一个网络不学习其他智能体的策略而是对它们的目标进行预测,SOM适用于多目标场景。

此外,博弈论和MARL的结合也是该类方法的重要组成部分,如神经虚拟自学习(Neural Fictitious Self-Play,NFSP)[56],算法设置了两个网络模拟两个智能体互相博弈的过程,智能体的目标是找到近似纳什均衡,该算法适用于不完美信息博弈对抗,如德州扑克。Minimax原则也是博弈论中的重要理论,清华大学多智能体团队将其与MADDPG算法相结合并提出了M3DDPG算法[57],其中Minimax原则用于估计环境中所有智能体的行为都完全敌对情况下的最坏结局,而智能体策略按照所估计的最坏结局不断更新,这就提升了智能体学习策略的鲁棒性,保证了学习的有效性。表5对建模学习型方法进行了总结分析。

建模学习型方法旨在对手或队友策略不可知的情况下以智能体建模的方式对行为进行预测,这类算法一般鲁棒性较强,可以应对多种不同的场景,但计算和建模的复杂度较高,无法适应大型复杂的多智能体系统,所以实际应用较少。表6对多智能体强化学习方法的分类进行了对比分析。

4　多智能体深度强化学习的关键问题

尽管MDRL方法在理论、框架、应用等层面都有不小的进展,但该领域的探索还处在起步阶段,与单智能体的诸多方法相同,MDRL方法在实验及应用层面也面临许多问题和挑战,本章对MDRL方法所面临的关键问题和现行解决方案及发展方向进行总结。

4.1　环境的非平稳性问题

与单智能体环境不同,在多智能体环境中,每个智能体不仅要考虑自己动作及回报,还要综合考虑其他智能体的行为,这种错综复杂的交互和联系过程使得环境不断地动态变化。在非平稳的环境中,智能体间动作及策略的选择是相互影响的,这使得回报函数的准确性降低,一个良好的策略会随着学习过程的推进不断变差。环境的非平稳性大大增加算法的收敛难度,降低算法的稳定性,并且打破智能体的探索和利用平衡。为解决环境非平稳问题,研究人员从不同角度对现有方法进行了改进,Castaneda[58]提出了两种基于DQN的改进方法,它们分别通过改变值函数和回报函数的方式增加智能体之间的关联性;Diallo等人[59]则将并行运算机制引入到DQN中,加速多智能体在非平稳环境中的收敛;Foerster等人[42]则致力于通过改进经验库机制让算法适用于不断变化的非平稳环境,为此他提出了两种方法:(1)为经验库中的数据设置重要性标记,丢弃先前产生而不适应当前环境的数据;(2)使用“指纹”为每个从经验库中取出的样本单元做时间标定,以提升训练数据的质量。目前针对环境非平稳性的解决方案较多,也是未来MDRL领域学术研究的热门方向。

4.2　非完全观测问题

在大部分多智能体系统中,智能体在交互过程中无法了解环境的完整信息,它们只能根据所能观测到的部分信息做出相对最优决策,这就是部分可观测马尔可夫决策过程(Partially Observable Markov Decison Process,POMDP),POMDP是MDP在多智能体环境中的扩展,它可由多元组 $G=\langle S,A,T,R,Q,O,\gamma,N\rangle$ 表示,其中 S 和 A 分别表示智能体的状态和动作集合,T 和 R 则表示状态转移方程和回报函数,Q 和 O 则为每个智能体 Q 值和部分观测值,每个智能体并不知道环境的全局状态 $s\in S$,只能将自己的部分观测值当作全局状态,即:

表5　建模学习型方法总结分析

算法名称	主要机制	适用场景
DRON	采用独立的神经网络学习对手智能体的策略	混合型场景
DPIQN	根据其他智能体观测进行建模,通过附属任务学习智能体策略	多子任务的对抗型场景
SOM	使用自身策略预测其他智能体的行为	多目标合作型场景
NFSP	双网络模拟智能体互博弈过程,通过找到纳什均衡完成算法收敛	棋牌类游戏等不完美信息的完全竞争场景
M3DDPG	改进自经典MDRL算法MADDPG,引入Minimax原则对环境中敌对智能体进行最坏估计,提升了生成策略的鲁棒性	混合型场景中强鲁棒策略的生成

表6　多智能体强化学习方法分类对比分析

方法分类	优点	缺点
无关联方法	简单,易实现	无法解决多智能体环境的非平稳性问题,难稳定收敛
通信规则型方法	建立的显式信道,学习最优策略	参数较多,算法的设计架构较为复杂
互相协作型方法	无需通信建模,兼具易实现性和高效性,通用性强	应用环境单一(无法应对完全竞争型环境)
建模学习型方法	可预测其他智能体行为,鲁棒性强,适用不同场景	计算复杂度高,无法适用大型系统

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$o_i = Q_i(s): S \rightarrow O_i \tag{23}$$

并以此为根据做出决策：

$$\pi_\theta(a_i|o_i): O_i \times A_i \rightarrow [0,1] \tag{24}$$

得到一个关于状态动作的回报值：

$$r_i = R(s,a_i): S \times A_i \rightarrow \mathrm{R} \tag{25}$$

之后智能体转移到了下一个状态：

$$s' = T(s,a_1,a_2,\cdots,a_N): S \times A_1 \times A_2 \times \cdots \times A_N \rightarrow S \tag{26}$$

每个智能体的目标都是最大化自己的总回报：

$$E[R_i] = E\left[\sum_{t=0}^{T} \gamma^t r_i^t\right] \tag{27}$$

其中，r_i^t 是第 i 个智能体在时间 t 上的总回报，T 为时间范围。现有研究中有多种方法用于求解POMDP问题，如DRQN[17]算法中的循环网络结构保证了智能体在非完全观测环境中高效学习和提升策略，其改进算法深度分布式循环 Q 网络（Distributed Deep Recurrent Q-Network，DDRQN）[17]在解决多智能体POMDP问题中也取得了很好的效果，算法主要有三点创新：（1）在训练过程中将智能体的上一步动作作为下一步的输入，从而加速算法的收敛；（2）在智能体间引入权重分享机制，降低学习参数的数量；（3）放弃经典DRL算法中的经验库机制以降低环境非平稳性带来的影响。与此同时，也有不少方法致力于解决大规模POMDP问题。Gupta提出了一种叫作课程学习（Curriculum Learning，CL）的训练机制，类似于人脑渐进的学习过程，该机制首先让少量智能体合作完成简单的任务，然后逐渐增加智能体数量和任务难度，整个训练过程还支持多种算法的融合。目前该领域的研究侧重于异构智能体POMDP问题的求解。

4.3　多智能体环境训练模式问题

早期的大部分MDRL算法都采用集中式或分散式两种训练模式，前者使用一个单独的训练网络总揽整个学习过程，算法很容易过拟合且计算负荷太大；后者采用多个训练网络，每个智能体之间完全独立，算法由于不存在中心化的目标函数，往往难以收敛。所以两种训练模式只支持少量智能体的小型系统。集中训练和分散执行（Centralized Learning and Decentralized Execution，CLDE）[50]融合了以上两种模式的特点，智能体一方面在互相通信的基础上获取全局信息进行集中式训练，然后根据各自的部分观测值独立分散执行策略，该模式最大的优点是允许在训练时加入额外的信息（如环境的全局状态、动作或者回报），在执行阶段这些信息又可被忽略，这有利于实时掌控和引导智能体的学习过程。近来采用CLDE训练模式的MDRL算法不断增加。以上述三种基本模式为基础，研究人员不断探索出新的多智能体训练模式，它们各有优长，可应用于不同的多智能体环境，限于篇幅原因本文就不做赘述。

4.4　多智能体信度分配问题

在合作型多智能体环境中，智能体的个体回报和全局回报都可以用来表征学习进程，但个体回报一般难以获得，所以大部分实验都使用全局回报计算回报函数。如何将全局回报分配给每个智能体，使其能够精准地反映智能体对整体行为的贡献，这就是信度分配问题。早起的方法如回报等分在实验中的效果很差。差分回报（Difference Rewards）[60]是一个比较有效的方法，其核心是将每个智能体对整个系统的贡献值进行量化，但这种方法的缺点是很难找到普适的量化标准，另外该方法容易加剧智能体间信度分配的不平衡性。COMA[51]中优势函数（Advantage Function）思想也是基于智能体的贡献大小进行信度分配，算法通常使用神经网络拟合优势函数，该方法无论是在分配效果还是效率上都好于一般方法。总之，信度分配是MDRL算法必须面临的重要问题，如何精确高效地进行信度分配直接关系到多智能体系统的成败，这也是近来多智能体领域研究的重点。

4.5　过拟合问题

过拟合最早出现在监督学习算法中，指的是算法只能在特定数据集中取得很好的效果，而泛化能力很弱。多智能体环境中同样存在过拟合问题，比如在学习过程中其中一个智能体的策略陷入局部最优，学习策略只适用于其他智能体的当前策略和当前环境。目前有3种比较成熟的解决方法：（1）策略集成（Policy Emsemble）[50]机制，即让智能体综合应对多种策略以提升适应性；（2）极小极大（Minimax）[57]机制，即让智能体学习最坏情况下的策略以增强算法的鲁棒性；（3）消息失活（Message Dropout）[61]机制，即在训练时随机将神经网络中特定节点进行失活处理以提升智能体策略的鲁棒性和泛化能力。

5　多智能体深度强化学习的测试平台

许多标准化的平台如OpenAI Gym已经支持在模拟环境中测试经典DRL和MARL算法，但由于MDRL起步较晚，目前来看还是一个较为新颖的领域，所以其配套测试平台还有待进一步发展完善。当前已有一些研究机构或个人开发了一部分开源的模拟器和测试平台用于MDRL算法的分析和测试，它们各有特点，且面向不同类型的环境，本章将进行简单介绍。

Bușoniu等人开发出一种基于matlab的多智能体物体运输（Coordinated Multi-agent Object Transportation，CMOT）环境[25]，其本质上是一个2D网格双智能体环境，Palmer等人在该环境原始版本的基础上进行了扩展，使其支持随机回报和噪声观测等复杂条件，该平台面向传统MARL合作型算法的测试工作（http://www.dcsc.tudelft.nl/）；炸弹人游戏（Pommerman）是由Facebook AI实验室和Google AI联合赞助的多智能体环境

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

测试平台，它同样也是一个二维网格环境，最多可以容纳四个智能体，支持合作型、竞争型、混合型等多种多智能体算法的测试，并且还支持非完全观测环境和智能体的通信建模，测试人员依托该平台不仅可以将自己的改进算法和基线算法进行对比，还可以与其他测试人员的算法实时对抗。另外该平台还支持python、Java等多语言编写（https://www.pommerman.com/）；MuJoCo最早是由华盛顿大学运动控制实验室开发的物理仿真引擎，可应用于具有丰富接触行为的复杂动态系统，平台支持多种可视化的多智能体环境，研究人员目前已将多智能体足球游戏（Multi-agent Soccer Game）应用到该引擎中，让环境模拟2对2比赛，该平台的优点是可支持三维动作空间；谷歌DeepMind和Blizzard公司联合开发了一个基于即时策略类游戏星际争霸2的DRL平台SC2LE，该平台提供基于Python的开源接口来与游戏引擎进行通信，其中的多智能体测试主要针对小型场景的微观管理，场景中的每个单位都由一个独立的智能体控制，该智能体基于自己的部分观测做出动作，该平台已经成功应用多种MDRL算法，如QMIX[48]、COMA[51]等；基于3D沙盒游戏《我的世界》的Malmo平台可用于完成多场景合作型任务，并支持多种开源项目，具备实时调试的功能；以卡牌类游戏Hanabi为背景的学习平台支持多玩家多任务竞争，该游戏的主要特点是玩家不仅分析自己手中的牌，同时也知晓其他玩家的部分信息，所以非常适合针对POMDP问题算法的测试；竞技场（Arena）是一个基于Unity引擎的多智能体搜索平台，该平台的支持多种经典多智能体场景（如社会难题、多智能体搬运等），并支持在智能体之间通信规则的搭建，目前该平台已能够实现如IDQN[41]、ITRPO[41]、IPPO[41]等几种简单的MDRL算法。

6　多智能体深度强化学习的实际应用及前景展望

6.1　多智能体深度强化学习的实际应用

MARL的实际应用领域十分广泛，涉及领域包括自动驾驶、能源分配、编队控制、航迹规划、路由规划、社会难题等，下文对此进行简要的介绍。

Prasad和Dusparic[62]将MDRL模型应用到能源分配领域，模拟场景为一个由数幢楼房组成的社区，并假定该社区中的每幢楼房每年消耗的能源不高于产生的能源，在该场景中，楼房由智能体表示，它们通过学习适当的多智能体策略优化能源在建筑物间的分配方式，环境中的全局回报由社区中的能源总量来表示，即：

$$reward=-\left(\sum_{i=1}^{n}c(h_i)-g(h_i)\right) \tag{28}$$

其中$c(h_i)$和$g(h_i)$分别表示第i幢楼房的能源消耗和能源产出，另外环境中设置一个控制智能体主导智能体数量的增减和能源的实时分配，实验表明该模型在保持楼房能源平衡的表现好于随机策略模型。但该模型的缺点为训练中不能实时观察智能体的行为，另外该模型也不能适用于大型环境（楼房数量的上限为10），模型的架构也有待完善（未能考虑能源分类等更为复杂的情况）。

Leibo等人[34]提出了解决贯序社会难题（Sequential Social Dilemmas，SSD）的模型，它用于解决POMDP环境下多智能体环境中的合作问题。Hüttenrauch等人[63]则尝试控制大量的智能体完成复杂的任务，该应用也被称为群体智能系统。系统使用的方法基于演员评论家框架，利用全局状态信息学习每个智能体的Q函数，研究人员还截取环境的实时图像用于收集分析智能群体的状态信息。该群体智能系统可以完成如搜索救援、分布式组装等多种复杂合作型任务。Calvo和Dusparic[64]则在群体智能系统中加入了多种对抗型MDRL算法使系统中的不同智能体独立并发的训练，改进后的系统能够胜任如城市交通信号控制等多种类型的任务。

通信规则型算法在实际问题中的应用较为广泛。Nguyen等[65]在智能体之间构建了一种特殊的通信通道以图片形式传输人类知识，场景使用A3C算法，其优点是支持异构型智能体间的合作；Noureddine等[66]基于合作型DRL算法构建了一套松耦合的分布式多智能体环境，环境中的智能体可以像人类团队一样互帮互助，适用于解决资源和任务的分配问题；CommNet算法因其强大的通信能力也多被用于高复杂度的大型任务分配问题并取得了不错的效果，但它也有计算复杂度高、通信开销大等缺点。

互相合作型算法主要在编队控制、交通规划、数据分析[67]等方面有所应用。其中Lin等人[68]将多种合作型算法应用在大型编队控制问题上，他们的方法聚焦于如何平衡分配交通资源以提升交通效率，减少拥堵，该方法使用参数共享机制保证多个车辆间的协同。Schmid等人[69]则将经济学中的交易规则引入到多智能体系统中，在该系统中，智能体的动作、状态、回报等参数都被看成可以互相交易的资源。该方法有效地抑制了每个独立智能体的贪婪行为，从而利于达到系统回报的最大化，该系统在社会福利分配等经济学问题中有可观的应用。

6.2　多智能体深度强化学习的前景展望

MDRL虽然在众多领域都有实际应用，但由于起步时间较晚，理论成熟度较低，其发展潜力十分巨大，前景相当可观。

现有的MDRL算法大部分采用无模型的结构，虽然简化了算法的复杂度，并且适用于复杂问题求解，但该类方法需要海量的样本数据和较长的训练时间为支撑，基于模型的方法则具有数据利用效率高、训练时间短、

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

泛化性强等优点,基于模型的强化学习算法在单智能体领域取得了较多进展,其必然是MDRL未来的重点研究方向[70];模仿学习(Imitation Learning)[71]、逆向强化学习(Inverse Reinforcement Learning)[72]、元学习(Meta Learning)[73]等新兴概念在单智能体领域已经有了不小的成果,解决了不少现实问题,其在多智能体领域的应用前景将相当可观;在城市交通信号控制、电子游戏竞技等实际应用中,同构型的智能体拥有如行为、目标和领域知识等较多的共性特点,可以通过集中训练的方式提升学习的效率和速率,但当环境是由大量异构型智能体组成时,如何学习到有效的协同策略并得到最优解成为了一大难题,这其中需要解决如异构型智能体信度分配、过估计、可扩展性等多种实质问题,总之大型异构多智能体系统也是一个非常有前景的研究方向[74];人机交互这个词正不断地被大众所接受,文献[75-77]中人机智能交互是MDRL未来的发展方向。因为在复杂环境中人类无法单独处理海量数据,而机器则难以解决非形式化的隐性问题,所以人类智慧与机器智慧的结合至关重要。近来,研究人员已经在尝试将人在回路(Human-On-The-Loop)[76]框架融合到MDRL算法中,即人类和智能体合作解决复杂问题,在传统的"人在回路"设定中,智能体自动地完成其所分配的任务,然后等待人类指挥员做出决策并继续自己的任务。未来将实现从"人在回路"到"人控回路"的飞跃,即从机器完成任务和人做决策的传统时序框架到机器与人智能化协作共同完成任务的新体系,人作为终极掌控者将会在多智能体领域中扮演愈发重要的角色。

7 结语

本文对按照由浅入深的次序对多智能体深度强化学习进行了分析,介绍了包括MDRL的相关概念、经典算法、主要挑战、实际应用和发展方向等。本文首先在引言部分简要介绍了MDRL的背景知识,随后按照从单智能体到多智能体的发展顺序简述了传统MARL的基本框架,并按照回报函数的不同将MDRL分为合作型、竞争型和混合型三类,接着对DRL及其代表算法进行了简要的概括,由此引入MDRL的概念,之后根据多智能体间的关联方式的不同将MDRL算法分为无关联型、通信规则型、互相协作型和建模学习型四大类,并分别对各类别的主要算法进行介绍和对比分析,最后对MDRL算法的测试平台、主要挑战、实际应用和未来展望进行简要的阐述。通过本文可以得出结论:多智能体深度强化学习是个新兴的、充满创新点的、快速发展的领域,无论是学术研究还是工程运用方面都较多空间亟待拓展,相信随着研究的不断深入,将会诞生更多方法解决各类复杂的问题,实现人工智能更美好的未来。

参考文献:

[1] 李杨,徐峰,谢光强,等.多智能体技术发展及其应用综述[J].计算机工程与应用,2018,54(9):13-21.

[2] 吴军,徐昕,王健,等.面向多机器人系统的增强学习研究进展综述[J].控制与决策,2011,26(11):1601-1610.

[3] 韩凯,孙金生.带虚拟领导者的多智能体系统的非光滑一致性[J].计算机工程与应用,2019,55(8):147-150.

[4] Sutton R S,Barto A G.Introduction to reinforcement learning[M].Cambridge:MIT Press,1998.

[5] LeCun Y,Bengio Y,Hinton G.Deep learning[J].Nature,2015,521(7553):436-444.

[6] Henderson P,Islam R,Bachman P,et al.Deep reinforcement learning that matters[C]//Thirty-Second AAAI Conference on Artificial Intelligence,2018.

[7] Egorov M.Multi-agent deep reinforcement learning[J].CS231n:Convolutional Neural Networks for Visual Recognition,2016.

[8] White C.Markov decision processes[M].New York:Springer,2001.

[9] Beard R W,Saridis G N,Wen J T.Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation[J].Automatica,1997,33(12):2159-2177.

[10] Watkins C,Dayan P.Q-learning[J].Machine Learning,1992,8(3/4):279-292.

[11] Sutton R S,McAllester D A,Singh S P,et al.Policy gradient methods for reinforcement learning with function approximation[C]//Advances in Neural Information Processing Systems,2000:1057-1063.

[12] Williams R J.Simple statistical gradient-following algorithms for connectionist reinforcement learning[J].Machine Learning,1992,8(3/4):229-256.

[13] Silver D,Lever G,Heess N,et al.Deterministic policy gradient algorithms[C]//Proceedings of the 31st International Conference on Machine Learning,2014.

[14] Mnih V,Kavukcuoglu K,Silver D,et al.Playing Atari with deep reinforcement learning[J].arXiv:1312.5602,2013.

[15] Van Hasselt H,Guez A,Silver D.Deep reinforcement learning with double q-learning[C]//Thirtieth AAAI Conference on Artificial Intelligence,2016.

[16] Wang Z,Schaul T,Hessel M,et al.Dueling network architectures for deep reinforcement learning[J].arXiv:1511.06581,2015.

[17] Hausknecht M,Stone P.Deep recurrent Q-learning for partially observable MDPs[C]//2015 AAAI Fall Symposium Series,2015.

[18] Schaul T,Quan J,Antonoglou I,et al.Prioritized experience replay[J].arXiv:1511.05952,2015.

[19] Lillicrap T P,Hunt J,Pritzel A,et al.Continuous control with deep reinforcement learning[J].arXiv:1509.02971,2015.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[20] Konda V R,Tsitsiklis J N.Actor-critic algorithms[C]//Advances in Neural Information Processing Systems,2000:1008-1014.

[21] Mnih V,Badia A P,Mirza M,et al.Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning,2016:1928-1937.

[22] Schulman J,Levine S,Abbeel P,et al.Trust region policy optimization[C]//International Conference on Machine Learning,2015:1889-1897.

[23] Schulman J,Wolski F,Dhariwal P,et al.Proximal policy optimization algorithms[J].arXiv:1707.06347,2017.

[24] Shapley S.Stochastic games[J].Proceedings of the National Academy of Sciences,1953,39(10):1095-1100.

[25] Buşoniu L,Babuška R,De Schutter B.Multi-agent reinforcement learning:an overview[M]//Innovations in multi-agent systems and applications-1.Berlin,Heidelberg:Springer,2010:183-221.

[26] Wang Y,De Silva C W.Multi-robot box-pushing:single-agent Q-learning vs.team Q-learning[C]//2006 IEEE/RSJ International Conference on Intelligent Robots and Systems,2006:3694-3699.

[27] Galindo A,Giupponi L.Distributed Q-learning for aggregated interference control in cognitive radio networks[J].IEEE Transactions on Vehicular Technology,2010,59(4):1823-1834.

[28] Littman M L.Markov games as a framework for multi-agent reinforcement learning[M]//Machine learning proceedings 1994.[S.l.]:Morgan Kaufmann,1994:157-163.

[29] Hu J,Wellman M P.Nash Q-learning for general-sum stochastic games[J].Journal of Machine Learning Research,2003,4:1039-1069.

[30] Greenwald A,Hall K,Serrano R.Correlated Q-learning[C]//ICML,2003:242-249.

[31] Littman M L.Friend-or-foe Q-learning in general-sum games[C]//ICML,2001:322-328.

[32] Harati A,Ahmadabadi M N,Araabi B N.Knowledge-based multiagent credit assignment:a study on task type and critic information[J].IEEE Systems Journal,2007,1(1):55-67.

[33] Tampuu A,Matiisen T,Kodelja D,et al.Multiagent cooperation and competition with deep reinforcement learning[J].PloS One,2017,12(4).

[34] Leibo Z,Zambaldi V,Lanctot M,et al.Multi-agent reinforcement learning in sequential social dilemmas[C]//Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems,2017:464-473.

[35] Lerer A,Peysakhovich A.Maintaining cooperation in complex social dilemmas using deep reinforcement learning[J].arXiv:1707.01068,2017.

[36] Heinz A.Self-play,deep search and diminishing returns[J].ICGA Journal,2001,24(2):75-79.

[37] Bansal T,Pachocki J,Sidor S,et al.Emergent complexity via multi-agent competition[J].arXiv:1710.03748,2017.

[38] Mahadevan S,Connell J.Automatic programming of behavior-based robots using reinforcement learning[J].Artificial Intelligence,1992,55(2/3):311-365.

[39] Ng Y,Harada D,Russell S.Policy invariance under reward transformations:theory and application to reward shaping[C]//ICML,1999:278-287.

[40] Raghu M,Irpan A,Andreas J,et al.Can deep reinforcement learning solve Erdos-Selfridge-Spencer games?[J].arXiv:1711.02301,2017.

[41] Gupta J K,Egorov M,Kochenderfer M.Cooperative multi-agent control using deep reinforcement learning[C]//International Conference on Autonomous Agents and Multiagent Systems.Cham:Springer,2017:66-83.

[42] Foerster J,Assael I A,de Freitas N,et al.Learning to communicate with deep multi-agent reinforcement learning[C]//Advances in Neural Information Processing Systems,2016:2137-2145.

[43] Sukhbaatar S,Fergus R.Learning multiagent communication with backpropagation[C]//Advances in Neural Information Processing Systems,2016:2244-2252.

[44] Peng P,Yuan Q,Wen Y,et al.Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games[J].arXiv:1703.10069,2017.

[45] Jiang J,Lu Z.Learning attentional communication for multi-agent cooperation[C]//Advances in Neural Information Processing Systems,2018:7254-7264.

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

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

[48] Rashid T,Samvelyan M,De Witt C S,et al.QMIX:monotonic value function factorisation for deep multi-agent reinforcement learning[J].arXiv:1803.11485,2018.

[49] Hostallero W,Son K,Kim D,et al.Learning to factorize with transformation for cooperative multi-agent reinforcement learning[C]//Proceedings of the 31st International Conference on Machine Learning,2019.

[50] Lowe R,Wu Y,Tamar A,et al.Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Advances in Neural Information Processing Systems,2017:6379-6390.

[51] Foerster J N,Farquhar G,Afouras T,et al.Counterfactual multi-agent policy gradients[C]//Thirty-Second AAAI

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Conference on Artificial Intelligence,2018.
[52] Pham H,Guan Y,Zoph B,et al.Efficient neural architecture search via parameter sharing[J].arXiv:1802.03268,2018.
[53] Zheng Y,Meng Z,Hao J,et al.Weighted double deep multiagent reinforcement learning in stochastic cooperative environments[C]//Pacific Rim International Conference on Artificial Intelligence.Cham:Springer,2018:421-429.
[54] Hong Z,Su S,Shann Y,et al.A deep policy inference Q-network for multi-agent systems[C]//Proceedings of the 17th International Conference on Autonomous-Agents and Multi-Agent Systems,2018:1388-1396.
[55] Raileanu R,Denton E,Szlam A,et al.Modeling others using oneself in multi-agent reinforcement learning[J].arXiv:1802.09640,2018.
[56] Heinrich J,Silver D.Deep reinforcement learning from self-play in imperfect-information games[J].arXiv:1603.01121,2016.
[57] Li S,Wu Y,Cui X,et al.Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient[C]//AAAI Conference on Artificial Intelligence,2019.
[58] Castaneda A O.Deep reinforcement learning variants of multi-agent learning algorithms[D].University of Edinburgh.School of Informatics,2016.
[59] Diallo E A O,Sugiyama A,Sugawara T.Learning to coordinate with deep reinforcement learning in doubles pong game[C]//2017 16th IEEE International Conference on Machine Learning and Applications(ICMLA),2017:14-19.
[60] Devlin S,Yliniemi L,Kudenko D,et al.Potential-based difference rewards for multiagent reinforcement learning[C]//Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems,2014:165-172.
[61] Srivastava N,Hinton G,Krizhevsky A,et al.Dropout:a simple way to prevent neural networks from overfitting[J].The Journal of Machine Learning Research,2014,15(1):1929-1958.
[62] Prasad A,Dusparic I.Multi-agent deep reinforcement learning for zero energy communities[J].arXiv:1810.03679,2018.
[63] Hüttenrauch M,Šošić A,Neumann G.Guided deep reinforcement learning for swarm systems[J].arXiv:1709.06011,2017.
[64] Calvo J A,Dusparic I.Heterogeneous multi-agent deep reinforcement learning for traffic lights control[C]//AICS,2018:2-13.
[65] Nguyen N D,Nguyen T,Nahavandi S.System design perspective for human-level agents using deep reinforcement learning:a survey[J].IEEE Access,2017,5:27091-27102.
[66] Noureddine B,Gharbi A,Ahmed S.Multi-agent deep reinforcement learning for task allocation in dynamic environment[C]//ICSOFT,2017:17-26.
[67] Golzadeh M,Hadavandi E,Chelgani S C.A new ensemble based multi-agent system for prediction problems:case study of modeling coal free swelling index[J].Applied Soft Computing,2018,64:109-125.
[68] Lin K,Zhao R,Xu Z,et al.Efficient large-scale fleet management via multi-agent deep reinforcement learning[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2018:1774-1783.
[69] Schmid K,Belzner L,Gabor T,et al.Action markets in deep multi-agent reinforcement learning[C]//International Conference on Artificial Neural Networks.Cham:Springer,2018:240-249.
[70] Doll B B,Simon D A,Daw N D.The ubiquity of model-based reinforcement learning[J].Current Opinion in Neurobiology,2012,22(6):1075-1081.
[71] Schaal S.Is imitation learning the route to humanoid robots?[J].Trends in Cognitive Sciences,1999,3(6):233-242.
[72] 陈希亮,曹雷,何明,等.深度逆向强化学习研究综述[J].计算机工程与应用,2018,54(5):24-35.
[73] Vilalta R,Drissi Y.A perspective view and survey of meta-learning[J].Artificial Intelligence Review,2002,18(2):77-95.
[74] Kapetanakis S,Kudenko D.Reinforcement learning of coordination in heterogeneous cooperative multi-agent systems[M]//Adaptive agents and multi-agent systems II.Berlin,Heidelberg:Springer,2004:119-131.
[75] Roth E M,Hanson M L,Hopkins C,et al.Human in the loop evaluation of a mixed-initiative system for planning and control of multiple UAV teams[C]//Proceedings of the Human Factors and Ergonomics Society Annual Meeting,2004:280-284.
[76] Nahavandi S.Trusted autonomy between humans and robots:toward human-on-the-loop in robotics and autonomous systems[J].IEEE Systems,Man,and Cybernetics Magazine,2017,3(1):10-17.
[77] Santhanam G R,Holland B,Kothari S,et al.Human-on-the-loop automation for detecting software side-channel vulnerabilities[C]//International Conference on Information Systems Security.Cham:Springer,2017:209-230.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net