多智能体的深度强化学习系统:挑战回顾,解决方案和应用

Thanh Thi Nguyen、Ngoc Duy Nguyen 和 Saeid Nahavandi ，IEEE 高级会员

摘要 强化学习 (RL) 算法已经存在了几十年,并被用于解决各种顺序决策问题。然而,这些算法在处理高维环境时面临着巨大的挑战。深度学习的最新发展使 RL 方法能够为复杂而有能力的代理提供最佳策略,这些代理可以在这些具有挑战性的环境中高效运行。本文讨论了深度 RL 的一个重要方面,它与需要多个代理进行通信和合作以解决复杂任务的情况有关。本文介绍了与多智能体深度 RL (MADRL) 相关问题的不同方法,包括非平稳性、部分可观测性、连续状态和动作空间、多智能体训练方案和多智能体迁移学习。本文将分析和讨论所审查方法的优缺点,并探讨其相应的应用。本文将提供有关各种 MADRL 方法的见解,并可在未来开发出更强大、更有用的多智能体学习方法来解决实际问题。

索引词 连续动作空间、深度学习、深度强化学习（RL）、多智能体、非平稳、部分可观测性、评论、机器人、调查。

一、引言

加强 学习是由试错法 (TE) 程序引发的,该程序由 Thorndike 在 1898 年的一项猫行为实验中进行 [1]。1954 年,明斯基 [2] 设计了第一台神经计算机,称为随机神经模拟强化计算器 (SNARC),它模拟老鼠的大脑来解决迷宫难题。SNARC 标志着 TE 学习提升到计算周期。近二十年后,Klopf [3] 将心理学中的时间差异 (TD) 学习机制整合到 TE 学习的计算模型中。这种整合成功地使 TE 学习成为一种可行的大规模方法系统。1989 年,Watkins 和 Dayan [4] 将包括贝尔曼方程和马尔可夫决策过程 (MDP)在内的最优控制理论 [5]与 TD 学习结合在一起,形成了著名的 Q 学习。此后， Q 学习被用于解决各种现实问题,但它无法解决高维问题,因为高维问题的计算量会随着输入数量的增加而急剧增加。这个问题被称为维数灾难,超出了传统计算机的计算限制。2015 年,Mnih等人[6] 通过将深度学习与强化学习 (RL)相结合,部分克服了维数灾难,取得了重要突破。从那时起,深度 RL 就引起了研究界的极大关注。图 1 展示了 RL 发展的里程碑,涵盖了从 TE 方法到深度 RL。

强化学习源于心理学中的动物学习,因此它可以模仿人类的学习能力,在与环境的交互中选择能够最大化长期利润的行为。强化学习在机器人技术和自主系统中得到了广泛的应用,例如,Mahadevan 和 Connell [7] 设计了一个可以推动立方体的机器人 (1992);Schaal [8] 创建了一个可以有效解决杆平衡任务的人形机器人 (1997);Benbrahim 和 Franklin [9] 制作了一个可以在不了解任何环境的情况下学会走路的双足机器人 (1997);Riedmiller等人[10] 组建了一支足球机器人队 (2009);Mulling等人[11] 训练机器人打乒乓球 (2013)。

2015 年深度强化学习的成功真正标志着现代强化学习的真正成功,当时 Mnih等人[6] 利用一种名为深度Q 网络 (DQN) 的结构创建了一个代理,该代理在 49 款经典 Atari 游戏中的表现优于专业玩家 [12]。

2016 年,谷歌的 DeepMind 创建了一个自学成才的 AlphaGo 程序,它可以击败最优秀的职业围棋选手,包括中国的柯洁和韩国的李诗杜 [13]。深度强化学习还被用于解决 MuJoCo 物理问题 [14] 和 3-D 迷宫游戏 [15]。2017 年,OpenAI 宣布了一款机器人,它可以击败在线游戏 Dota 2 上最优秀的职业玩家,这款游戏应该比围棋游戏更复杂。更重要的是,由于其实用的方法,深度强化学习已成为解决现实问题的一种有前途的方法,例如非线性系统的最优控制 [16]、行人监管 [17] 或交通网格信号控制 [18]。

谷歌、特斯拉和优步等企业一直在竞相制造自动驾驶汽车。此外,最近

稿件于 2019 年 2 月 7 日收到;修订于 2019 年 7 月 11 日、2019 年 10 月 18 日和 2019 年 12 月 15 日;接受于 2020 年 2 月 25 日。出版日期 2020 年 3 月 20 日;当前版本日期 2020 年 8 月 18 日。本文由副主编 D. Liu 推荐。 （通讯作者:Thanh Thi Nguyen。）

Thanh Thi Nguyen 就职于迪肯大学 (Burwood 校区) 信息技术学院,地址:澳大利亚维多利亚州 Burwood 3125 (电子邮件:thanh.nguyen@deakin.edu.au)。

Ngoc Duy Nguyen 和 Saeid Nahavandi 就职于迪肯大学（Waurn Ponds 校区)智能系统研究与创新研究所,地址:澳大利亚维多利亚州 Waurn Ponds 3216（电子邮件:duy.nguyen@deakin.edu.au;saeid.nahavandi@deakin.edu.au）。

本文中有一张或多张图的彩色版本可供下载在线网址:http://ieeexplore.ieee.org。
数字对象标识符 10.1109/TCYB.2020.2977374

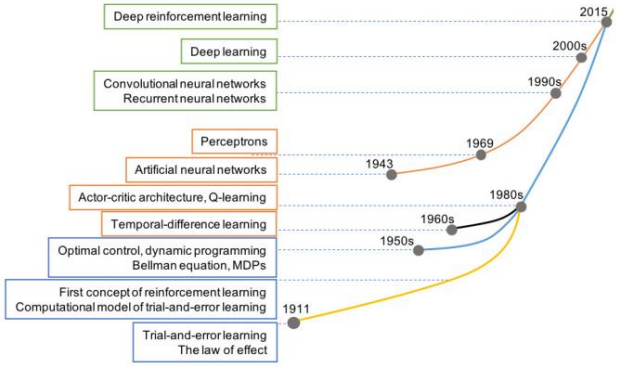

图 1. 深度强化学习通过不同的重要里程碑的出现。

深度强化学习的进展已经扩展到解决 NP 难题,例如车辆路径问题,这在物流中至关重要[19],[20]。

随着现实世界的问题变得越来越复杂,许多情况下单个深度强化学习智能体无法应对。在这种情况下,多智能体系统 (MAS) 的应用必不可少。在 MAS 中,智能体必须竞争或合作才能获得最佳整体结果。此类系统的示例包括多人在线游戏、生产工厂中的合作机器人、交通控制系统以及无人驾驶飞行器 (UAV)、监视和航天器等自主军事系统。

在文献中深度强化学习的众多应用中,有大量研究在 MAS 中使用深度强化学习,即多智能体深度强化学习 (MADRL)。从单智能体领域扩展到多智能体环境带来了一些挑战。先前的研究考虑了不同的角度,例如,Busoniu等人[21] 研究了智能体的稳定性和适应性方面,Bloembergen等人[22] 分析了进化动力学,Hernandez-Leal等人[23] 考虑了突发行为、沟通和合作学习角度,da Silva等人[24] 回顾了多智能体强化学习 (MARL) 中知识重用自主性的方法。

本文概述了多智能体学习中的技术挑战以及应对这些挑战的深度强化学习方法。我们涵盖了多智能体学习的众多视角,包括非平稳性、部分可观测性、多智能体训练方案、MAS 中的迁移学习以及多智能体学习中的连续状态和动作空间。本研究还回顾和分析了多智能体学习在各个领域的应用。在最后一节中,我们介绍了多智能体学习的广泛讨论和有趣的未来研究方向。

II.背景:强化学习

A. 初步

RL 是一种 TE 学习,其学习方式为 1)直接与环境交互;2)随着时间的推移进行自我学习;3)最终实现指定目标。具体而言,RL 将任何决策者(学习者)定义为代理,将代理之外的任何事物定义为环境。代理与环境之间的交互通过三个基本元素来描述:1)状态s;2)动作a;3)奖励r [25]。时间步骤t时的环境状态表示为st。因此,代理检查st并执行相应的动作。然后环境将其状态st更改为st+1 ,并向代理提供反馈奖励rt+1 。

通过定义策略的概念,可以将代理的决策形式化。策略 π 是从任何感知状态s到从该状态采取的动作a的映射函数。如果从s 中选择动作a的概率为:对于所有有状态s ,p(a|s) = 1,则策略是确定性的。相反,如果存在状态s,使得p(a|s) < 1,则策略是随机的。无论哪种情况,我们都可以将策略 π 定义为从某个状态中选择的候选动作的概率分布,如下所示

$$\pi = (s) \quad = p(ai|s) \ \forall ai \ \varepsilon \ {}_{p} \ \wedge \ p(ai|s) = 1 \qquad (1)$$

其中π表示策略 π 的所有候选动作(动作空间) 。为清楚起见,我们假设动作空间是离散的,因为连续情况可以通过使用积分符号直接推断出来。此外,我们假设下一个状态st+1和反馈奖励rt+1完全由当前状态-动作对(st, at)决定,而与历史无关。任何满足此“无记忆”条件的 RL 问题称为 MDP。因此, RL 问题的动态(模型)完全通过给出所有转移概率p(ai|s)来指定。

B.贝尔曼方程

提醒一下,代理在每个时间步骤t都会收到反馈奖励rt+1 ,直到到达终止状态sT 。但是,即时奖励rt+1并不代表长期利润,我们改为利用时间步骤t的广义回报值Rt Rt = rt+1 + γ rt+2 + γ 2rt+3 + · · · + γ T−t−1rT

$$= \sum_{我=0}^{T-t-1} \gamma^{i} rt+i+1 \qquad (2)$$

其中 γ 是一个折扣因子,因此 0 ⩽ γ < 1。当 γ 接近于 1 时,代理将变得远视,反之亦然,当 γ 接近于 0 时,代理将变得近视。

下一步是定义一个价值函数,用于评估某个状态s或某个状态-动作对(s, a) 的“好”程度。具体来说,通过从 s 获取预期回报值来计算策略 π 下状态s的价值函数: Vπ (s) = E [Rt|st = s, π]。

同样,状态-动作对(s, a) 的价值函数为Qπ (s, a) = E [Rt|st = s, at = a, π]。我们可以利用价值函数来比较两个策略 π 和 π 之间的“好坏” ,规则如下 [25]:

$$\pi \leqslant \pi \Longleftrightarrow [V\pi (s) \leqslant V\pi(s) \ \forall s] \vee [Q\pi (s, a) \leqslant Q\pi(s, a) \ \forall (s, a)]。 \qquad (3)$$

基于(2) ,我们可以展开Vπ (s)和Qπ (s, a),以表示两个连续状态s = st和s = st+1之间的关系为 [25]

$$V\pi (s) = \sum_{a \in \rightarrow} \pi(s, a) \sum_{s} ps\,|s, 且Ws \rightarrow s\,|a + \gamma V\pi\, s \qquad (4)$$

和

$$Q\pi(s,a) = \sum_{s'} p(s'|s,a)W_{s\to s'|a} + c\sum_{s'} p(s'|s,a)Q\pi(s',\pi(s')) \quad (5)$$

其中$W_{s\to s'|a} = E[r_{t+1}|s_t = s, a_t = a, s_{t+1} = s']$.
解（4)或（5），我们可以分别找到值函数V（s)或Q（s，a)。方程（4)和（5)称为贝尔曼方程。

动态规划及其变体[26]–[31]可用于近似贝尔曼方程的解。

然而,它需要问题的完整动态信息,因此当状态数量很大时,由于传统计算机的内存和计算能力不足,这种方法是不可行的。在下一节中,我们将回顾两种无模型强化学习方法 [不需要转移概率p(ai|s) 的知识]来近似值函数。

表一
强化学习方法的特点

Category	Pros	Cons
Model-free	● Dynamics of environment is unknown ● Deal with larger state-space environments	○ Requires "exploration" condition
On-policy	● Stable when using with function approximator ● Suitable with continuous state-space problems	○ Policy must be stochastic
Off-policy	● Simplify algorithm design ● Can tackle with different kinds of problems ● Policy can be deterministic	○ Unstable when using with function approximator
Bootstrapping	● Learn faster in most cases	○ Not as good as nonbootstrapping methods on mean square error

表二
动态规划与强化学习方法的比较

Category	Dynamic programming	RL Methods			
		MC	Sarsa	Q-learning	AC
Model-free		✓	✓	✓	✓
On-policy		✓	✓		✓
Off-policy		✓		✓	✓
Bootstrapping	✓		✓	✓	✓

C.强化学习方法

在本节中,我们回顾了强化学习中的两种著名学习方案:1) 蒙特卡罗 (MC) 和 2) TD 学习。这些方法不需要环境的动态信息,也就是说,它们可以处理比动态规划方法更大的状态空间问题。

1) 蒙特卡罗方法:该方法通过重复生成情节并记录每个状态或每个状态-动作对的平均回报来估计价值函数。 MC 方法不需要任何转移概率知识,也就是说,MC 方法是无模型的。然而,这种方法做出了两个基本假设来确保收敛发生:1) 情节数量很大;2) 每个状态和每个动作都必须被访问相当多次。

一般来说,MC 算法分为两类:1)在策略和 2)离策略。在在策略方法中,我们使用策略 π 进行评估和探索。

因此,策略 π 必须是随机的或软的。相比之下,离策略使用不同的策略 π = π 来生成情节,因此 π 可以是确定性的。虽然离策略因其简单性而受到青睐,但在处理连续状态空间问题以及与函数逼近器（如神经网络)一起使用时,在策略方法更稳定 [32]。

2)时间差分法:与MC类似,TD方法也是从经验中学习（无模型方法）。

然而,与 MC 不同,TD 学习不会等到情节结束才进行更新。它利用贝尔曼方程 (4) 对情节中的每一步进行更新,因此可能提供更快的收敛速度。等式 (6) 表示单步 TD 方法

$$V_i(s_t) \leftarrow\!- \alpha V_{i-1}(s_t) + (1-\alpha)\,r_{t+1} + \gamma\, V_{i-1}(s_{t+1}) \quad (6)$$

其中 α 是步长参数,0 <α< 1。TD 学习使用先前估计的值Vi－1来更新当前值Vi,这称为引导方法。

基本上,在大多数情况下,引导方法比非引导方法学习得更快[25]。TD学习也分为两类:1)在策略TD控制（Sarsa)和2)离策略TD控制（Q学习）。

在实践中,MC 和 TD 学习通常使用表内存结构（表格方法)来保存每个状态或每个状态-动作对的值函数。由于内存不足,在解决状态数量很大的复杂问题时,它们效率低下。因此,设计了一个参与者-评论家 (AC) 架构来克服这一限制。具体来说,AC 包括两个独立的代理内存结构:1)参与者和 2)评论家。参与者结构用于根据观察到的状态选择合适的动作并转移到评论家结构进行评估。评论家结构使用 TD 误差函数来决定所选动作的未来趋势。AC 方法可以是在线策略或离线策略,具体取决于实施细节。表一总结了 RL 方法的特点及其优缺点。表二重点介绍了动态规划和 RL 方法之间的差异,包括 MC、Sarsa、 Q 学习和 AC。与动态规划相反,RL 算法是无模型的 [33]。虽然其他 RL 方法（例如 Sarsa、 Q-learning 和 AC)使用 boot-strapping 方法,但 MC 需要重新启动 episode 来更新其价值函数。值得注意的是,基于 AC 的算法是最通用的,因为它们可以属于任何类别。

III.深度强化学习:单一代理

深度 Q 网络

深度强化学习是一个广义的术语,指的是深度学习与强化学习的结合,用于处理高维环境[34]–[36]。2015 年,Mnih等人[6]首次

time 宣布这一组合的成功,创建了一个可以熟练玩 49 款 Atari 游戏的自主代理。简而言之,作者提出了一种名为 DQN 的新结构,它利用卷积神经网络 (CNN) [37] 直接解释来自环境的输入状态s的图形表示。DQN 的输出产生在状态 s 下采取的所有可能动作a ∈的 Q 值,其中表示动作空间 [38]。因此,DQN 可以看作是一个由 β 参数化的策略网络 τ,它不断训练以近似最佳策略。从数学上讲,DQN 使用贝尔曼方程将损失函数 L(β) 最小化为

t

t

L(β) = E r + γ 最大 质量 ，a |β − Q(s, a|β) 2 . (7)

然而,使用神经网络来近似值函数已被证明是不稳定的,并且可能由于来自相关样本的偏差而导致发散 [32]。为了使样本不相关,Mnih等人[6] 创建了一个目标网络τ ,由 β 参数化,该网络每N步从估计网络 τ更新一次。此外,生成的样本存储在经验重放存储器中。然后从经验重放中随机检索样本并输入到训练过程中 [39],如图 2 所示。

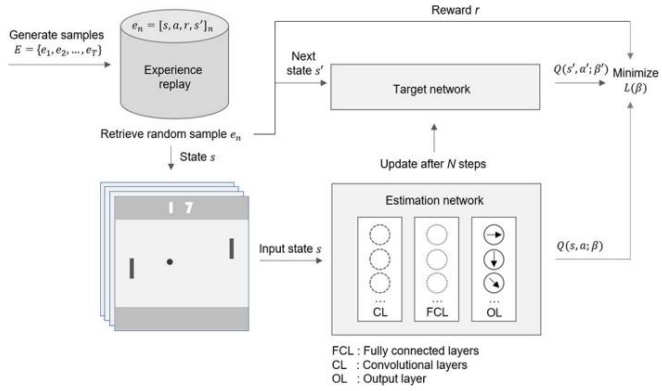

图 2. 具有经验重放记忆的深度Q 网络结构和目标网络,其参数每N步从估计网络更新一次,以确保稳定的学习过程。

尽管 DQN 基本上解决了强化学习中的一个难题,即维数灾难,但这只是解决完全现实世界应用的第一步。DQN 有许多缺点,可以通过不同的方案来弥补,从简单的形式到复杂的修改,我们将在下一节中讨论。

B.DQN 变体

DQN 变体的第一个也是最简单的形式是 [40] 和 [41] 中提出的双 DQN (DDQN)。DDQN 的思想是将“贪婪”动作的选择与动作评估分开。通过这种方式,DDQN 希望减少训练过程中对 Q 值的高估。换句话说， (7) 中的最大运算符被解耦为两个不同的运算符,如以下损失函数所示:

LDDQN(β)

= E r + γ Q s 精氨酸含量 最大 ，a |β − Q(s, a|β) 2 。

(8)

在 57 款 Atari 游戏上进行的经验实验结果表明,未经调整的 DDQN 的归一化性能是 DQN 的两倍,调整后是 DQN 的三倍。

其次,DQN 中的经验回放对于打破样本间的相关性起着重要作用,同时还能提醒策略网络可能很快忘记的“稀有”样本。然而,从经验回放中随机选择样本并不能完全分离样本数据。具体来说,我们更希望稀有和与目标相关的样本比冗余样本出现得更频繁。

因此,Schaul等人[42] 提出了优先经验

根据样本 i 的 TD 误差绝对值,对样本i进行优先排序

pi = |δi|=|ri + γ max Q si, a|β − Q(si−1, ai−1|β)|。 (9)

优先经验重放与 DDQN 相结合,实现了策略网络的稳定收敛,并在 57 款 Atari 游戏的归一化平均分数方面实现了比 DQN 高出五倍的性能。

DQN 的策略评估过程很难在“冗余”情况下发挥作用,也就是说,有两个以上的候选动作可以选择而不会产生任何负面结果。例如,当开车时,如果前方没有障碍物,我们可以沿着左车道或右车道行驶。

如果左车道前方有障碍物,我们必须驶入右车道以避免撞车。因此,如果我们只关注前方的道路和障碍物会更有效率。为了解决这种情况,Wang等[43] 提出了一种称为决斗网络的新型网络架构。在决斗架构中,有两个共存的附属网络:一个网络由 θ 参数化,估计状态值函数V(s|θ);另一个网络由 θ 参数化,估计优势动作函数A(s, a|θ)。然后使用以下方程将两个网络聚合起来,以近似Q 值函数:

,

Q s, a|θ,θ = V(s|θ)

+ A s, a|θ − |π | 1 一个s,一个|th .(10)

由于决斗网络代表动作价值函数,因此它与 DDQN 相结合,并优先考虑经验重放,在 Atari 域中将性能提升至标准 DQN 的六倍 [43]。

DQN 的另一个缺点是它使用四帧历史作为策略网络的输入。因此,DQN 无法有效解决当前状态依赖于大量历史信息的问题,例如“Double Dunk”或“Frostbite”[44]。这些游戏通常被称为部分可观察的 MDP 问题。直接的解决方案是用循环长短期记忆替换策略网络最后一个卷积层之后的全连接层,如 [44] 中所述。这个 DQN 的变体称为

深度循环Q 网络 (DRQN) 在“Double Dunk”和“Frostbite”游戏中的表现比标准 DQN 高达 700%。此外,Lample 和 Chaplot [45] 通过在 DRQN 中添加游戏特征层,成功创建了一个在 3D FPS（第一人称射击游戏)环境“Doom”上击败普通玩家的代理。DRQN 的另一个有趣变体是深度注意循环Q 网络 (DARQN) [46]。在那篇文章中,Sorokin等人在 DRQN 中添加了注意力机制,以便网络可以只关注游戏中的重要区域,从而允许更小的网络参数,从而加快训练过程。结果,在游戏“Seaquest”上,DARQN 取得了 7263 分,而 DQN 和 DRQN 分别取得了 1284 分和 1421 分。

深度强化学习:多智能体

MAS 之所以受到广泛关注,是因为它们能够通过各个智能体的协作来解决复杂任务。在 MAS 中,智能体相互通信并与环境交互。在多智能体学习领域,MDP 被推广为随机博弈或马尔可夫博弈。我们将n表示为智能体的数量， S表示为环境状态的离散集， Ai， i = 1, 2,..., n,表示每个智能体的一组动作。所有智能体的联合动作集定义为A = A1 × A2 × · · · × An。

状态转移概率函数表示为p : S × A × S → [0, 1],奖励函数表示为r : S × A × S → Rn。每个代理的价值函数取决于联合动作和联合策略,其特征为Vπ : S × A → Rn。以下各节介绍了挑战和 MADRL 解决方案以及它们在解决实际问题中的应用。

A. MADRL:挑战与解决方案

1) 非平稳性:与单智能体环境相比,控制多个智能体带来了一些额外的挑战,例如智能体的异质性、如何定义合适的集体目标或可扩展到大量智能体（这需要设计紧凑的表示）,更重要的是非平稳性问题。在单智能体环境中,智能体只关注其自身行为的结果。在多智能体领域,智能体不仅观察其自身行为的结果,还观察其他智能体的行为。智能体之间的学习很复杂,因为所有智能体都可能相互交互并同时学习。多个智能体之间的交互不断重塑环境并导致非平稳性。在这种情况下,智能体之间的学习有时会导致智能体的策略发生变化,并可能影响其他智能体的最佳策略。对某个行为的潜在回报的估计是不准确的,因此,在多智能体环境中,某一时刻的良好策略在未来不可能一直如此。在单智能体环境中应用的 Q 学习收敛理论无法保证适用于大多数多智能体问题,因为马尔可夫特性在非平稳环境中不再成立 [47]。因此,信息的收集和处理必须以一定的递归方式进行,同时确保它不会影响智能体的稳定性。

在多智能体环境下,探索利用困境可能会更加复杂。

流行的独立 Q 学习[48] 或基于经验重放的 DQN [6] 不是为非平稳环境设计的。Castaneda [49] 提出了两种 DQN 变体,即深度重复更新Q 网络 (DRUQN) 和深度松耦合Q 网络 (DLCQN),以处理 MAS 中的非平稳性问题。DRUQN 是基于[50] 和 [51] 中介绍的重复更新 Q 学习 (RUQL) 模型开发的。它旨在通过以与选择动作的可能性成反比的方式更新动作值来避免策略偏差。另一方面,DLCQN 依赖于[52] 中提出的松耦合 Q 学习,它使用每个代理的负奖励和观察来指定和调整其独立度。通过这种独立度,代理可以学会决定在不同情况下是需要独立行动还是与其他代理合作。同样,Diallo等人 [9] 提出了一种基于 DQN 的独立学习方法。 [53] 将 DQN 扩展为多智能体并发 DQN ,并证明该方法可以在非平稳环境中收敛。

Foerster等人[54] 分别介绍了两种在 MADRL 中稳定 DQN 经验重放的方法。第一种方法使用重要性采样方法自然衰减过时数据,而第二种方法使用指纹消除从重放记忆中检索到的样本的年龄歧义。

最近,为了处理 MAS 中由于多个智能体并发学习而产生的非平稳性,Palmer等人[55] 提出了一种方法,即 lenient-DQN (LDQN),该方法应用衰减温度值的宽容性来调整从经验重放记忆中采样的策略更新。

多智能体环境中的宽容性描述了这样一种情况:一个正在学习的智能体会忽略同伴的不良行为（这会导致较低的奖励）,但仍然与同伴合作,希望同伴将来能够改进自己的行为。例如,智能体 A 和 B 正在学习踢足球。由于失误或训练不足,智能体 B 无法处理智能体 A 传给他的球。在这种情况下,有了宽容性,智能体 A 会认为智能体 B 可以提高自己的技能,因此智能体 A 会继续将球传给智能体 B,而不是认为智能体 B 没有踢足球的技能而不再将球传给智能体 B [56]。LDQN 被应用于协调多智能体物体运输问题,并将其性能与滞后 DQN（HDQN）进行了比较 [57]。实验结果表明,在随机奖励环境下,LDQN 在收敛到最优策略方面优于 HDQN。 [58] 中,宽容的概念以及预定的重放策略也被纳入加权 DDQN (WDDQN),以处理 MAS 中的非平稳性。实验表明,在具有随机奖励和大状态空间的两个多智能体环境中,WDDQN 的性能优于 DDQN。

2)部分可观测性:在实际应用中,很多情况下智能体对环境只有部分可观测性。这个问题在多智能体问题中更为严重,因为它们通常更复杂,

大规模。换句话说,当代理不知道与环境相关的状态的完整信息时

它们与环境相互作用。在这种情况下,
代理观察环境的部分信息,并
需要在每个时间步骤中做出“最佳”决策。这
这类问题可以用部分可观测模型来建模
MDP(POMDP)。

在目前的文献中,许多深度强化学习模型已经
被提议用来处理 POMDP。Hausknecht 和 Stone [44]
提出了基于长短期记忆网络的DRQN。
借助循环结构,基于 DRQN 的代理能够
在部分
可观察环境。与 DQN 不同,DRQN 近似
Q(o, a),是一个Q 函数,具有观测值o和
动作a,由循环神经网络执行。DRQN 将网络ht-1的隐藏状态视为内部状态。
DRQN
因此以Q 函数(ot, ht−1, a; θi)为特征
其中θi是第 i 次训练时网络的参数
步骤。在 [59] 中,DRQN 扩展为深度分布式循环
Q 网络(DDRQN)用于处理多智能体 POMDP 问题。
DDRQN 的成功依赖于三个显著的特点,
即最后行动输入、跨智能体权重共享和禁用经验重放。第一个特征,即最后行动
输入,需要提供每个输入的先前操作
代理作为其下一步的输入。代理间权重共享
意味着所有代理都只使用一个网络的权重,
是在训练过程中学习的。禁用经验回放只是排除了经验回放功能

DQN 的Q 函数。因此,DDRQN 学习的是
形式Q(om 鸣, 嗯t−1, m,上午t−1, 是; θi),其中每个代理接收
其自身索引m作为输入。权重共享减少了学习时间,因为它减少了参数的数量

学习。虽然每个代理都有不同的观察
和隐藏状态,但是这种方法假设代理
采取相同的行动。为了解决复杂问题,
自主代理通常有不同的动作集。对于
例如,无人机在空中机动,而机器人则在
地面。因此,无人机和机器人的行动空间是
不同,因此跨代理权重共享功能无法
得到应用。

在部分可观察的领域中扩展到多智能体系统是一个具有挑战性的问题。
Gupta等人[60]
将课程学习技术扩展到 MAS,
集成了三类深度强化学习,包括策略梯度、TD误差和AC方法。课程原则

是先学习完成简单的任务以积累知识,然后再执行复杂的任务。

这适用于代理较少的 MAS 环境
在扩展之前先进行协作,以容纳更多
代理完成越来越困难的任务。实验结果显示了课程学习的活力

将深度强化学习算法扩展到复杂多智能体的方法
问题。

Hong等人[61] 提出了一种深度策略推理
Q 网络(DPIQN)用于对 MAS 进行建模,其增强版本深度循环策略推理Q
网络(DRPIQN)用于
应对部分可观测性。DPIQN 和 DRPIQN
通过调整网络注意力来学习策略
特征及其在各个阶段的 Q 值
训练过程。实验表明,总体而言
DPIQN 和 DRPIQN 的性能优于基线 DQN 和 DRQN [44]。在部分

可观测性,但扩展到多任务、多智能体问题,Omidshafiei等人[57] 提出了一种方法,称为
多任务 MARL(MT-MARL)集成了滞后学习器 [62]、DRQN [44]、蒸馏 [63] 和并发
经验重放轨迹 (CERT),这是经验重放策略的去中心化扩展,

在[6]中。代理没有明确地提供任务
身份(因此是部分可观察性),同时他们合作
学习使用以下方法完成一组分散的 POMDP任务
稀疏奖励。然而,这种方法有一个缺点
无法在异构环境中执行
代理。

除了部分可观察性之外,还有一些情况
代理必须处理极其嘈杂的观察结果,这
与环境的真实状态相关性较弱。
Kilinc 和 Montana [64] 提出了一种方法,表示为
结合深度确定性策略梯度的 MADDPG-M
(DDPG)和通信媒介来解决这些情况。代理需要决定他们的观察结果是否

与其他代理共享信息,并且通信策略与主要策略同时学习

通过经验。最近,Foerster等人[65] 提出
用于学习的贝叶斯动作解码器(BAD)算法
具有合作部分可观测设置的多个代理。
引入了一个新的概念,即公众信念MDP
基于采用近似贝叶斯更新的 BAD
获得公众信任,具有公开可观察的特征
环境。BAD 依赖于分解和近似的信念状态来发现惯例,从而使代理能够

有效地学习最优策略。这与
人类通常用来解释他人行为的心理理论。实验结果基于原理证明

两步矩阵博弈和合作部分信息
纸牌游戏 Hanabi 证明了所提方法相对于传统策略梯度的效率和优越性

算法。

3)新加坡金融管理局培训计划:单一
Agent Deep RL 应用于多智能体环境就是学习每个智能体
通过将其他代理视为环境的一部分,独立地实现 Q 学习算法

在[66]中。这种方法容易出现过度拟合[67],而且计算成本高昂,因此代理的数量
参与的程度有限。另一种流行的方法是
集中学习和分散执行,其中
可以通过应用
通过开放的沟通渠道采用集中式方法[68]。
分散策略,每个代理可以根据
在局部观测方面具有优势
可观察性和执行过程中的有限通信。
去中心化政策的中心化学习已经成为
多智能体设置中的标准范式,因为学习过程可能发生在模拟器和实验室中

没有沟通限制,也没有额外的状态
相关信息 [68]。

三种不同的 MAS 训练方案包括
集中学习、并行学习和参数

算法1 PS-TRPO

1:初始化策略网络0的参数,并信任
区域大小2:对
于i ←− 0, 1,...,为所有代理生
3: 成轨迹,因为 τ πθi
使用具有共享参数的策略。
4: 对于每个代理m,计算优势值
Aπθi (om, m, am),其中m为代理索引。
5: 搜索最大化πθ (a|o,m) πθk的πθi+1

$$L(\theta) = E_{o \quad p\theta k, a \quad \pi\theta k}(a|o,m)\overline{A\theta k}(o, m, a) \text{ subject to } D^{-}$$

KL(πθi πθi+1) ⩽ 其中DKL是两个策略分布之间的 KL 散度,
pθ是由πθ引起的状态访问的折扣频率。6 :结束

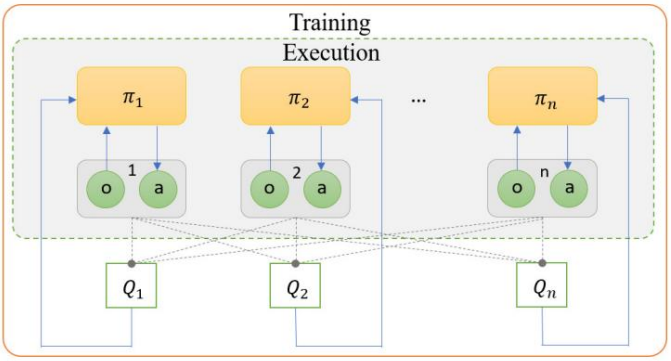

图 3. 基于集中学习和分散执行的 MADDPG,其中代理的策略由集中评论家利用来自其他代理的观察和行动的增强信息来学习。

在[60]中对共享进行了研究。集中式策略试图从所有代理的联合观察中获得联合行动,而并发学习则使用联合奖励信号同时训练代理。在后者中,每个代理根据私人观察独立学习自己的策略。

另外,参数共享方案允许使用所有代理的经验同时训练代理,尽管每个代理都可以获得独特的观察结果。凭借执行分散策略的能力,参数共享可用于扩展单代理深度强化学习算法,以适应多代理系统。具体而言,参数共享和 TRPO 的结合,即 PS-TRPO,已在 [60] 中提出,并在算法 1 中简要总结。

PS-TRPO 在处理部分可观测性下的高维观测和连续动作空间时表现出色。

Foerster等人[69] 引入了基于集中式学习方法的强化跨智能体学习 (RIAL) 和可微分跨智能体学习 (DIAL) 方法,以改善智能体学习通信。在 RIAL 中,深度Q 学习具有一个循环结构来解决部分可观测性问题,其中独立 Q 学习让各个智能体学习自己的网络参数。DIAL 通过通道将梯度从一个智能体推送到另一个智能体,从而实现跨智能体端到端反向传播。同样,Sukhbaatar等人[70] 开发了一个通信神经网络 (CommNet),使动态智能体能够在学习完全协作任务的策略的同时学习持续通信。

与 CommNet 不同的是,何等人[71] 提出了一种方法,即深度强化对手网络（DRON）,将对手代理的观察编码到 DQN 中,以在没有领域知识的情况下联合学习对手的策略和行为。

在 [72] 和 [73] 中,将分散式和集中式视角融入到分层主从架构中,形成一个称为主从 MARL (MS-MARL) 的模型,以解决 MAS 中的通信问题。主代理接收并集体处理来自从属代理的消息,然后为每个从属代理生成唯一的指导性消息。从属代理使用自己的来自主代理的信息和指导性消息,以便采取行动。与对等架构相比,该模型大大减少了 MAS 内的通信负担,尤其是当系统有许多代理时。

[74] 提出了基于 AC 策略梯度算法的多智能体深度确定性策略梯度（MADDPG)方法。MADDPG 具有集中学习和分散执行的范式,其中批评者使用额外信息来简化训练过程,而参与者则根据自己的本地观察采取行动。

图 3 说明了 MADDPG 的多智能体分散式参与者和集中式评论家组件,其中在执行阶段仅使用参与者。

最近,[75] 中引入了另一种多智能体 AC 方法,即反事实多智能体 (COMA),该方法也依赖于集中式学习和分散式执行方案。与 MADDPG [74] 不同,COMA 可以处理多智能体信用分配问题 [76],其中智能体很难从合作环境中联合行动产生的全局奖励中计算出他们对团队成功的贡献。然而,COMA 的缺点是只关注离散动作空间,而 MADDPG 能够有效地学习连续策略。

4) 连续动作空间:大多数深度强化学习模型只能应用于离散空间 [77]。例如,DQN [6] 仅限于离散和低维动作空间的问题,尽管它可以处理高维观测空间。DQN 旨在找到具有最大动作价值的动作,因此需要在连续动作（状态)空间中的每一步进行迭代优化过程。

离散化动作空间是将深度强化学习方法应用于连续域的一种可能解决方案。然而,这会产生许多问题,尤其是维数灾难:动作数量相对于自由度数量呈指数增长。

Schulman等人[78] 提出了一种信任域策略优化 (TRPO) 方法,该方法可以扩展到连续状态和动作,用于优化机器人运动和基于图像的游戏领域的随机控制策略。Lillicrap等人[77] 提出了一种离策略算法,即 DDPG,它利用 AC 架构 [79],[80] 来处理连续动作空间。基于

表三
多智能体学习挑战及其解决方法

Challenges	Value-based	Actor-critic	Policy-based
Partial observ-ability	DRQN [44]; DDRQN [59]; RIAL and DIAL [69]; Action-specific DRQN [89]; MT-MARL [57]; PS-DQN [60]; RL as a Rehearsal (RLaR) [68]	PS-DDPG and PS-A3C [60]; MADDPG-M [64]	DPIQN and DRPIQN [61]; PS-TRPO [60]; Bayesian action decoder (BAD) [65]
Non-stationarity	DRUQN and DLCQN [49]; Multi-agent concurrent DQN [53]; Recurrent DQN-based multi-agent importance sam-pling and fingerprints [54]; Hysteretic-DQN [57]; Lenient-DQN [55]; WDDQN [58]	MADDPG [74]; PS-A3C [60]	PS-TRPO [60]
Continuous action spaces		Recurrent DPG [82]; DDPG [77]	TRPO [78]; PS-TRPO [60]
Multi-agent training schemes	Multi-agent extension of DQN [66]; RIAL and DIAL [69]; CommNet [70]; DRON [71]; MS-MARL [72], [73]; Linearly fuzzified joint Q-function for MAS [90]	MADDPG [74]; COMA [75]	
Transfer learning in MAS	Policy distillation [63]; Multi-task policy distillation [85]; Multi-agent DQN [87]	Progressive networks [83]	Actor-Mimic [86]

确定性策略梯度（DPG)[81],DDPG
使用参数化的参与者函数确定性地将状态映射到特定动作,同时保持 DQN 学习

批评者的观点。然而,这种方法需要大量的训练集才能找到解决方案,这在无模型强化方法中很常见。Heess等人[82]

将 DDPG 扩展为循环 DPG（RDPG）,用于处理部分可观测性下的连续动作空间问题,其中代理无法获得真实状态

决策时。最近,Gupta等人[60] 提出了 PS-TRPO 方法用于多智能体学习（见

算法1）。 该方法基于
TRPO,以便处理连续动作空间
有效地。
5)MADRL 的迁移学习:训练Q 网络
或者一般来说,单个代理的深度强化学习模型通常
计算成本非常高。这个问题
对于多代理系统来说,这是非常严重的。为了改善
为了提高多个深度强化学习模型的训练性能并降低计算成本,已有多项研究

促进了深度强化学习的迁移学习。Rusu等人[63],[83]
提出了一种策略蒸馏方法和渐进式神经网络
网络在促进迁移学习的背景下
深度强化学习。然而,这些方法计算复杂且成本高昂 [84]。Yin 和 Pan [85] 同样引入了
另一种应用知识的策略蒸馏架构
迁移到深度强化学习。该方法减少了训练时间
并优于 DQN,但其探索策略仍然
效率不高。Parisotto等人[86] 提出了演员模仿
一种多任务和迁移学习方法,可提高深度策略网络的学习速度。该网络可以获得

同时在许多游戏中表现出专家水平,尽管
它的模型并不那么复杂。但是该方法需要
源和目标之间具有足够的相似度
任务并且容易受到负转移的影响。
多智能体环境被重新表述为类似图像的表示,并在 [87] 中利用 CNN 来估计

针对每个相关代理的 Q 值。该方法可以解决
迁移学习中 MAS 的可扩展性问题
方法来加速训练过程。
在不同但相关的环境下训练的策略网络
用于其他代理的学习过程,以减少
计算费用。在
追踪-逃避问题[88]证明了
多智能体领域的迁移学习方法。
表三总结了所评论的文章,
解决不同的多智能体学习挑战。可以看出
文献中已经提出了许多 DQN 的扩展,而基于策略或 AC 的方法还没有得到充分的

已在多智能体环境中进行了探索。

B. MADRL 申请

由于深度强化学习的成功以 DQN 的提出为标志
在 [6] 中,已经提出了许多算法来整合深度
学习到多智能体学习。这些算法可以解决
各个领域的复杂问题。本节提供了
对这些应用的调查,重点是整合
深度学习和 MARL。表 IV 总结了这些特征
以及这些应用方法的局限性。
[91] 中引入了 MADRL 模型来处理
零能耗社区的能源共享问题
包括一系列零能耗建筑,
一年的总能源使用量小于或等于
每栋建筑内的可再生能源发电。深度 RL
代理用于表征每栋建筑,以学习与其他建筑共享能源的适当行动。

全局奖励由社区能量状态的负数建模,即奖励 = −($\sum_{i=1}^{n}$ c(hi) − g(hi)),
其中c(hi)和g(hi)分别是消耗的能量和能量
由第 i 栋建筑产生。社区监测
引入服务来管理群组成员活动,例如加入、离开群组或维护

活动代理列表。实验表明,与随机动作相比,所提出的模型具有优越性

净零能量平衡选择策略
社区。
分层 RL 与 MADRL 方法的组合
是为了协调和控制多个代理而开发的
优先考虑代理隐私的问题 [92]。此类分布式调度问题可能是多任务对话

自动助手需要帮助用户规划

表IV
不同领域中典型MADRL应用总结

Applications	Basic DLR	Features	Limitations
Energy sharing optimization [91]	DQN	•Each building is characterized by a DRL agent to learn appropriate actions independently. •Agents' actions include: consume and store excess energy, request neighbor or supply grid for additional energy, grant or deny requests. •Agents collaborate via shared or global rewards to achieve a common goal, *i.e.* zero-energy status.	•Agents' behaviors cannot be observed in an online fashion. •Limited number of houses, currently ten houses at maximum were experimented. •Energy price is not considered.
Federated control [92]	Hierarchical-DQN (h-DQN) [133]	•Divide the control problem into disjoint subtasks and leverage *temporal abstractions*. •Use *meta-controller* to guide decentralized controllers. •Able to solve distributed scheduling problems such as multi-task dialogue, urban traffic control.	•Does not address non-stationarity problem. •Number of agents is currently limited at six. •Meta-controller's optimal policy becomes complicated and inefficient when the number of agents increases.
Sequential social dilemma (SSD) [93]	DQN	•Introduce an SSD model to extend MGSD to capture sequential structure of real-world social dilemmas. •Describe SSDs as general-sum Markov games with partial observations. •Multi-agent DQN is used to find equilibria of SSD problems.	•Assume agent's learning is independent and regard the others as part of the environment. •Agents do not recursively reason about one another's learning.
Common-pool resource (CPR) appropriation [98]	DQN	•Introduce a new CPR appropriation model using an MAS containing spatially and temporally environment dynamics. •Use the *descriptive agenda* [134] to describe the behaviors emerging when agents learn in the presence of other learning agents. •Simulate multiple independent agents with each learned by DQN.	•Single agent DQN is extended to multi-agent environment where the Markov assumption is no longer hold. •Agents do not do anything of *rational negotiation*, e.g. bargaining, building consensus, or making appeals.
Swarm systems [100]	DQN and DDPG	•Agents can only observe local environment (partial observability) but not the global state. •Use guided approach for multi-agent learning where actors make decisions based on locally sensed information whilst critic has central access to global state.	•Can only work with homogeneous agents. •Unable to converge to meaningful policies in huge dimensionality and partial observed problem.
Traffic lights control [102]	DDDQN and IDQN	•Learning multiple agents is performed using IDQN where each agent is modelled by DDDQN. •First approach to address heterogeneous multi-agent learning in urban traffic control. •Fingerprint technique is used to stabilize the experience replay memory to handle non-stationarity.	•The proposed deep RL approach learns ineffectively in high traffic conditions. •The fingerprint does not improve the performance of experience replay although the latter is required for efficient learning.
Keepaway soccer [103]	DQN	•Low-dimensional state space, described by only 13 variables. •Heterogeneous MAS, each agent has different experience replay memory and different network policy.	•Number of agents is limited, currently setting with 3 keepers vs. 2 takers. •Heterogeneous learning speed is significantly lower than homogeneous case.
Task and resources allocation [105]	CommNet [70]	•Propose distributed task allocation where agents can request help from cooperating neighbors. •Three types of agents are defined: manager, participant and mediator. •Communication protocols are learned simultaneously with agents' policies through CommNet.	•May not be able to deal with heterogeneous agents. •Computational deficiencies regarding the decentralization and reallocation characteristics. •Experiments only on small state action spaces.
Large-scale fleet management [106]	Actor-critic and DQN	•Reallocate vehicles ahead of time to balance the transport demands and supplies. •Geographic context and collaborative context are integrated to coordinate agents. •Two proposed algorithms, *contextual multi-agent actor-critic* and *contextual deep Q-learning*, can achieve explicit coordination among thousands of agents.	•Can only deal with discrete actions and each agent has a small (simplified) action space. •Assume that agents in the same region at the same time interval (*i.e.* same spatial-temporal state) are homogeneous.
Action markets [107]	DQN	•Agents can exchange their atomic actions for environmental rewards. •Reduce greedy behavior and thus negative effects of individual reward maximization. •The proposed approach significantly increases the overall reward compared to methods without action trading.	•Agents cannot find prices for actions by themselves because they are given at design time. •Strongly assume that agents cannot make offers that they do not eventually hold.

完成几个独立的任务,例如购买一列火车
一张城市门票,订一张电影票,做一顿晚餐
餐厅预订。这些任务中的每一个都由
由分散控制器控制,而助手是元控制器,它受益于时间抽象来减轻
通信复杂性,从而能够为用户找到全局一致的解决方案（图 4）。另一方面

另一方面,Leibo等人[93] 引入了一个连续的社会困境
基于一般和马尔可夫博弈的（SSD)模型
部分可观测性,以解决合作的演变
MAS。SSD 能够捕捉现实世界社会困境的序列结构,是矩阵博弈的延伸
社会困境（MGSD)已被应用于社会科学和生物学中的各种现象[22],[52],[94]。

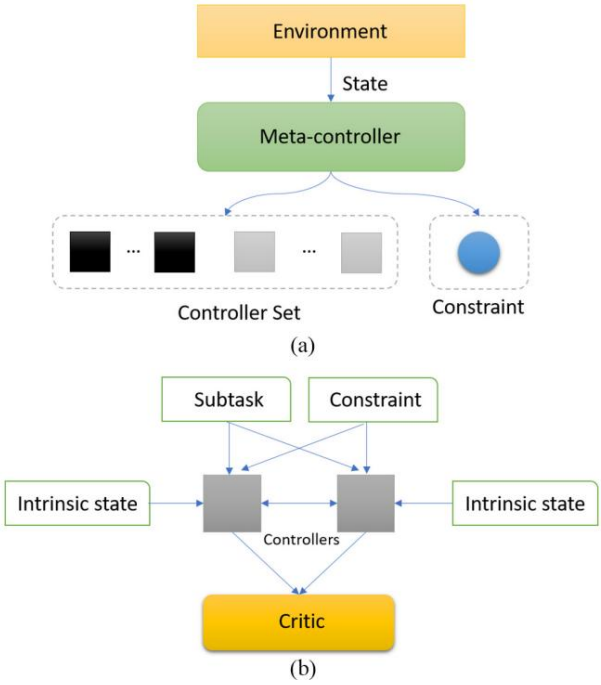

图 4. 采用分层 MADRL 方法的联合控制。 (a) 元控制器从环境接收状态并将子任务及其相关约束分配给控制器。 (b) 控制器具有环境的单独部分视图,但需要进行通信才能完成子任务。

一般和建模需要求解算法来追踪每个代理的不同潜在均衡,或者能够找到由使用不同状态空间扫描学习到的多个策略组成的循环策略 [95],[96]。DQN 用于表征自利独立学习代理,以找到 SSD 的均衡,而这无法通过用于 MGSD 的标准进化和学习方法来解决 [97]。Pérolat等人[98] 还展示了 MADRL 在社会科学现象中的应用,即公共池资源 (CPR) 占用。所提出的方法包括一个空间和时间动态的 CPR 环境 [99] 和一个由独立的自利 DQN 组成的 MAS。CPR 占用问题通过自组织来解决,自组织会随着时间的推移调整独立个体代理感受到的激励。

在 [100] 中,群体系统被表述为去中心化 POMDP [101] 的一个特例,并使用 AC 深度强化学习方法来控制一组合作的智能体。Q 函数是使用全局状态信息来学习的,在群体机器人示例中,全局状态信息可以是摄像机捕捉场景的视图。尽管单个智能体的感知能力有限,但该群体可以执行复杂的任务,例如搜索和救援或分布式组装。该模型有一个缺点,因为它假设智能体是同质的。[102] 提出了使用 IDQN 来解决城市交通信号灯控制多智能体环境中的异质性问题。每个智能体都通过决斗 DDQN (DDDQN) 进行学习,该网络集成了决斗网络、DDQN 和优先经验重放。将其他智能体视为环境的一部分,对异质智能体进行独立和同时的训练。多智能体环境的非平稳性通过以下技术来处理

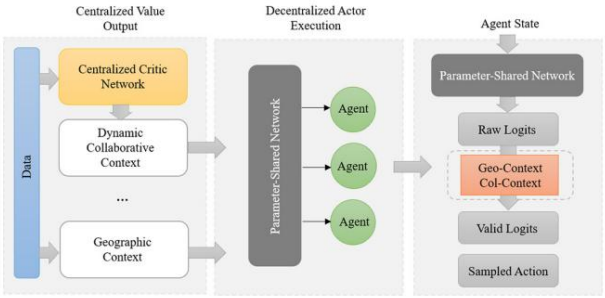

图 5. [106] 中提出的上下文多智能体 AC 架构,其中分散式执行通过集中式价值网络的输出进行协调,如左侧部分所示,而右侧部分则显示了上下文如何嵌入到策略网络中。

指纹识别可以消除训练样本的年龄歧义并稳定重放记忆。

[103] 介绍了 DQN 在状态空间为低维的异构 MAS 中的一种特殊应用。实验针对多智能体 Keepaway 足球问题进行,该问题的状态仅包含 13 个变量。为了处理异构性,每个 DQN 智能体都设置了不同的经验重放记忆和神经网络。

智能体之间无法相互沟通,只能观察他人的行为。虽然 DQN 可以在低维环境下的异质团队学习设置中提高游戏得分,但其学习过程明显慢于同质情况。

在学习过程中建立代理之间的通信渠道是设计和构建 MADRL 算法的重要步骤。Nguyen等[104] 通过图像表示的人类知识来表征通信渠道,并允许深度强化学习代理使用这些共享图像进行通信。异步优势 AC (A3C) 算法 [80] 用于学习每个代理的最优策略,该算法可以扩展到多个异构代理。另一方面,Noureddine等[105] 引入了一种方法,即使用协作深度强化学习的任务分配过程,以允许多个代理相互交互并有效分配资源和任务。代理可以在松散耦合的分布式多代理环境中向其协作邻居请求帮助。CommNet 模型 [70] 用于促进代理之间的通信,其特征是 DRQN [44]。

[106] 使用 MADRL 通过两种算法(即上下文深度 Q 学习和上下文多智能体 AC)解决了大规模车队管理问题。这些算法旨在通过重新分配运输资源来平衡需求和供应之间的差异,从而有助于减少交通拥堵并提高运输效率。

图 5 说明了上下文多智能体 AC 模型,其中使用参数共享策略网络来协调智能体,这些智能体代表可用车辆或空闲驾驶员。

最近,在 [107] 中引入了一种有趣的 MAS 方法,其中代理可以交易他们的行动以换取其他资源,例如环境奖励。行动交易的灵感来自福利经济学的基本定理,即竞争市场会朝着

帕累托效率。具体来说,智能体需要扩展其行动空间并同时学习两种策略:一种用于原始随机奖励,另一种用于交易环境奖励。从行动交易中实现的行为市场有助于缓解贪婪行为(如 [108] 中提出的针锋相对博弈论策略),使智能体能够激励其他智能体并减少个人奖励最大化的负面影响。

五、结论与研究方向

本文概述了多智能体学习中的不同挑战以及使用深度强化学习方法解决这些挑战的方法。我们将调查的文章分为五类,包括非平稳性、部分可观测性、多智能体训练方案、多智能体迁移学习以及连续状态和动作空间。我们强调了应对挑战的方法的优缺点。我们还全面回顾了 MADRL 方法在不同领域的应用。我们发现,将深度学习与传统的 MARL 方法相结合已经能够解决许多复杂问题,例如城市交通信号灯控制、零能耗社区中的能源共享问题、大规模车队管理、任务和资源分配、群体机器人以及社会科学现象。结果表明,基于深度强化学习的方法为处理 MAS 领域的复杂任务提供了一种可行的方法。

从示范中学习,包括模仿学习和逆向强化学习,在单智能体深度强化学习中是有效的 [109]。一方面,模仿学习试图将状态映射到动作,这是一种监督方法。它直接将专家策略推广到未访问的状态,以便在有限动作集的情况下接近多类分类问题。另一方面,逆向强化学习智能体需要从专家示范中推断出奖励函数。

逆强化学习假设专家策略在未知奖励函数方面是最优的 [110],[111]。然而,这些方法在多智能体环境中尚未得到充分探索。模仿学习和逆强化学习在 MAS 中都具有巨大的应用潜力。它们有望缩短学习时间并提高 MAS 的有效性。这些应用带来的一个非常直接的挑战是需要多位专家能够协作演示任务。此外,专家的沟通和推理能力很难由 MAS 领域的自主智能体来描述和建模。这些对将模仿学习和逆强化学习扩展到 MADRL 方法提出了重要的研究问题。此外,对于人类难以演示的复杂任务或行为,需要替代方法将人类偏好融入深度强化学习 [104],[112],[113]。

深度强化学习极大地促进了自主性,这使得许多应用程序可以在机器人或自动驾驶汽车中部署。然而,深度强化学习模型最常见的缺点是通过人机协作技术与人类互动的能力。在复杂和对抗的环境中,迫切需要人类智慧与技术相结合,因为只有人类无法维持这种规模,当引入新情况时,机器本身无法做出创造性的反应。人机回路架构 [114] 的最新进展可以与 MADRL 融合,将人类和自主代理整合在一起,以处理复杂问题。在人机回路中,代理自主执行任务直至完成,而处于监控或监督角色的人保留干预代理执行的操作的能力。如果人类监督者允许代理完全独立地完成任务,那么基于人机回路的架构可以完全自主 [114]。

无模型深度强化学习已经能够解决单智能体和多智能体领域的许多复杂问题。然而,该类方法需要大量样本和较长的学习时间才能取得良好的效果。

基于模型的方法在使用单智能体和多智能体模型解决各种问题时,在样本效率、可转移性和通用性方面都是有效的。

尽管最近在单智能体领域研究了基于模型的方法的深度学习扩展,例如 [115]–[120],但这些扩展在多智能体领域尚未得到广泛研究。这造成了一个研究空白,可以发展为基于模型的 MADRL 的研究方向。此外,使用基于模型的方法处理高维观测或结合基于模型的规划和无模型策略的元素是另一个活跃、令人兴奋但尚未得到充分探索的研究领域。

自诞生之日起,扩展到大型系统(尤其是处理许多异构代理)一直是强化学习研究领域面临的主要挑战。随着世界动态变得越来越复杂,这一挑战始终需要解决。由于代理具有共同的行为,例如动作、领域知识和目标(同质代理),因此可以通过(部分)集中训练和分散执行来实现可扩展性 [121],[122]。在具有许多代理的异构环境中,关键挑战是如何在代理之间采用有效的协调和合作策略,通过自学习提供最优解决方案并最大化任务完成成功率。解决这一难题的研究方向值得研究。

关于多智能体学习的应用,已经有许多研究使用传统的 MARL 方法来解决各种问题,例如控制一组自动驾驶汽车或无人机 [123]、机器人足球 [124]、控制交通信号 [125]、协调工厂和仓库中的协作机器人 [126]、控制电力网络 [127] 或优化分布式传感器网络 [128]、自动交易 [129]、竞争性电子商务和金融市场中的机器竞价 [130]、资源管理 [131] 和运输 [132]。自 DQN [6] 出现以来,文献中已经发现了将传统 RL 扩展到多智能体领域的深度 RL 的努力,但这些努力仍然非常有限(有关当前文献中可用的应用,请参阅表 IV)。现在,MADRL 可以基于其高维处理能力有效地解决 MARL 的许多应用。因此,需要进一步的实证研究,以应用 MADRL 方法有效地

解决上述应用等复杂的现实问题。

参考

[1] EL Thorndike，“动物智力:对动物相关过程的实验研究”，美国心理学,第 53 卷,第 10 期,第 1125-1127 页,1898 年。

[2] M L Minsky，《神经模拟强化系统理论及其在脑模型问题中的应用》，普林斯顿大学,美国新泽西州普林斯顿,1954 年。

[3] A. Klopf，《大脑功能和自适应系统:异质理论》,英国剑桥空军研究中心,1972 年。

[4] CJ Watkins 和 P. Dayan，《Q 学习》，《Mach. Learn.》,第 8 卷,第 3-4 期,第 279-292 页,1992 年。

[5] R.Bellman，“论动态规划理论”，Proc. Nat. Acad. Sci. USA,第 38 卷,第 8 期,第 716-719 页,1952 年。

[6] V. Mnih等人，“通过深度强化学习实现人类水平的控制”，《自然》,第 518 卷,第 7540 期,第 529-533 页,2015 年。

[7] S. Mahadevan 和 J. Connell，“使用强化学习实现基于行为的机器人自动编程”，Artif. Intell.,第 55 卷,第 2-3 期,第 311-365 页,1992 年。

[8] S. Schaal，“从示范中学习”，Proc。副词。神经信息。Process. Syst. 1997,第 1040-1046 页。

[9] H. Benbrahim 和 JA Franklin，“利用强化学习实现双足动态行走”，Robot. Auton. Syst.,第 22 卷,第 3-4 期,第 283-302 页,1997 年。

[10] M. Riedmiller、T. Gabel、R. Hafner 和 S. Lange，《机器人足球的强化学习》，《Auton. Robots》,第 27 卷,第 1 期,第 55-73 页,2009 年。

[11] K. Mulling、J. Kober、O. Kroemer 和 J. Peters，“学习选择和概括机器人乒乓球的击球动作”，Int. J. Robot。研究,卷。32、没有。第 3 页。263-279,2013。

[12] MG Bellemare、Y. Naddaf、J. Veness 和 M. Bowling，“街机学习环境:面向一般代理的评估平台”，J. 艺术家英特尔。研究,卷。47,页。253–279,2013 年 5 月。

[13] D. Silver等人，“利用深度神经网络和树搜索掌握围棋游戏”，《自然》,第 529 卷,第 7587 期,第 484-489 页,2016 年。

[14] Y. Duan、X. Chen、R. Houthooft、J. Schulman 和 P. Abbeel，“针对连续控制的深度强化学习基准测试”，载于Proc. Int. Conf. Mach. Learn.，2016 年 6 月,第 1329-1338 页。

[15] C. Beattie等人，“DeepMind 实验室”,2016 年。[在线]。可访问:arXiv:1612.03801。

[16] B. Luo、D. Liu 和 H.-N. Wu，“具有仅批评结构的基于数据的非线性离散时间系统的自适应约束最优控制设计”，IEEE 神经网络学习系统汇刊,第 29 卷,第 6 期,第 2099-2111 页,2017 年 6 月。

[17] Z. Wan, C. Jiang, M. Fahad, Z. Ni, Y. Guo, 和 H. He，“基于深度强化学习的机器人辅助行人监管”，IEEE Trans. Cybern.,第 50 卷,第 4 期,第 1669-1682 页,2020 年 4 月,doi：10.1109/TCYB.2018.2878977。

[18] T. Tan, F. Bao, Y. Deng, A. Jin, Q. Dai, J. Wang，“协同深度强化学习在大规模交通网格信号控制中的应用”，IEEE Trans. Cybern.,早期访问,doi：10.1109/TCYB.2019.2904742。

[19] M. Nazari、A. Oroojlooy、LV Snyder 和 M. Takác，“强化学习用于解决车辆路线问题”,载于Proc. Adv. Neural Inf. Process. Syst.，2018 年,第 9839-9849 页。

[20] W. Kool、H. van Hoof 和 M. Welling，“注意,学会解决路由问题！”2018 年。[在线]。可用:arXiv:1803.08475。

[21] L. Busoniu、R. Babuska 和 B. De Schutter，“多智能体强化学习综合调查”，IEEE 系统、人与控制论汇刊。C, Appl. Rev.,第 38 卷,第 2 期,第 156-172 页,2008 年 3 月。

[22] D. Bloembergen、K. Tuyls、D. Hennes 和 M. Kaisers，“多智能体学习的进化动力学:一项调查”，J. Artif。英特尔。研究,卷。53,页。659–697,1999 年 8 月；2015年。

[23] P. Hernandez-Leal、B. Kartal 和 ME Taylor，“多智能体深度强化学习是答案还是问题?简要调查”，2018 年。[在线]。可访问网址:arXiv:1810.05587。

[24] FL da Silva、ME Taylor 和 AHR Costa，“在多智能体强化学习中自主重用知识”,载于第 27 届国际人工智能联合会议论文集，2018 年,第 5487-5493 页。

[25] RS Sutton 和 AG Barto，《强化学习:导论》。美国马萨诸塞州剑桥:麻省理工学院出版社,1998 年。

[26] B. Luo、D. Liu、H.-N. Wu、D. Wang 和 FL Lewis，“基于数据的最优控制的策略梯度自适应动态规划”，IEEE 控制论汇刊,第 47 卷,第 10 期,第 3341-3354 页,2017 年 10 月。

[27] KG Vamvoudakis、FL Lewis 和 GR Hudas，“多智能体差异图形游戏:用于优化同步的在线自适应学习解决方案”，Automatica,第 48 卷,第 8 期,第 1598-1611 页,2012 年。

[28] H. Zhang、H. Jiang、C. Luo 和 G. Xiao，“使用基于策略迭代的自适应动态规划算法的多人离散时间非零和游戏”，IEEE 控制论汇刊,第 47 卷,第 10 期,第 3331-3340 页,2017 年 10 月。

[29] W. Gao、ZP Jiang、FL Lewis 和 Y. Wang，“使用自适应动态规划实现多智能体系统的合作最优输出调节”，载于Proc. Amer. Control Conf. (ACC)，2017 年 5 月,第 2674-2679 页。

[30] J. Zhang、H. Zhang 和 T. Feng，“具有未知动态的非线性多智能体系统的分布式最优共识控制”，IEEE 神经网络学习系统汇刊,第 29 卷,第 8 期,第 3339-3348 页,2018 年 8 月。

[31] 张红,苏红,张可,罗毅，“基于广义模糊双曲模型的未知非线性系统非零和博弈事件触发自适应动态规划算法”，IEEE Trans. Fuzzy Syst.,第 27 卷,第 11 期,第 2202-2214 页,2019 年 11 月,doi：10.1109/TFUZZ.2019.2896544。

[32] JN Tsitsiklis 和 B. Van Roy，“通过函数逼近分析时间差分学习”,载于Proc. Adv. Neural Inf. Process。Syst. 1997,第 1075-1081 页。

[33] B. Luo, D. Liu, T. Huang 和 D. Wang，“通过仅评判Q 学习实现无模型最优跟踪控制”，IEEE 神经网络学习汇刊。Syst.,第 27 卷,第 10 期,第 2134-2144 页,2016 年 10 月。

[34] K. Arulkumaran、MP Deisenroth、M. Brundage 和 AA Bharath，“深度强化学习:简要调查”，IEEE 信号过程。杂志,卷。34、没有。第 6 页。11 月 26 日至 38 日2017年。

[35] Y. Li，“深度强化学习:概述”,2017 年。[在线]。可获取于:arXiv:1701.07274。

[36] ND Nguyen、T. Nguyen 和 S. Nahavandi，“使用深度强化学习的人类级代理的系统设计视角:一项调查”，IEEE Access,第 5 卷,第 27091-27102 页,2017 年。

[37] A. Krizhevsky、I. Sutskever 和 GE Hinton，《使用深度卷积神经网络进行 ImageNet 分类》，《Proc. Adv. Neural Inf.》Process. Syst. 2012,第 1097-1105 页。

[38] T. Nguyen，“多目标深度强化学习框架”,2018 年。[在线]。可访问:arXiv:1803.02965。

[39] B. Luo、Y. Yang 和 D. Liu，“基于经验重放的基于数据的最优输出调节的自适应 Q 学习”，IEEE Trans. Cybern.,第 48 卷,第 12 期,第 3337-3348 页,2018 年 12 月。

[40] HV Hasselt，《双Q 学习》，《神经信息处理高级期刊》Syst.，2010 年,第 2613-2621 页。

[41] HV Hasselt、A. Guez 和 D. Silver，“通过双Q 学习进行深度强化学习”,载于第 30 届 AAAI Conf. Artif. Intell. 会议论文集，2016 年 2 月,第 2094-2100 页。

[42] T. Schaul、J. Quan、I. Antonoglou 和 D. Silver，“优先经验重播”,2015 年。[在线]。可访问:arXiv:1511.05952。

[43] Z. Wang、T. Schaul、M. Hessel、H. Hasselt、M. Lanctot 和 ND Freitas，“深度强化学习的决斗网络架构”，载于Proc. Int. Conf. Mach. Learn.，2016 年 6 月,第 1995-2003 页。

[44] MJ Hausknecht 和 P. Stone，“深度循环Q 学习用于部分可观察的 MDP”,载于Proc. AAAI 秋季研讨会系列，2015 年 9 月,第 29-37 页。

[45] G. Lample 和 DS Chaplot，“通过深度强化学习玩 FPS 游戏”,载于第 31 届 AAAI Conf. Artif. Intell. 会议论文集，2017 年 2 月,第 2140-2146 页。

[46] I. Sorokin、A. Seleznev、M. Pavlov、A. Fedorov 和 A. Ignateva，“深度注意循环Q 网络”,2015 年。[在线]。可访问网址:arXiv:1512.01693。

[47] P. Hernandez-Leal、M. Kaisers、T. Baarslag 和 EM de Cote，“多智能体环境中的学习调查:处理非平稳性”,2017 年。[在线]。可访问网址:arXiv:1707.09183。

[48] M.Tan，“多智能体强化学习:独立智能体与合作智能体”,载于第 10 届国际机器学习会议论文集，1993 年,第 330-337 页）。

[49] AO Castaneda，“多智能体学习算法的深度强化学习变体”,硕士论文,爱丁堡大学信息学院,英国爱丁堡,2016 年。

[50] S. Abdallah 和 M. Kaisers，“通过重复更新解决 Q 学习的策略偏差”,载于第 12 届国际会议论文集 Auton. Agents Multiagent Syst.，2013 年 5 月,第 1045-1052 页。

[51] S. Abdallah 和 M. Kaisers，“通过重复Q 学习更新解决环境非平稳性问题”，J. Mach. Learn. Res.,第 17 卷,第 1 期,第 1582-1612 页,2016 年。

[52] C. Yu, M. Zhang, F. Ren, 和 G. Tan, “空间社交困境中的情感多智能体强化学习”, IEEE 神经网络汇刊。
Learn. Syst.,第 26 卷,第 12 期,第 3083-3096 页,2015 年 12 月。
[53] EAO Diallo、A. Sugiyama 和 T. Sugawara,“学习在双打乒乓游戏中利用深度强化学习进行协调”,载于第 16 届 IEEE 国际会议机器学习与应用会议 (ICMLA), 2017 年 12 月,第 14-19 页。

[54] J. Foerster等人, “深度多智能体强化学习的稳定经验重放”,载于Proc. Int. Conf. Mach. Learn., 2017 年 7 月,第 1146-1155 页。

[55] G. Palmer、K. Tuyls、D. Bloembergen 和 R. Savani,“宽松的多智能体深度强化学习”,载于Proc. Int. Conf. Auton. Agents Multiagent Syst., 2018 年 7 月,第 443-451 页。

[56] L. Panait、K. Tuyls 和 S. Luke,“宽容学习者的理论优势:进化博弈论视角”, J. Mach. Learn。
研究,卷。第 9 页。 423–457,2008 年 3 月。
[57] S. Omidshafiei、J. Pazis、C. Amato、JP How 和 J. Vian,“部分可观测性下的深度分散多任务多智能体强化学习”,载于Proc. Int. Conf. Mach. Learn., 2017 年 7 月,第 2681-2690 页。

[58] Y. Zheng、Z. Meng、J. Hao 和 Z. Zhang,“随机合作环境中的加权双深度多智能体强化学习”,载于Proc. Pac. Rim Int. Conf. Artif. Intell., 2018 年 8 月,第 421-429 页。

[59] JN Foerster、YM Assael、N. de Freitas 和 S. Whiteson,“学习通过深度分布式循环Q 网络进行沟通以解答谜题”,2016 年。[在线]。可访问网址:arXiv:1602.02672。

[60] JK Gupta、M. Egorov 和 MJ Kochenderfer,“使用深度强化学习实现合作多智能体控制”,载于Proc. Int. Conf.
Auton. Agents Multiagent Syst., 2017 年 5 月,第 66-83 页。
[61] Z.-W. Hong、S.-Y. Su、T.-Y. Shann、Y.-H. Chang 和 C.-Y. Lee,“用于多智能体系统的深度策略推理 Q 网络”,载于第 17 届国际会议论文集 Auton. Agents Multiagent Syst., 2018 年 7 月,第 1388-1396 页。

[62] L. Matignon、G. Laurent 和 NL Fort-Piat,“滞后Q 学习:一种用于合作多智能体团队的分散强化学习的算法”, IEEE/RSJ 国际智能机器人系统会议论文集, 2007 年 10 月,第 64-69 页。

[63] AA Rusu等, “政策蒸馏”,2015 年。[在线]。可访问网址:arXiv:1511.06295。

[64] O. Kilinc 和 G. Montana,“具有极度噪声观测的多智能体深度强化学习”,2018 年。[在线]。可访问网址:arXiv:1812.00922。

[65] JN Foerster等人, “深度多智能体强化学习的贝叶斯动作解码器”,2018 年。[在线]。可访问网址:arXiv:1811.01458。
[66] A.Tampuu等人, “深度强化学习的多智能体合作与竞争”, PLoS ONE,第 12 卷,第 4 期,2017 年,文章编号 e0172395。

[67] M. Lanctot等人, “多智能体强化学习的统一博弈论方法”,载于Proc. Adv. Neural Inf. Process. Syst., 2017 年,第 4193-4206 页。

[68] L. Kraemer 和 B. Banerjee,“多智能体强化学习作为分散规划的演练”, 《神经计算》,第 190 卷,第 82-94 页,2016 年 5 月。

[69] JN Foerster、YM Assael、N. de Freitas 和 S. Whiteson,“学习通过深度多智能体强化学习进行通信”,载于Proc. Adv. Neural Inf. Process. Syst., 2016 年,第 2137-2145 页。

[70] S. Sukhbaatar、A. Szlam 和 R. Fergus,“通过反向传播学习多智能体通信”,载于Proc. Adv. Neural Inf. Process。
Syst., 2016 年,第 2244-2252 页。
[71] H. He、JL Boyd-Graber、K. Kwok 和 H. Daumé III,《深度强化学习中的对手建模》,载于Proc. Int. Conf. Mach.
学习, 2016 年 6 月,第 1804-1813 页。
[72] X. Kong、B. Xin、F. Liu 和 Y. Wang,“多智能体深度强化学习系统中的有效主从通信”,载于第 31 届分层强化学习研讨会 (NIPS) 论文集, 2017 年,第 1-6 页。

[73] X. Kong, B. Xin, F. Liu, 和 Y. Wang,“重新审视多智能体深度强化学习中的主从架构”,2017 年。
[在线]。可访问网址:arXiv:1712.07305。
[74] R. Lowe、Y. Wu、A. Tamar、J. Harb、P. Abbeel 和 I. Mordatch,“面向混合合作竞争环境的多智能体参与者-评论家”,载于Proc. Adv. Neural Inf. Process. Syst., 2017 年,第 6379-6390 页。

[75] JN Foerster、G. Farquhar、T. Afouras、N. Nardelli 和 S. Whiteson,“反事实多主体政策梯度”,载于Proc。第22届AAAI会议阿蒂夫。情报, 2018,第 101-1 页2974–2982。

[76] A. Harati、MN Ahmadabadi 和 BN Araabi,“基于知识的多智能体信用分配:关于任务类型和评论信息的研究”, IEEE Syst. J.,第 1 卷,第 1 期,第 55-67 页,2007 年 9 月。

[77] TP Lillicrap等人, “深度强化学习的持续控制”,2015 年。[在线]。可访问网址:arXiv:1509.02971。

[78] J. Schulman、S. Levine、P. Abbeel、M. Jordan 和 P. Moritz,“信任区域策略优化”,载于Proc. Int. Conf. Mach. Learn., 2015 年 6 月,第 1889-1897 页。

[79] VR Konda 和 JN Tsitsiklis,“演员-评论家算法”,
Adv. Neural Inf. Process. Syst., 2000 年,第 1008-1014 页。
[80] V. Mnih等人, “深度强化学习的异步方法”,载于Proc. Int. Conf. Mach. Learn., 2016 年 6 月,第 1928-1937 页。
[81] D. Silver、G. Lever、N. Heess、T. Degris、D. Wierstra 和 M. Riedmiller,“确定性策略梯度算法”,载于Proc. Int.
Conf. Mach. Learn., 2014 年 1 月,第 387-395 页。
[82] N. Heess、JJ Hunt、TP Lillicrap 和 D. Silver,“基于记忆的循环神经网络控制”,2015 年。[在线]。可访问网址:arXiv:1512.04455。

[83] AA Rusu等人, “渐进式神经网络”,2016 年。[在线]。
可获取于:arXiv:1606.04671。
[84] J. Kirkpatrick等人, “克服神经网络中的灾难性遗忘”,美国自然科学院院刊,第 114 卷,第 13 期,第 3521-3526 页,2017 年。

[85] H. Yin 和 SJ Pan,“通过分层经验重放实现深度强化学习的知识转移”,第 31 届 AAAI 会议论文集。
艺术家英特尔,一月2017 年,第 17 页。 1640–1646。
[86] E. Parisotto、JL Ba 和 R. Salakhutdinov,“Actor–MIMIC:深度多任务和迁移强化学习”,2015 年。[在线]。可访问网址:arXiv:1511.06342。

[87] M. Egorov,多智能体深度强化学习,斯坦福大学,美国加利福尼亚州斯坦福,2016 年。

[88] TH Chung、GA Hollinger 和 V. Isler,“移动机器人中的搜索与追击-规避”, Auton. Robots,第 31 卷,第 4 期,第 299 页,2011 年。
[89] P. Zhu、X. Li 和 P. Poupart,“论改进 POMDP 的深度强化学习”,2017 年。[在线]。可访问网址:arXiv:1704.07978。
[90] D. Luviano-Cruz、F. Garcia-Luna、L. Perez-Dominguez 和 S. Gadi,“使用线性模糊模型的多智能体强化学习应用于合作移动机器人”, 《对称性》,第 10 卷,第 10 期,第 461 页,2018 年。

[91] A. Prasad 和 I. Dusparic,“面向零能耗社区的多智能体深度强化学习”,2018 年。[在线]。可访问网址:arXiv:1810.03679。

[92] S Kumar、P. Shah、D. Hakkani-Tur 和 L. Heck,“基于分层多智能体深度强化学习的联邦控制”,2017 年。
[在线]。可访问网址:arXiv:1712.08266。
[93] JZ Leibo、V. Zambaldi、M. Lanctot、J. Marecki 和 T. Graepel,“连续社交困境中的多智能体强化学习”,载于第 16 届国际会议论文集 Auton. Agents Multiagent Syst., 2017 年 5 月,第 464-473 页。

[94] EM de Cote、A. Lazaric 和 M. Restelli,“学习在多智能体社交困境中合作”,载于Proc. ACM 第五届国际联合会议 Auton。
Agents Multiagent Syst., 2006 年 5 月,第 783-785 页。
[95] M. Zinkevich、A. Greenwald 和 ML Littman,“马尔可夫博弈中的循环均衡”,载于Proc. Adv. Neural Inf. Process. Syst., 2006 年,第 1641-1648 页。

[96] J. Perolat、B. Piot、B. Scherrer 和 O. Pietquin,“论使用非平稳策略解决双人零和马尔可夫博弈”,载于第 19 届国际人工智能统计会议论文集, 2016 年 5 月,第 893-901 页。

[97] M. Kleiman-Weiner、MK Ho、JL Austerweil、ML Littman 和 J. Tenenbaum,“协调合作或竞争:社会互动中的抽象目标和共同意图”,第 38 届年度会议纪要。
Cogn. Sci. Soc., 2016 年 1 月,第 1679-1684 页。
[98] J. Pérolat、JZ Leibo、VF Zambaldi、C. Beattie、K. Tuyls 和 T. Graepel,“公共池资源占用的多智能体强化学习模型”,载于Proc. Adv. Neural Inf. Process. Syst., 2017 年,第 3643-3652 页。

[99] MA Janssen、R. Holahan、A. Lee 和 E. Ostrom,“社会生态系统研究的实验室实验”, 《科学》,第 328 卷,第 5978 期,第 613-617 页,2010 年。

[100] M. Huttenrauch、A. Sosic 和 G. Neumann,“引导式深度强化学习用于群体系统”,2017 年。[在线]。可访问网址:arXiv:1709.06011。

[101] FA Oliehoek,“去中心化 POMDP”,强化学习。
德国柏林:Springer,2012 年,第 471-503 页。
[102] JA Calvo 和 I. Dusparic,“异构多智能体深度强化学习用于交通信号灯控制”,第 26 届爱尔兰会议论文集。
艺术家英特尔。认知。科学, 2018,第1-12。

[103] M. Kurek 和 W. Jaskowski，“低维多智能体环境中的异构团队深度 Q 学习”， IEEE Conf. 帐户英特尔。游戏 (CIG)， 9 月2016 年,第 17 页。 1-8。

[104] T. Nguyen、ND Nguyen 和 S. Nahavandi，“具有人类策略的多智能体深度强化学习”,2018 年。[在线]。可访问网址:arXiv:1806.04562。

[105] D. Noureddine、A. Gharbi 和 S. Ahmed，“多智能体深度强化学习在动态环境中的任务分配”，第 12 届国际软件技术会议论文集 (ICSOFT)， 2017 年,第 17-26 页。

[106] K. Lin, R. Zhao, Z. Xu, 和 J. Zhou, “通过多智能体深度强化学习实现高效的大规模车队管理”, 2018 年。[在线]。可获取于:arXiv:1802.06444。

[107] K. Schmid、L. Belzner、T. Gabor 和 T. Phan，“深度多智能体强化学习中的动作市场”,载于 Proc. Int. Conf. Artif. Neural Netw.， 2018 年 10 月,第 240-249 页。

[108] A. Lerer 和 A. Peysakhovich，“利用深度强化学习在复杂的社会困境中维持合作”,2017 年。[在线]。可获取于:arXiv:1707.01068。

[109] B. Piot、M. Geist 和 O. Pietquin，“弥合模仿学习与逆向强化学习之间的差距”， IEEE 神经网络汇刊。Learn. Syst.,第 28 卷,第 8 期,第 1814-1826 页,2017 年 8 月。

[110] D. Hadfield-Menell、SJ Russell、P. Abbeel 和 A. Dragan，“合作逆向强化学习”,载于Proc. Adv. Neural Inf. Process. Syst.， 2016 年,第 3909-3917 页。

[111] D. Hadfield-Menell、S. Milli、P. Abbeel、SJ Russell 和 A. Dragan，“逆向奖励设计”，载于Proc. Adv. Neural Inf. Process. Syst.， 2017 年,第 6765-6774 页）。

[112] PF Christiano、J. Leike、T. Brown、M. Martic、S. Legg 和 D. Amodei，“从人类偏好中进行深度强化学习”,载于Proc. Adv. Neural Inf. Process. Syst.， 2017 年,第 4299-4307 页。

[113] ND Nguyen、S. Nahavandi 和 T. Nguyen，“深度强化学习的人性化混合策略方法”， IEEE Int. Conf. Syst. 人控制论， 2018 年,第 4023-4028 页。

[114] S. Nahavandi，“人与机器人之间的可信自主性:迈向机器人和自主系统中的人机交互”， IEEE Trans. Syst., Man, Cybern., Syst.,第 3 卷,第 1 期,第 10-17 页,2017 年 1 月。

[115] S. Levine、C. Finn、T. Darrell 和 P. Abbeel，“深度视觉运动策略的端到端训练”， J. Mach. Learn. Res.,第 17 卷,第 1 期,第 1334-1373 页,2016 年。

[116] S. Gu、TP Lillicrap、I. Sutskever 和 S. Levine，“基于模型加速的持续深度 Q 学习”,载于Proc. Int. Conf. Mach. 学习， 2016 年 6 月,第 2829-2838 页。

[117] C. Finn 和 S. Levine，“深度视觉预见用于规划机器人运动”,载于IEEE 国际机器人与自动化会议论文集 (ICRA)， 2017 年 5 月,第 2786-2793 页。

[118] A. Nagabandi、G. Kahn、RS Fearing 和 S. Levine，“基于模型的深度强化学习与无模型微调的神经网络动力学”,载于IEEE 国际机器人与自动化会议论文集 (ICRA)， 2018 年 5 月,第 7559-7566 页。

[119] IV Serban、C. Sankar、M. Pieper、J. Pineau 和 Y. Bengio，“瓶颈模拟器:基于模型的深度强化学习方法”,2018 年。[在线]。可用:arXiv:1807.04723。

[120] D. Corneil、W. Gerstner 和 J. Brea，“基于变分状态制表的高效模型深度强化学习”,2018 年。[在线]。可获取于:arXiv:1802.04325。

[121] T. Rashid、M. Samvelyan、CS de Witt、G. Farquhar、J. Foerster 和 S. Whiteson，“QMIX:深度多智能体强化学习的单调值函数分解”,2018 年。[在线]。可访问网址:arXiv:1803.11485。

[122] J. Foerster等人， “通过对手学习意识进行学习”,载于第 17 届国际会议论文集 Auton. Agents Multiagent Syst.， 2018 年 7 月,第 122-130 页。

[123] S.-M. Hung 和 SN Givigi，“随机环境下无人机集群的 Q 学习方法” ,IEEE 控制论汇刊,第 47 卷,第 1 期,第 186-197 页,2017 年 1 月。

[124] D. Schwab、Y. Zhu 和 M. Veloso，“机器人足球的零样本迁移学习”,载于第 17 届国际自动代理多代理系统会议论文集， 2018 年 7 月,第 2070-2072 页。

[125] SS Mousavi、M. Schukat 和 E. Howley，“使用深度策略梯度和基于价值函数的强化学习进行交通灯控制”，IET 英特尔。运输。系统,卷。 11、没有。第 7 页。 417-423,2017。

[126] A. Kattepur、HK Rath、A. Simha 和 A. Mukherjee，“面向工业 4.0 仓库的多智能体机器人分布式优化”,载于第 33 届 ACM 应用计算研讨会文集， 2018 年 4 月,第 808-815 页。

[127] MS Rahman、MA Mahmud、HR Pota、MJ Hossain 和 TF Orchi，“基于分布式多代理的电力系统暂态稳定性增强保护方案”， Int. J. Emerg. Elect. Power Syst.,第 16 卷,第 2 期,第 117-129 页,2015 年。

[128] J. He, J. Peng, F. Jiang, G. Qin 和 W. Liu，“认知无线电传感器网络的分布式 Q 学习频谱决策方案”， Int. J. Distrib. Sensor Netw.,第 11 卷,第 5 期,2015 年,货号 301317。

[129] PC Pendharkar 和 P. Cusatis，“利用强化学习代理交易金融指数”， Expert Syst. Appl.,第 103 卷,第 1-13 页,2018 年 8 月。

[130] O. Brandouy、P. Mathieu 和 I. Veryzhenko，“论基于代理的人工股票市场的设计”,载于 Proc. Int. Conf. Agents Artif. Intell.， 2011 年 1 月,第 350-364 页。

[131] M. Hussin、NAWA Hamid 和 KA Kasmiran，“通过分布式系统的自适应强化学习提高资源管理的可靠性”,并行分布计算杂志,第 75 卷,第 93-100 页,2015 年 1 月。

[132] B. Fernandez-Gauna、I. Etxeberria-Agiriano 和 M. Grana，“通过分布式循环Q 学习来学习多机器人软管运输和部署”， PLoS ONE,第 10 卷,第 7 期,2015 年,文章编号 e0127129。

[133] TD Kulkarni、K. Narasimhan、A. Saeedi 和 J. Tenenbaum，《分层深度强化学习:整合时间抽象和内在动机》，《 Proc. Adv. Neural Inf. Process》 Syst.， 2016 年,第 3675-3683 页。

[134] Y. Shoham、R. Powers 和 T. Grenager，“如果多智能体学习是答案,那么问题是什么？” Artif. Intell.,第 171 卷,第 7 期,第 365-377 页,2007 年。

Thanh Thi Nguyen于 2013 年获得澳大利亚维多利亚州墨尔本莫纳什大学数学和统计学博士学位。

2015 年,他曾任美国斯坦福大学计算机科学系访问学者,2019 年，他曾任美国马萨诸塞州剑桥市哈佛大学约翰·保尔森工程与应用科学学院边缘计算实验室访问学者。他目前是澳大利亚维多利亚州伯伍德市迪肯大学信息技术学院的高级讲师。他在人工智能、深度学习、深度强化学习、网络安全、物联网和数据科学等多个领域拥有专业知识。

阮博士于 2016 年获得阿尔弗雷德·迪肯博士后研究奖学金,并于 2018 年获得欧盟委员会颁发的欧洲太平洋伙伴关系信息和通信技术专家交流计划奖。

Ngoc Duy Nguyen于 2011 年获得韩国水原成均馆大学计算机工程硕士学位。他目前正在澳大利亚维多利亚州沃恩庞兹迪肯大学智能系统研究与创新研究所攻读博士学位。

他的研究兴趣包括机器学习、优化问题和系统设计。

阮先生荣获信息系最佳硕士论文奖成均馆大学通信工程系。

Saeid Nahavandi （IEEE 高级会员)于 1991 年获得英国达勒姆大学博士学位。

他是阿尔弗雷德·迪肯教授、副校长（国防技术）、工程系主任以及澳大利亚维多利亚州沃恩庞兹迪肯大学智能系统研究与创新研究所所长。他在各种国际期刊和会议上发表了 600 多篇论文。他的研究兴趣包括复杂系统建模、机器人技术和触觉。

Nahavandi 教授是 IEEE SYSTEMS JOURNAL 的联合主编、 IEEE/ASME TRANSACTIONS ON MECHATRONICS和 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS的副主编以及 IEEE ACCESS的编委会成员。

他是澳大利亚工程师协会和工程技术学会的会员。