多智能体强化学习研究综述

1. 简介：包括其背景、发展趋势、热点等进行简述

多智能体系统（Multi-agent System, MAS）是多个智能体组成的集合，其目标是将大而复杂的系统建设成小而彼此互相通信协调的易于管理的系统。多智能体系统自20世纪70年代被提出以来，就在智能机器人、交通控制、分布式决策、商业管理、软件开发、虚拟现实等各个领域迅速地得到了应用，目前已经成为一种对复杂系统进行分析与模拟的工具。多智能体系统由分布式人工智能演化而来，其研究目的是解决大规模的、复杂的现实问题。在现实问题中，单智能体的决策能力远远不够。使用一个中心化的智能体解决问题时，会遇到各种资源和条件的限制，导致单个智能体无法应对错综复杂的现实环境；而使用多个智能体相互协作可以解决很多问题[1]。

强化学习（Reinforcement Learning, RL）是机器学习的一种重要方法，它是一种以环境反馈作为输入目标，用试错方法发现最优行为策略的学习方法。在强化学习的数学基础研究取得了突破性的进展后，对强化学习的研究和应用日益增多[2]。目前，强化学习已被广泛应用于手工业制造、机器人控制、优化与调度、仿真模拟、游戏博弈等领域[3]。目前，结合多智能体系统和强化学习方法形成的多智能体强化学习正逐渐成为强化学习领域的研究热点之一，并在各个领域得到广泛应用[4-6]。

多智能体强化学习（Multi-agent Reinforcement Learning , MARL）是将强化学习的思想和算法应用到多智能体系统中。20世纪90年代，Littman[7]提出了以马尔可夫决策过程（Markov Decision Process, MDP）为环境框架的MARL，为解决大部分强化学习问题提供了一个简单明确的数学框架，后来研究者们大多在这个模型的基础上进行了更进一步的研究。最近，随着深度学习的成功，人们将深度学习的方法与传统的强化学习算法相结合，形成了许多深度强化学习算法，使单智能体强化学习的研究和应用得到迅速发展。比如，DeepMind公司研制出的围棋博弈系统AlphaGO已经在围棋领域战胜了人类顶级选手，并以较大优势取得了胜利，这极大地震撼了社会各界[8]，也促使研究人员在多智能体强化学习领域投入更多的精力。以DeepMind, Open AI公司为代表的企业和众多高校纷纷开发MARL的新算法，并将其应用到实际生活中，目前主要应用于机器人系统[9-10]、人机对弈[11-13]、自动驾驶[1-4]、互联网广告[15]和资源利用[16-17]等领域。

本文第 2 节介绍了多智能体强化学习算法的分类方法，并对各分类下的代表性算法进行了探讨；第 3 节对多智能体强化学习目前遇到的挑战进行了概括；第 4 节对多智能体强化学习在现实领域中的应用和前景进行了探讨；最后总结全文。

2. 算法分类

本节遵循综述 Is multiagent deep reinforcement learning the answer or the question? A brief survey对多智能体强化学习算法的分类方法，将 MARL 算法分为四类。下面按一一进行简单讲解。

- Analysis of emergent behaviors（行为分析）

行为分析类别的算法主要是将单智能体强化学习算法（SARL）直接应用到多智能体环境之中，每个智能体之间相互独立，遵循 Independent Q-Learning [2] 的算法思路。这类作品比较早，这里主要讨论以下两部作品。

在[4]中，作者首次将DQN算法与IQL结合，并将其应用于ALE环境[5]中的Pong游戏。为了全面观察DQN应用于多智能体环境的性能，作者通过设计奖励函数设计了完全协作环境、完全竞争环境和不完全协作/竞争环境。最终的实验结果表明，在完全协作环境中，智能体学到的策略是尽可能长时间的不失球；而在完全竞争环境中，智能体学到的是如何更好的得分（即让对方失球）。

从这个结果可以看出，在将 DQN 直接应用到多智能体环境中，也能够达到一个比较好的性能，即便 IQL 算法是一个十分简单的算法，没有办法处理环境非平稳问题，但是依旧是一个比较强的基准算法。

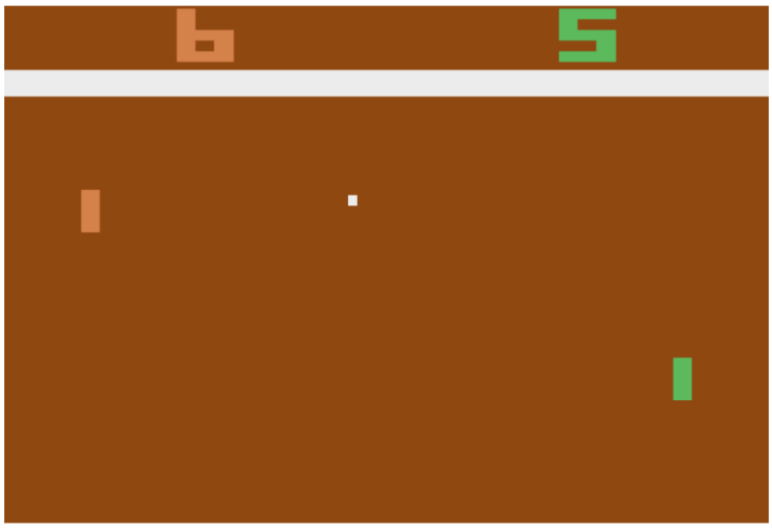


Figure 1: The *Pong* game. Each agent corresponds to one of the paddles.

在[3]中，作者将基于值函数的算法 DQN、基于策略梯度的算法 TRPO 以及演员-评论家算法 DDPG 与 IQL 算法以及循环神经网络（或前向神经网络）相结合，应用到局部观察的多智能体环境中。与前文不同的是，为了提高算法在大规模场景下的可扩展性，所有的agent共享同一组参数。对所有agent在训练过程中采集的样本进行汇总，用于更新共享模型参数。

实验结果表明，在使用前向神经网络构建模型时，基于策略梯度的 TRPO 算法在最终性能上超越了另外两种算法；另外，使用循环神经网络构建模型时，性能超过使用前向神经网络的情况。

- Learning communication（通信学习）

Learning communication在Learning communication中，显式假设智能体之间存在信息的交互，并在训练过程中学习如何根据自身的局部观察来生成信息，或者来确定是否需要通信、与哪些智能体通信等等。在训练完毕后运行的过程中，需要显式依据其余智能体传递的信息来进行决策。这一部分主要论述了以下五部作品。

在[6]中，首次将通信学习引入深度多智能体强化学习，解决的问题是Dec-POMDP问题，使用CTDE框架（Centralized Training Decentralized Execution）。本文假设通信信道是离散的，即只有离散的信息(one-hot向量)可以在代理之间传播。作者提出了两种算法，后者是前者的改进版本，即

- Reinforced Inter-Agent Learning (RIAL)
- Differentiable Inter-Agent Learning (DIAL)

在[7]中，提出了一种新的算法CommNet，假设agent之间传递的消息是连续变量(不像RIAL或DIAL是离散的)，本文使用的强化学习算法是policy gradient 。下图是算法框架，我们可以看出该算法接收所有agent的局部观测值作为输入，然后输出所有agent的决策。

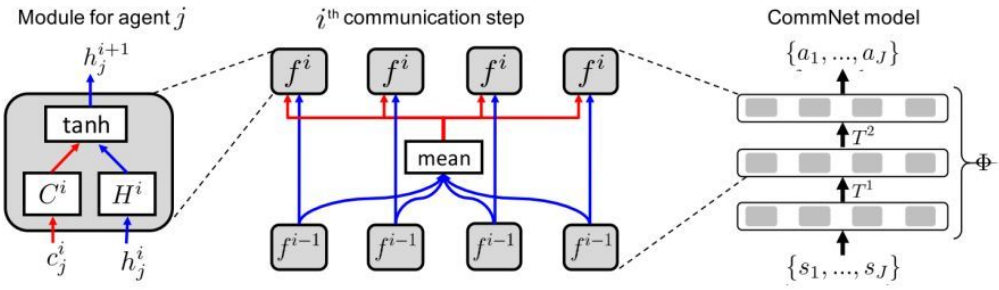


Figure 1: An overview of our CommNet model. Left: view of module f^i for a single agent j. Note that the parameters are shared across all agents. Middle: a single communication step, where each agents modules propagate their internal state h, as well as broadcasting a communication vector c on a common channel (shown in red). Right: full model Φ, showing input states s for each agent, two communication steps and the output actions for each agent.

在[8]中，作者提出了一种新的算法BiCNet，该算法也假设agent之间传输的信息是离散的，旨在解决zero-sum Stochastic Game (SG) 问题，同样遵循CTDE框架。算法是基于actor-critic算法框架,使用DDPG算法,并且考虑到算法在大规模多智能体环境下的可扩展性问题，智能体之间共享模型参数，并且算法假设每个智能体都拥有同样的全局观察（全局状态），这也是本文的局限之一。

在[9]中，笔者认为前三种算法R(D)IAL、CommNet和BiCNet在每一个时间步中都必须在所有agent之间进行通信，或者每个agent与它的相邻agent进行通信。在本文中，将它们视为一种预定义的通信模式，但这种模式不够灵活。作者提出了一种基于注意机制的通信模型ATOC，允许agent随时决定是否需要与其他agent通信以及与哪些agent通信。

在[10]中，作者认为在现实中，通信信道的带宽是有限的。如果所有的座席都向这个带宽有限的通道发送信息，一旦超过容量，信息就会丢失或阻塞。ATOC限制每个启动器最多只能选择加入通信组的agent，但选择的方法非常简单。本文将通信领域中的MAC(Medium Access Control)方法引入到多智能体强化学习中来解决上述问题。本文提出的SchedNet算法框架分为Actor网络、Scheduler网络和Critic网络三部分。对于演员网络部分，每个代理都有自己独立的演员网络。

算法框架如下：

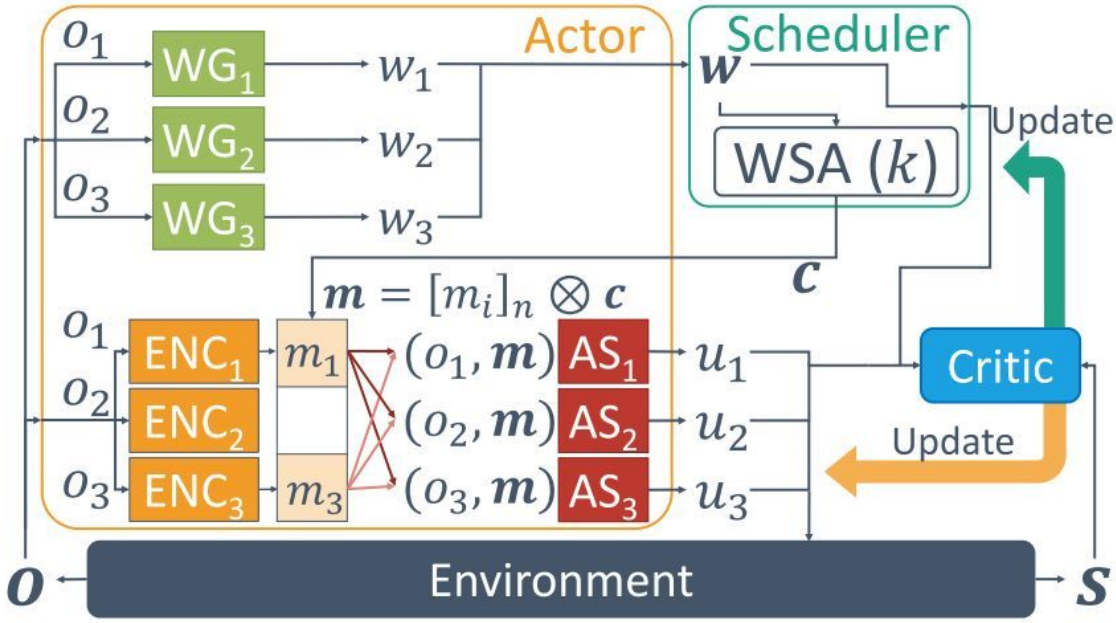


Figure 1: Architecture of SchedNet with three agents. Agents 1 and 3 have been scheduled for this time step.

- Learning cooperation（协作学习）

在Learning cooperation中，并不显式地学习智能体之间的通信，而是将 multi-agent learning 领域的一些思想引入到 MARL 中。而这类方案又可以分为以下三个类别：基于值函数的方法、基于演员-评论家的方法的和基于经验回放缓存的方法。

基于值函数的方法：

基于值函数的方法可以说是多智能体强化学习算法最开始的尝试（例如 Independent Q-Learning， IQL 算法）。本文提出的 VDN 方法的基本思想是，中心化地训练一个联合的 Q network，但是这个联合的网络是由所有智能体局部的 Q networks 加和得到，这样不仅可以通过中心化训练处理由于环境非平稳带来的问题，而且由于实际是在学习每个智能体的局部模型，因而解耦智能体之间复杂的相互关系。最后，由于训练完毕后每个智能体拥有只基于自己局部观察的 Q network，可以实现去中心化执行，即 VDN 遵循 CTDE 框架，并且解决的是 Dec-POMDP 问题。

基于演员-评论家的方法：

将单智能体强化学习算法扩展到多智能体环境中，最简单就是 IQL 类别方法，但是此类方法在复杂环境中无法处理由于环境非平稳带来的问题；另一方面，虽然中心化方法能够处理上述问题，但是与 IQL 相比又失去了较好的可扩展性。

本文提出的 MADDPG 算法将 DDPG 算法扩展到多智能体环境中，MADDPG 算法假定每一个智能体拥有自己独立的 critic network 以及 actor network，并且假定每个智能体拥有自己独立的回报函数，这样 MADDPG 算法就可以同时解决协作环境、竞争环境以及混合环境下的多智能体问题。

但是 MADDPG 算法假定每个智能体在训练时都能够获取其余所有智能体的局部观察以及动作，因而即使每个智能体的 critic network 是独立的，训练时也需要中心化训练，因而遵循 CTDE 框架。

MADDPG 算法框架如下图所示（图片来源原论文）：

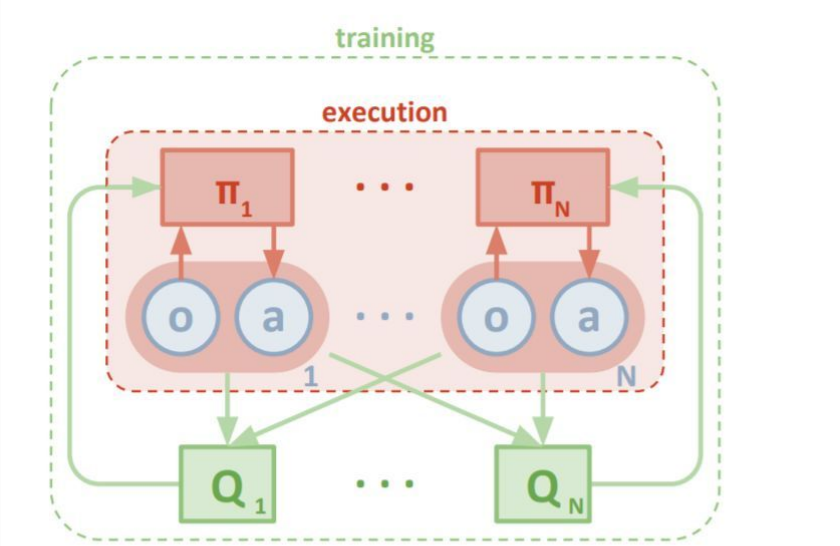
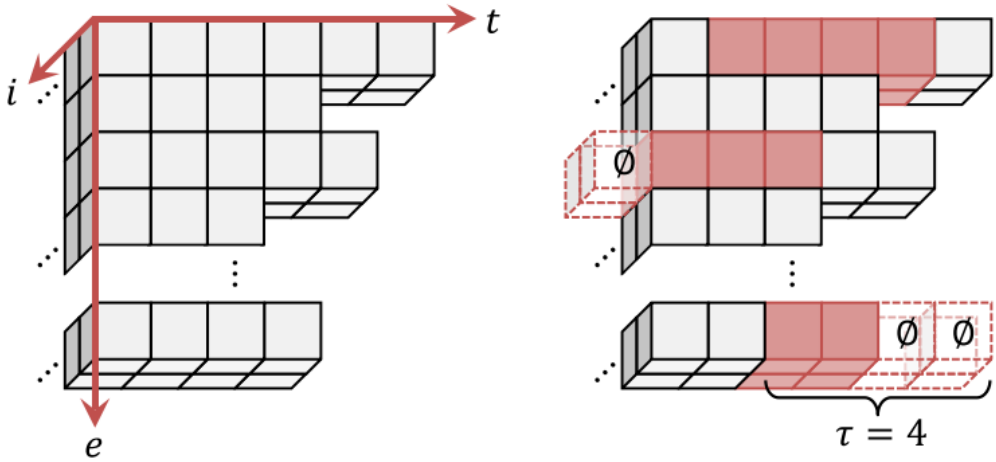


Figure 1: Overview of our multi-agent decentralized actor, centralized critic approach.

基于经验回放缓存的方法：

这部分要介绍的工作主要聚焦于使用 ER 训练 Q-function 时增加稳定性（CommNet 甚至因为 ER 在 multi-agent 环境下的不稳定性而禁用了 ER），这两方法遵循 CTDE 框架，并且类似 MADDPG 方法一样，均假设每个智能体拥有自己独立的 Q-function。Omidshafiei, Shayegan, et al. 的工作致力解决 partial observation 的问题，因而采用的是 DRQN 算法，本文提出采用 ER 训练 DRQN 时应当采用如下方式，并提出了 concurrent experience replay trajectories 的概念（图片来源原论文）：



(a) CERT structure. (b) CERT minibatches.

Figure 1. Concurrent training samples for MARL. Each cube signifies an experience tuple $\langle o_t^{(i)}, a_t^{(i)}, r_t, o_{t+1}^{(i)} \rangle$. Axes e, t, i correspond to episode, timestep, and agent indices, respectively.

即每个智能体在独立训练自己的 Q-function 时，从 ER 中 sample 出来的数据需要从 episode 层面以及时间层面上对齐。

- Agents modeling agents（智能体建模）

这一类方法主要聚焦于通过对其他智能体的策略、目标、类别等等建模来进行更好的协作或者更快地打败竞争对手。这里简单的讨论几个估计其他智能体策略的相关工作。

在[12]中，我们之前已经讨论过这篇文章，但是这篇文章也提出了一种方法来估计其他agent的策略。本文首先从hyper Q-Learning算法出发，该算法使用贝叶斯估计来估计其他agent的策略。但这种方法会增加Q函数的输入维数，使Q函数更难学习。

在[13]中，本文将行为和脑科学领域的心理理论(Theory of Mind, ToM)引入到多智能体强化学习中，以其他智能体的历史行为数据为基础预测未来行为。ToM理论认为，要预测一个agent未来的行为，我们需要知道它的特性、心态和当前状态。它使用特定的方法对个体的个性表示进行编码。

3. 目前遇到了哪些挑战

1)非平稳性:与单个智能体相比，控制多个智能体带来了几个额外的挑战，例如智能体的异构性、如何定义合适的集体目标或需要设计紧凑表示的大量智能体的可扩展性，以及更重要的非平稳性问题。在单智能体环境中，智能体只关心自己行为的结果。在多智能体域中，智能体不仅观察自己行为的结果，还观察其他智能体的行为。智能体之间的学习是复杂的，因为所有智能体都可能相互交互并同时学习。多个主体之间的相互作用不断重塑环境，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一个智能体的策略发生变化，并会影响其他智能体的最优策略。一个行动的潜在回报的估计是不准确的，因此，在多主体环境中给定点的好策略在未来不可能保持这样。在单智能体环境中应用的Q学习收敛理论不能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重复性进行，同时确保不影响试剂的稳定性。在多主体环境下，exploration–exploitation 困境可能会更加复杂。

流行的独立问答学习[48]或基于经验回放的DQN [6]不是为非平稳环境设计的。卡斯塔涅达[49]提出了DQN的两个变种，即深度重复更新Q网络(DRUQN)和深度松散耦合Q网络(DLCQN)，以处理多智能体系统中的非平稳性问题。DRUQN是基于在[50]和[51]中介绍的重复更新Q学习(RUQL)模型开发的。它旨在通过更新与选择行动的可能性成反比的行动值来避免政策偏差。另一方面，DLCQN依赖于[52]中提出的松散耦合的Q学习，它使用每个代理的负面奖励和观察来指定和调整每个代理的独立性程度。通过这种独立性程度，代理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，迪亚洛等人[53]将DQN扩展到多智能体并发DQN，并证明了这种方法可以在非平稳环境中收敛。福尔斯特等人[54]交替介绍了两种方法来稳定DQN在马德尔的经验重播。第一种方法使用重要性抽样方法来自然衰减过时的数据，而第二种方法使用指纹来消除从重放存储器中检索的样本的年龄的歧义。

最近，为了处理由于多智能体系统中多个智能体的并发学习引起的非平稳性，帕尔默等人[55]提出了一种方法，即宽松DQN (LDQN)，该方法利用衰减温度值的宽松度来调整从经验重放存储器中采样的策略更新。多主体环境中的宽容描述了这样一种情况，即学习智能体忽略了合作学习者的不良行为，这导致了低回报，但仍然与合作学习者合作，希望合作学习者能够在未来改进自己的行为。比如智能体A和智能体B在学踢足球。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，在宽大处理的情况下，智能体A会认为智能体B可以提高自己的技能，从而智能体A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，不会再将球传给智能体B[56]。LDQN应用于协调多智能体对象运输问题，并将其性能与hystereticDQN (HDQN) [57]进行了比较。实验结果表明，在随机报酬环境下，LDQN相对HDQN在收敛到最优策略方面具有优势。在[58]中，宽大处理的概念以及

预定的重放策略也被结合到加权的决策支持网络中，以处理多智能体系统中的非平稳性。实验表明，在两个具有随机回报和大状态空间的多智能体环境中，WDDQN比DDQN具有更好的性能。

2)部分可观测性:在现实世界的应用中，有很多情况下，agents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、规模更大。换句话说，当智能体与环境交互时，智能体 不知道与环境有关的状态的完整信息。在这种情况下，智能体会观察关于环境的部分信息，并需要在每个时间步骤中做出“最佳”决策。这种类型的问题可以用部分可观测的MDP模型来模拟。

在当前的文献中，已经提出了许多深度RL模型来处理POMDP。Hausknecht和Stone [44]提出了基于长短期记忆网络的DRQN。通过递归结构，基于DRQN的智能体能够在部分可观察的环境中以健壮的方式学习改进的策略。与DQN不同，DRQN通过递归神经网络逼近Q(o，a)，Q(o，a)是一个带有观测值o和动作a的Q函数。DRQN将网络ht1的隐藏状态视为内部状态。因此，DRQN的特征是Q函数(ot，ht1，a；θi),其中θI是在第ith训练步骤中的网络参数。在[59]中，DRQN被扩展到深度分布式递归Q网络(DDRQN)来处理多代理POMDP问题。DDRQN的成功依赖于三个显著的特征，即最后动作输入、智能体间重量共享和禁用体验回放。第一个特征，即最后动作输入，要求提供每个智能体的前一个动作作为其下一步的输入。智能体间权重共享意味着所有智能体仅使用一个网络的权重，这是在训练过程中学习的。禁用体验重放只是排除了DQN的体验重放功能。因此，DDRQN学习形式为Q(om t，hm t 1，m，am t 1，am t；θi)，其中每个智能体接收其自己的索引m作为输入。权重分担减少了学习时间，因为它减少了要学习的参数数量。虽然每个智能体都有不同的观察和隐藏状态，但是这种方法假设智能体具有相同的一组动作。为了解决复杂的问题，自治智能体通常有不同的操作集。例如，无人机在空中机动，而机器人在地面操作。因此，无人机和机器人的动作空间是不同的，因此不能应用智能体间的重量分担功能。

扩展到部分可观察域中的许多代理的系统是一个具有挑战性的问题。古普塔等人[60]将课程学习技术扩展到MAS，该技术集成了三类深度学习方法，包括策略梯度、TD错误和AC方法。课程原则是先开始学习完成简单的任务，积累知识，再着手执行复杂的任务。这适用于多智能体环境，在该环境中，在扩展以容纳更多智能体来完成越来越困难的任务之前，最初协作的智能体较少。实验结果表明，该课程学习方法在将深度学习算法扩展到复杂多智能体问题方面具有活力。

洪等[61]引入深度策略推理Q-网络(DPIQN)对MASs进行建模，并引入其增强型深度递归策略推理Q-网络(DRPIQN)处理部分可观测性。DPIQN和DRPIQN都是通过在训练过程的不同阶段使网络的注意力适应策略特征和它们自己的Q值来学习的。实验表明，DPIQN和DRPIQN的整体性能优于基线DQN和DRQN [44]。同样在部分可观测性的背景下，但扩展到多任务、多主体问题，Omidshafiei等人[57]提出了一种称为多任务MALL(MT-MALL)的方法，该方法集成了滞后学习者[62]、DRQNs [44]、提炼[63]和并发经验重放轨迹(CERTs)，它们是[6]中提出的经验重放策略的分散扩展。当智能体协作学习用稀疏的奖励完成一组分散的POMDP任务时，智能体没有被明确地提供任务标识(因此是部分可观察的)。然而，这种方法有一个缺点，即不能在具有异构智能体的环境中执行。

除了部分可观测性之外，在某些情况下，智能体必须处理与环境真实状态弱相关的极其嘈杂的观察。基林茨和蒙大纳[64]介绍了一种称为MADDPG-M的方法，该方法结合了深度确定性政策梯度(DDPG)和通信介质来解决这些情况。代理需要决定他们的观察是否有助于与其他智能体共享，并且通信策略是通过经验与主要策略同时学习的。最近，Foerster等人[65]提出了一种贝叶斯动作解码器(BAD)算法，用于学习具有合作部分可观察设置的多个智能体。一个新的概念，即公众信念MDP，被引入到基于贝叶斯估计的贝叶斯估计中，它使用一个近似的贝叶斯更新来获得环境中具有可公开观察特征的公众信念。BAD依赖于分解的近似信念状态来发现约定，以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密切相关。在原理证明两步矩阵博弈和合作部分信息卡牌博弈Hanabi上的实验结果证明了该方法相对于传统策略梯度算法的有效性和优越性。

4. MARL的应用和前景

近几年，MARL在许多领域都得到了实际应用，目 前 在机器人系统、人机博弈和自动驾驶等领域开展的研究较多。在机器人系统领 域，Gu等[53]提出了一种基于离线策略的深度强化学习算法，其可以有效地训练真实的物理机器人。机器人可以在没有任何演示或手动设计的情况下，学习各种仿真与复杂的操作技巧。Foerster等[10]将多智能体强化学习方法应用到机器人交流领域，他们采用集中学习分散执行的方式，第一次实现了机器人之间的深层次交流。Ｄｕａｎ等［５４］在机器人控制上的实践工作对多智能体强化学习的应用也具有启发意义。人机博弈一直都是人工智能领域最具挑战性也最令人兴奋的工作，２０１８年ＯｐｅｎＡＩ和ＤｅｅｐＭｉｎｄ相 继 取 得重大进展。ＯｐｅｎＡＩ在实时５ｖ５策 略 游 戏ｄｏｔａ２中 战 胜 人 类 顶级玩家；ＤｅｅｐＭｉｎｄ在复杂的第一人称多人游戏 ＱｕａｋｅＩＩＩ中达到人类水平，还能与人类玩家合作［１１－１３］。在自动驾驶领域，Ｓｈａｌｅｖ－Ｓｈｗａｒｔｚ等［１４］针对自动驾驶的安全性和环境的不可预测性问题进行了改进和优化，并展示了如何在没有ＭＤＰ 假设的情况下使用策略梯度迭代，以及用随机梯度上升来最小化梯度估计的方差。

可以预见的是，未来ＭＡＲＬ还将被更广泛地应用到各行各业，如互联网、资源管理、交通系统、医疗和金融等领域。目前，在互联网领域，Ｊｉ等［１５］将聚类的思想与ＭＡＲＬ方 法 相 结合，面对大量的广告商，对即时在线投标的性能进行了优化；为了平衡广告商之间的竞争与合作，提出并实现了一种实用的分布式协调多智能体竞标算法。在资源管理方面，Ｘｉ等［１６］提出了一种ＭＡＲＬ新算 法，新算法不基于马尔可夫假设，具有更快的收敛速度和更强的鲁棒性，使得电网系统能够在更复杂的条件下提高新能源的利用率。Ｐｅｒｏｌａｔ等［１７］运 用 部 分马尔可夫观测模型对公共资源的占用主体进行了建模，揭示了排他性、可持续性和不平等性之间的关系，并提出了解决方案，提高了资源管理能力。Ｋｏｆｉｎａｓ等［５５］提 出 的 模 糊 Ｑ 学 习方法有效地提高了分散式微电网的能量管理能力。在交通控制领 域，Ｃｈｅｎ等［５６］提出了一种基于ＭＡＲＬ 的 协 同 控制 框架，并用其来实时缓解公交车道上的公交 拥 挤。Ｖｉｄｈａｔｅ等［５７］提出了一种基于协同多智能体强化学习的交通流模型用于控制优化交通系统，模型能够很好地处理未知的复杂状态。目前，多智能体强化学习应用在医疗和金融等领域的研究还较少，需要人们进行更进一步的探索。

ＭＡＲＬ虽然已经在诸多领域中实现了应用，但 依 然 存 在很多问题。在ＭＡＲＬ未来的研究工作和实际应用中，有一些方向需要进行进一步的关注和探索。首 先，可 扩 展 性 依 然 是最核心的问题。目前，ＭＡＲＬ在电脑游戏、人 机 博 弈（包 括 小型机器人体系）中已经取得了不错的 成 果，其 中 DeepMind公司和OpenAI公司做出了重要的贡献。但是，ＭＡＲＬ在自动驾驶、船舶制造、商业交 易、资源配置等大型机器人体系中的应用还不够成熟，因此将ＭＡＲＬ应用到海量智能体以及更复杂的环境中是未来研究的重要方向。其 次，现 在 普 遍 的 多 智能体强化学习都是基于马尔可夫随机过程环境的，它给许多模型提供了一个简单明确的数学框架，但是现实环境中很多场景和问题是非马尔可夫的，在大部分多智能体场景中，其他智能体的行为是不可预测的。上 文 提 到 的 Shalev等［１４］和 Ｘｉ等［１６］提出的ＭＡＲＬ新算法都不基于马尔可夫假设。因 此，在没有马尔可夫假设的情况下建模实现 MARL，需 要 进 行 进一步的研究与探索。最 后，无监督的生成模 型、计算图模型、注意力机制等其他机器学习方法与ＭＡＲＬ的结合可以解决许多 实 际 应 用 的 问 题。 迁移学习也将被更多地应用到ＭＡＲＬ方法中，以缓解真实任务场景中训练数据缺乏的问题。借助云服务器端的分布式的多智能体协同学习也是未来一个重要的方向。