基于通信的协作型多智能体强化学习算法综述

田 琪,吴 飞
浙江大学计算机科学与技术学院, 杭州 310000

摘 要 多智能体系统在许多实际领域中得到了广泛应用,包括机器人技术、分布式控制和多人游戏等。这些领域中的许多复杂任务无法通过预定义的智能体行为来解决,而基于通信的多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)技术是应对这些挑战的有效方法之一。该领域存在2个核心问题:1)如何建立有效的多智能体通信机制,从而提升多智能体系统的整体性能;2)在带宽受限的场景下,如何设计高效的通信调度方案从而压缩通信过程中冗余信息。本文首先对处理这两个核心问题的文献进行了概述并重点介绍具有代表性的一些工作,接着说明其在航天领域的应用前景,最后进行总结。

关键词 强化学习;通信机制;多智能体系统
中图分类号:TP3 **文献标识码**: A
文章编号:1006-3242(2023)04-0013-07

A Survey of Communication Based Cooperative Multi-Agent Reinforcement Learning Algorithms

Tian Qi, Wu Fei
College of Computer Science and Technology, Zhejiang University, Hangzhou 310000, China

Abstract *Multi-agent systems are widely used in many practical fields, including robotics, distributed control, and multiplayer games. Many complex tasks in these fields can not be solved by predefined agent behaviors, and communication based multi-agent reinforcement learning (MARL) technology is one of the effective methods to deal with these challenges. There are two core research issues in this field: 1) How to establish an effective multi-agent communication mechanism to improve the overall performance of the multi-agent system; 2) In the scenario under limited bandwidth, how to design an efficient communication schedule to compress redundant information in the communication process. The literature is summarized for dealing with these two core issues and some representative works are focused, then its application prospects in the aerospace field is presented, and finally the points of this research are shown.*

Key words *Reinforcement learning; Communication mechanism; Multi-agent system*

0 引言

令智能体拥有类似人类的行为决策能力一直是人工智能研究人员追求的终极目标之一,近年来深度强化学习技术的快速发展使得这个目标成为可能,例如2017年5月,基于深度强化学习技术训练的AlphaGo[1]智能体在中国乌镇围棋峰会上击败

收稿日期:2023-03-29
作者简介:田 琪(1995-),男,博士研究生,主要研究方向为多智能体强化学习、计算机视觉等人工智能技术;
吴 飞(1970-),男,博士,教授,主要从事多媒体、推荐系统等人工智能技术,本文通信作者,E-mail:wufei@zju.edu.cn。

了排名世界第一的世界围棋冠军柯洁,这预示着单智能体在特定决策任务上已经拥有超越人类的能力。自从 AlphaGo 出现后,激发了深度强化学习社区的研究热潮,其中一个重要的研究方向就是协作型多智能体强化学习[2](Cooperative Multi-Agent Reinforc-ement Learning, CMARL)技术。不同于 AlphaGo 这种单智能体决策模型,协作型多智能体强化学习旨在为多个智能体训练其对应的策略模型,从而使得这些智能体能够合作以完成一个共同的目标任务。

传统的协作型多智能体强化学习在训练阶段允许访问环境的全局信息和每个智能体的局部信息,但在执行阶段只允许每个智能体根据自身的局部观测执行下一步的动作,如图 1(a)所示。这显然不是最优的方式,因为在多智能体环境中每个智能体的决策不仅仅与自身观测有关,还与其他智能体有关。为了缓解这个问题,如图 1(b)所示,在传统的协作型多智能体强化学习的基础上,近期许多研究者指出如果允许多个智能体在训练和执行期间相互交换信息,那么每个智能体就能更好地执行下一步的动作,这种学习范式被称为基于通信的多智能体强化学习算法。

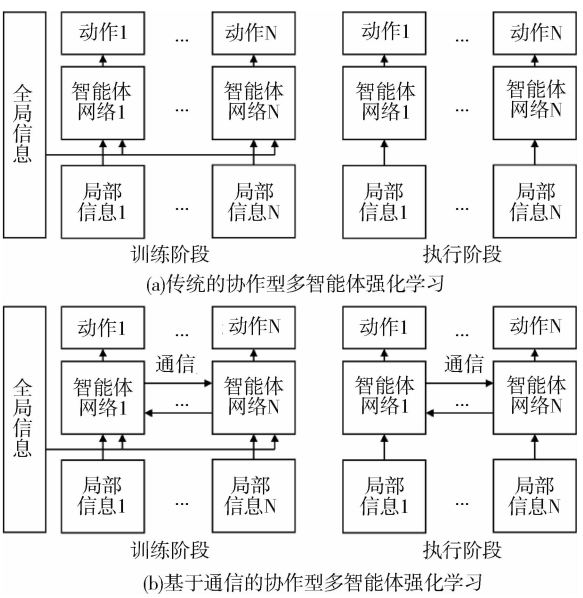

图 1 传统/基于通信的多智能体强化学习训练/执行阶段

本文将针对这种通信类的算法进行综述。即首先介绍基于通信的协作型多智能体强化学习技术基础,然后列举出这个领域中具有代表性的工作,包括传统的通信算法、受限带宽下的通信算法,接着说明基于通信的协作型多智能体强化学习技术在航天领域的应用,最后对本文的内容进行总结。

1 通信机制

1.1 问题定义

基于通信的协作型多智能体强化学习由去中心化的部分可观察马尔可夫决策过程[3](Decentralized Partially Observable Markov Decision Process, Dec-POMDP)扩展而来,它可以被定义为一个元组 $\langle N,S,A,T,R,O,M,\Omega,\gamma\rangle$,其中 N 表示智能体的数量、S 表示环境的全局状态空间、$A=\{\boldsymbol{a}_i\}_{i=1,2,\cdots,N}$ 表示动作集合、$T(\boldsymbol{s}'\mid\boldsymbol{s},\boldsymbol{a}):S\times A\to S$ 表示状态转移函数、$\boldsymbol{a}=[a_1,a_2,\cdots,a_N]$ 表示联合动作空间、$R=\{r_i\}_{i=1,2,\cdots,N}:S\times A\to\mathbb{R}^N$ 表示一组奖励函数,在某些设置下可以是 1 个共享奖励、$O=\{\boldsymbol{o}_i\}_{i=1,2,\cdots,N}$ 表示所有智能体的局部观测集合、$\Omega(s,i):\mathcal{S}\to O_i$ 是决定智能体 i 局部观测的观测函数、γ 表示折扣因子、$M=\{\boldsymbol{m}_i\}_{i=1,2,\cdots,N}$ 表示消息空间,其中 m_i 表示智能体 i 发送的消息,它通常通过神经网络编码局部观测 o_i 获得。每个智能体都会收到由其他智能体发送消息 $\boldsymbol{m}_{-i}=[\boldsymbol{m}_1,\cdots,\boldsymbol{m}_{i-1},\boldsymbol{m}_{i+1},\cdots,\boldsymbol{m}_N]$ 以做出更好的决策,最终的目标是最大化奖励函数。

1.2 训练框架

目前基于通信的协作型多智能体强化学习主要使用 Q-学习[4]风格的训练框架和演员家-评论家[5]的训练框架。Q-学习风格的训练框架将智能体网络建模为 Q 函数 $Q_i(\boldsymbol{o}_i,\boldsymbol{a}_i,\boldsymbol{m}_{-i})$,典型的训练方法是 QMIX[6],该方法的损失函数如下:

$$\begin{cases}\mathcal{L}_{\text{QMIX}}=(y_{\text{tot}}-Q_{\text{tot}}(s,\cdots,Q_i(\boldsymbol{o}_i,\boldsymbol{a}_i,\boldsymbol{m}_{-i}),\cdots))^2\\ y_{\text{tot}}=r+\gamma\max\limits_{\boldsymbol{a}'}Q^-_{\text{tot}}(\boldsymbol{s}',\cdots,Q^-_i(\boldsymbol{o}'_i,\boldsymbol{a}'_i,\boldsymbol{m}'_{-i}),\cdots)\end{cases}\tag{1}$$

其中:Q_{tot} 表示混合网络,$(\cdot)^-$ 表示目标网络,$(\cdot)'$表示下一个时刻的变量。这种智能体的建模方式主要处理离散动作空间的问题。

演员家-评论家的训练框架将智能体网络建模为策略 $\pi_i(\boldsymbol{a}_i\mid\boldsymbol{o}_i,\boldsymbol{m}_{-i})$,典型的训练方法是 MAPPO[7],该方法中具有参数 θ_i 的每个策略 π_i 的更新策略梯度如下:

$$\begin{cases}\nabla_{\theta_i}\mathcal{L}_{\text{MAPPO},i}=\mathbb{E}_{\pi_{i,\text{old}}}[(\min(\zeta_1,\zeta_2)\cdot\kappa)]\\ \kappa=\nabla_{\theta_i}\log\pi_i(\boldsymbol{a}_i\mid\boldsymbol{o}_i,\boldsymbol{m}_{-i})\\ \zeta_1=\rho_i(\sum\limits_{k=t}^{T}\gamma^{k-t}r^t-V(\boldsymbol{s}))\\ \zeta_2=\text{clip}(\rho_i,1-\epsilon,1+\epsilon)(\sum\limits_{k=t}^{T}\gamma^{k-t}r^t-V(\boldsymbol{s}))\end{cases}\tag{2}$$

其中:$\mathbb{E}_{\pi_{i,\text{old}}}(\cdot)$ 表示从 $\pi_{i,\text{old}}$ 产生的轨迹中采样的训练数据、$\rho_i = \pi_i(\boldsymbol{a}_i \mid \boldsymbol{o}_i, \boldsymbol{m}_{-i}) / \pi_{i,\text{old}}(\boldsymbol{a}_i \mid \boldsymbol{o}_i, \boldsymbol{m}_{-i})$ 表示校正偏差的重要性权重、$\text{clip}(\cdot)$ 表示截断函数、ϵ 表示控制策略更新可行范围的超参数、$V(s)$ 表示中心化价值函数,其通过最小化均方差损失 $\mathcal{L}_{\text{MSE}} = \left(\sum_{k=t}^{T} \gamma^{k-t} r^t - V(\boldsymbol{s})\right)^2$ 进行优化。

2　深度强化学习算法

基于通信的协作型多智能体强化学习的重点是如何处理智能体之间传递的消息。具体来说,对于智能体 i,其接收的消息可以被表示为 $\boldsymbol{m}_{-i} = [\boldsymbol{m}_1, \cdots, \boldsymbol{m}_{i-1}, \boldsymbol{m}_{i+1}, \cdots, \boldsymbol{m}_N]$,这些来自其他智能体的消息 $\boldsymbol{m}_j(\forall j \neq i)$ 应该通过怎样的交流模块进行聚合是基于通信的协作型多智能体强化学习算法关注的重点,当 $\boldsymbol{m}_{-i}$ 聚合完成后,代入到式(1)或式(2)的训练框架中即可完成多智能体系统的训练。目前已经涌现了许多基于通信的协作型多智能体强化学习算法的文献,本文将这些文献分为传统通信方法和受限带宽通信方法 2 类,前者的目的是希望多智能体系统在加入通信模块后可以最大化提升系统的整体性能,后者是希望通信模块在增益系统性能的同时尽量占用更少的通信带宽,从而压缩冗余的通信消息。下面本文将依次介绍这 2 类算法。

2.1　传统通信方法

传统通信方法旨在通过交流模块帮助多智能体能够更好地完成一个合作任务,如图 2 所示,其可以分为全连接通信、局部连接通信和加权连接通信 3 种类别。全连接通信是指每个智能体会接收来自其他智能体传输的所有消息,如图 2(a)。局部连接通信是指每个智能体只会接收部分智能体传来的信息,因为并非所有消息都对自身决策有用,过多的冗余消息反而会成为噪声,对决策产生负面影响,如图 2(b)。加权连接通信是每个智能体按重要性权重采纳其他智能体传来的消息,而不是完全接受或者完全否定,是一种更合理的方式,如图 2(c)。下面介绍这 3 类传统通信方法的代表方法。

对于全连接通信,CommNet[8] 是其典型代表,也是该领域最早的工作之一,后续的许多工作都是基于该方案的改进。

所图 3 所示,CommNet 假设通信消息通过聚合模型循环执行 k 轮,对于智能体 i,其在第 k 轮的输出 h_i^k 可以表示为:

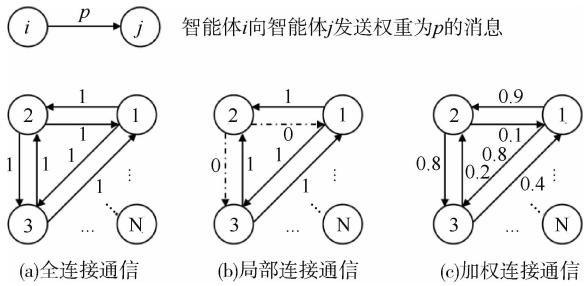

图 2　传统通信方法的分类

$$\begin{cases} \boldsymbol{m}_{-i}^{k-1} = \dfrac{1}{J-1} \sum_{j \neq i} \boldsymbol{h}_j^{k-1} \\ \boldsymbol{h}_i^k = f^k(\boldsymbol{h}_i^{k-1}, \boldsymbol{m}_{-i}^{k-1}) \end{cases} \tag{3}$$

其中:$\boldsymbol{h}_j^{k-1}$ 表示($k-1$)轮通信中除智能体 i 外其他所有智能体 j 经过上一轮函数 f^{k-1} 的输出、J 表示其他所有智能体的总数量、$\boldsymbol{m}_{-i}^{k-1}$ 表示第($k-1$)轮通信聚合的消息、第 k 轮的 f^k 函数中包含 H^k,M^k 两个多层感知器以及一个 tanh 激活函数。通信的初始化函数为 $\boldsymbol{h}_i^0 = r(\boldsymbol{o}_i)$,其中 r 可以为多层感知器或者循环神经网络,o_i 为每个智能体的局部观测。可以发现 CommNet 的聚合方法就是简单的将其他智能体的信息求均值,由于并非所有消息都是有用的,因此这种方式训练出的模型性能并不好,除了同期的文献 DIAL[9] 也采用这种通信方法,后期的所有文献都摈弃了这种全连接式通信。

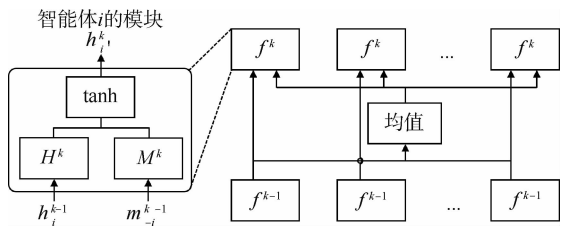

图 3　CommNet 的消息聚合机制

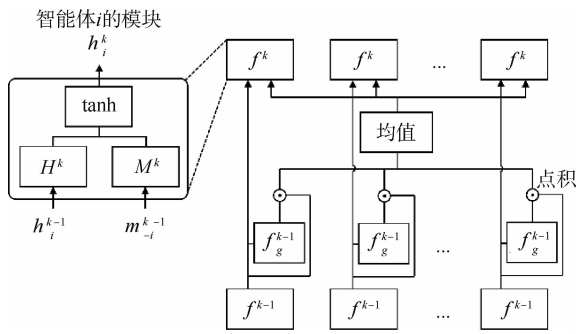

图 4　IC3 Net 的消息聚合机制

对于局部连接通信,IC3 Net[10] 是典型代表。如

图 4 所示，该方法的总体训练框架和 CommNet 非常相似，主要的不同是每个智能体在第 k 轮交流时都会学习一个神经网络 f_g^k，该神经网络将智能体上一层的输出数据作为输入，并从 0 或者 1 中预测一个值，如果结果为 0 表示对应智能体的信息不参与聚合，如果结果为 1 表示对应智能体的信息参与聚合。它就像一个门控开关一样，因此被称为基于门机制的神经网络。

对于智能体 i，其在第 k 轮的输出 h_i^k 可以表示为：

$$\begin{cases} \boldsymbol{g}_j^{k-1} = f_g^k(\boldsymbol{h}_j^{k-1}),\ \forall j \neq i \\ \boldsymbol{m}_{-i}^{k-1} = \dfrac{1}{G\text{-}1}\sum_{j\neq i}\boldsymbol{h}_j^{k-1}\odot\boldsymbol{g}_j^{k-1} \\ \boldsymbol{h}_i^k = f^k(\boldsymbol{h}_i^{k-1},\boldsymbol{m}_{-i}^{k-1}) \end{cases} \tag{4}$$

式中：大多数符号的含义与式(3)中的符号一致，主要的不同是门机制神经网络 f_g^k 会输出门控元素 $\boldsymbol{g}_j^{k-1}\in\{0,1\}$ 来控制消息是否参与聚合，f_g^k 本身作为一个策略网络与智能体策略网络叠加在一起进行训练，G 表示通信通道打开的数量。这种建模门控网络的思想被后续的许多工作借鉴，如 VBC[11]，TMC[12]，I2C[13]和 SMS[14]，但是它们具体的建模方式略有不同，本文将在表 1 中进行简单总结。

对于加权连接通信，TarMAC[15]是其典型的代表，它主要是将自然语言处理领域中的注意力机制引入到多智能体交流模块中。

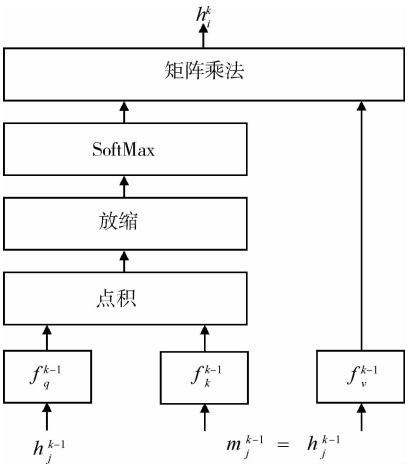

图 5 TarMAC 的消息聚合机制

如图 5 所示，假设第 $(k-1)$ 轮通信后智能体 i 的输出为 $\boldsymbol{h}_i^{k-1}$，其他智能体 $j(j\neq i)$ 的输出为 $\boldsymbol{h}_j^{k-1}(\boldsymbol{m}_j^{k-1})$，根据注意力机制的工作方式，首先通过 $\boldsymbol{q}_i^{k-1}$，$\boldsymbol{k}_j^{k-1}$ 和 $\boldsymbol{v}_j^{k-1}$ 这三个多层感知器学习查询向量 $\boldsymbol{q}_i^{k-1}$、键值向量 $\boldsymbol{k}_j^{k-1}$ 和值向量 $\boldsymbol{v}_j^{k-1}$，具体计算方式如下：

$$\begin{cases} \boldsymbol{q}_i^{k-1} = f_q^{k-1}(\boldsymbol{h}_i^{k-1}) \\ \boldsymbol{k}_j^{k-1} = f_k^{k-1}(\boldsymbol{m}_j^{k-1}) \\ \boldsymbol{v}_j^{k-1} = f_v^{k-1}(\boldsymbol{m}_j^{k-1}) \end{cases} \tag{5}$$

然后将 $\boldsymbol{q}_i^{k-1}$ 和 $\boldsymbol{k}_j^{k-1}$ 进行点积、放缩和 SoftMax 计算以获得注意力(重要性)权重：

$$\boldsymbol{\alpha}_i^{k-1} = \mathrm{SoftMax}[\alpha_{i1}^{k-1},\cdots,\alpha_{iN}^{k-1}] = \mathrm{SoftMax}\left[\frac{\boldsymbol{q}_i^{k-1\mathrm{T}}\boldsymbol{k}_1^{k-1}}{\sqrt{d_{k-1}}},\cdots,\frac{\boldsymbol{q}_i^{k-1\mathrm{T}}\boldsymbol{k}_N^{k-1}}{\sqrt{d_{k-1}}}\right] \tag{6}$$

其中：$(\cdot)^{\mathrm{T}}$ 表示向量转置，d_{k-1} 表示查询向量 $\boldsymbol{q}_i^{k-1}$ 的向量维度，目的是对点积数值在向量维度上进行归一化，SoftMax 表示归一化指数函数，目的是对重要性权重进行归一化。接着利用获得的注意力权重就可以进行如下信息聚合：

$$\boldsymbol{h}_i^k = \sum_{j=1}^{N}\alpha_{ij}^{k-1}\boldsymbol{v}_j^{k-1} \tag{7}$$

这种基于注意力的方法也启发了后续一系列工作，如 DGN[16]，SymbC[17]，它们具体的建模方式与 TarMAC 略有不同。实际上局部连接通信和加权连接通信并不矛盾，一些工作将它们的思想融合，实现了更优的性能，比如 G2A[18]，MAGIC[19]，MAIC[20]。表 1 简单总结了传统通信方法，其中类型 A/B/C 分别表示全/局部/加权连接通信。

表 1 传统通信方法的总结

名称	日期	类型	特点
CommNet[8]	2016	A	引入均值聚合
DIAL[9]	2016	A	引入可微分消息
IC3 Net[10]	2019	B	引入门机制
TarMAC[15]	2019	C	引入注意力机制
VBC[11]	2019	B	引入动作影响机制
DGN[16]	2020	C	引入图卷积网络
G2A[18]	2020	B/C	引入软硬注意力机制
TMC[12]	2020	B	引入消息存储机制
SymbC[17]	2020	C	引入神经符号
I2C[13]	2020	B	引入因果影响机制
MAGIC[19]	2021	B/C	引入图注意力网络
MAIC[20]	2022	B/C	引入表征学习
SMS[14]	2022	B	引入沙普利值

2.2 受限带宽通信

传统通信方法不限制通信带宽，只要对多智能体系统有利的消息都允许进行传递，然而在现实场景中如果通信占用太多的带宽，将消耗大量的资源，因此以受限带宽下的多智能体交流为主题的研

究方向逐渐受到研究者的重视。这一研究领域的关键是如何对通信消息进行压缩,为了将这部分的工作放在一种统一的视角下讨论,可以将多智能体交流看作交流图上的信息流动,其中图的节点就是每个智能体需要传递的消息,图的边是信息流动的方向,那么交流图的信息压缩可以被分为结构压缩和节点压缩。结构压缩是指每个智能体应该尽可能少地和它们智能体交流,节点压缩是指当确定两个智能体需要交流时,传输的信息也应该是简洁的。

对于结构压缩类的工作,GACML[21] 是典型代表。

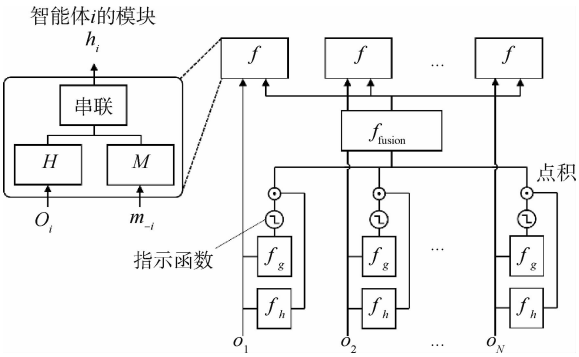

图 6　GACML 的消息聚合机制

如图 6 所示,GACML 与图 4 中的 IC3 Net 非常相似,对于智能体 i,其输出 $\boldsymbol{h}_i$ 可以表示为:

$$\begin{cases} \boldsymbol{v}_j = f_h(\boldsymbol{o}_j),\ \forall j \neq i \\ \boldsymbol{g}_j = f_g(\boldsymbol{o}_j) \\ \boldsymbol{m}_{-i} = \dfrac{1}{G-1}\sum_{j\neq i}\boldsymbol{v}_j \odot \boldsymbol{g}_j \\ \boldsymbol{h}_i = f(\boldsymbol{o}_i, \boldsymbol{m}_{-i}) \end{cases} \tag{8}$$

GACML 与 IC3 Net 主要有如下几点区别:1) 由于带宽受限,GACML 只考虑单轮通信而不像 IC3 Net 那样建模为多轮通信。2) 消除了 IC3 Net 中 f 函数的 tanh 激活单元。3) 消息融合方式从 IC3 Net 中简单求均值变为一个融合网络 f_{fusion}。4) 门控单元函数的学习不再是与智能体策略网络联合训练,而是建模为二分类的监督学习,其标签可以表示为:

$$\begin{cases} Y(\boldsymbol{o}_i) = \mathbb{I}(\Delta Q(\boldsymbol{o}_i) > T) \\ \Delta Q(\boldsymbol{o}_i) = Q(\boldsymbol{o}_i, \boldsymbol{a}_i^C, \boldsymbol{o}_{-i}, \boldsymbol{a}_{-i}^C) - Q(\boldsymbol{o}_i, \boldsymbol{a}_i^I, \boldsymbol{o}_{-i}, \boldsymbol{a}_{-i}^C) \end{cases} \tag{9}$$

其中:$\mathbb{I}$ 表示指示函数,$\boldsymbol{a}_i^C$ 表示智能体 i 在接收消息后做出的动作,$\boldsymbol{a}_i^I$ 表示智能体 i 在不接收消息而独立做出的动作,$\boldsymbol{a}_{-i}^C$ 表示其他智能体在接收消息后做出的动作。这个式子的含义可以理解为如果智能体 i 接收消息后 Q 函数的变化大于阈值 T,说明这个消息是重要的,因此此时的门控单元应该打开,反之关闭,当获得这个标签后,门控网络 f_g 可以用如下二分类损失函数进行训练:

$$\begin{cases} \mathcal{L}_{f_g} = -\mathbb{E}_{o_i}[\boldsymbol{A} + \boldsymbol{B}] \\ \boldsymbol{A} = Y(\boldsymbol{o}_i)\log f_g(\boldsymbol{o}_i) \\ \boldsymbol{B} = (1 - Y(\boldsymbol{o}_i))\log(1 - f_g(\boldsymbol{o}_i)) \end{cases} \tag{10}$$

其中:$\mathbb{E}_{o_i}$ 表示采样数据来自智能体 i 的轨迹。这部分是 GACML 与 IC3 Net 相比最明显的区别,阈值 T 可以主动控制带宽的限制量,从而强制每个智能体只能接受少量其他智能体发送的信息,实现了对门控单元的有效控制,类似的改进工作还有 SchedNet[22] 和 ETCNet[23]。

对于节点压缩类的工作,IMAC[24] 是典型代表。它的主要思想是利用信息瓶颈理论[25] 构建如下有约束的优化问题:

$$\mathcal{L}_{\text{MARL}} \text{ s. t. } \text{MI}(\boldsymbol{o}_i; \boldsymbol{m}_i) \leqslant I_c \tag{11}$$

其中:优化目标是原始的多智能体强化学习损失函数 $\mathcal{L}_{\text{MARL}}$,它可以被实例化为式(1)或式(2)、MI(·;·)表示互信息、I_c 表示约束的信息项、s. t. 表示在优化过程中局部观测 o_i 和消息表征 $\boldsymbol{m}_i$ 的互信息应该限制到信息量 I_c 以下。从总体上看,该式子的含义是希望消息表征 $\boldsymbol{m}_i$ 在满足多智能体任务的前提下,尽可能地压缩消息表征 $\boldsymbol{m}_i$ 的信息量,从而减少消息对带宽的占用。为了求解这个问题,需要用拉格朗日方程将这个带约束的优化问题转换为无约束优化问题,然而即使进行了这样的转换,损失函数中的互信息项依然难以优化,因此研究者通过变分推理获得了互信息项的变分上界,使整个损失可以实现端到端的优化。后续提出的 NDQ[26] 借鉴了上述思想,并设计了 2 个互信息项以同时约束消息表征的简洁性和紧凑性。表 2 简单总结了所有受限带宽的通信方法,其中类型 A/B 分别属于结构压缩/节点压缩类型的文献。

表 2　受限带宽通信方法的总结

名称	日期	类型	特点
SchedNet[22]	2019	A	引入排序式门控机制
GACML[21]	2020	A	引入监督学习式门控机制
IMAC[24]	2020	B	引入信息瓶颈正则项
NDQ[26]	2020	B	引入简洁和紧凑性正则项
ETCNet[23]	2021	A	引入触发机制

3 在航天领域的应用

基于通信的多智能体强化学习在合作型的卫星群控制领域拥有广阔的应用前景,比如美国太空探索技术公司 SpaceX 近期提出了一种名为“星链”的项目,该项目计划 2019 ~ 2024 年间在太空搭建由约 1.2 万颗卫星组成的“星链”网络提供互联网服务,其中 1584 颗将部署在地球上空 550 km 处的近地轨道,并从 2020 年开始工作。在这个场景下,每个“星链”卫星可以看作为一个智能体,为了使卫星群尽可能多地有效覆盖地球表面,需要精密控制卫星群的行动轨迹;另一方面,避免不同卫星间的碰撞也是非常重要的环节,而交流机制可以很好地使每个智能体理解其他智能体下一步可能的动作,从而优化多个卫星的群体运行轨迹。

4 结论

首先介绍了基于通信的协作型多智能体强化学习与传统协作型多智能体深度强化学习的区别,然后详细说明了多智能体强化学习中的通信机制,接着对常见的基于通信的协作型多智能体深度强化学习算法进行了分类和介绍,指出这类算法在航天领域的应用前景,最后对文章进行总结。

参 考 文 献

[1] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484-489.

[2] Canese L, Cardarilli G C, Di Nunzio L, et al. Multi-agent reinforcement learning: A review of challenges and applications[J]. Applied Sciences, 2021, 11(11): 4948-4960.

[3] Amato C, Chowdhary G, Geramifard A, et al. Decentralized control of partially observable Markov decision processes[C]//52nd IEEE Conference on Decision and Control. IEEE, 2013: 2398-2405.

[4] Watkins C J C H, Dayan P. Q-learning[J]. Machine learning, 1992, 8: 279-292.

[5] Konda V, Tsitsiklis J. Actor-critic algorithms[J]. Advances in neural information processing systems, 1999, 12.

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

[7] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[J]. arXiv preprint arXiv:2103.01955, 2021.

[8] Sukhbaatar S, Fergus R. Learning multiagent communication with backpropagation[J]. Advances in neural information processing systems, 2016, 33: 721-730.

[9] Foerster J, Assael I A, De Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning[J]. Advances in neural information processing systems, 2016, 14: 275-283.

[10] Singh A, Jain T, Sukhbaatar S. Learning when to communicate at scale in multiagent cooperative and competitive tasks [C]//International Conference on Learning Representations, 2019.

[11] Zhang S Q, Zhang Q, Lin J. Efficient communication in multi-agent reinforcement learning via variance based control[J]. Advances in Neural Information Processing Systems, 2019, 33: 3233-3241.

[12] Zhang S Q, Zhang Q, Lin J. Succinct and robust multi-agent communication with temporal message control[J]. Advances in Neural Information Processing Systems, 2020, 33: 17271-17282.

[13] Ding Z, Huang T, Lu Z. Learning individually inferred communication for multi-agent cooperation [J]. Advances in Neural Information Processing Systems, 2020, 33: 22069-22079.

[14] Di Xue, Lei Yuan, Zongzhang Zhang, et al. Efficient multi-agent communication via shapley message value [C]// Proceedings of the 31 st International Joint Conference on Artificial Intelligence, 2022: 578-584.

[15] Das A, Gervet T, Romoff J, et al. Tarmac: Targeted multi-agent communication [C]//International Conference on Machine Learning. PMLR, 2019: 1538-1546.

[16] Jiang J, Dun C, Huang T, et al. Graph convolutional reinforcement learning[C]//International Conference on Learning Representations, 2020.

[17] Inala J P, Yang Y, Paulos J, et al. Neurosymbolic transformers for multi-agent communication [J]. Advances in Neural Information Processing Systems, 2020, 33: 13597-13608.

[18] Liu Y, Wang W, Hu Y, et al. Multi-agent game abstraction via graph attention neural network [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(5): 7211-7218.

[19] Niu Y, Paleja R R, Gombolay M C. Multi-agent graph-attention communication and teaming [C]//AAMAS.

2021: 964-973.

[20] Yuan L, Wang J, Zhang F, et al. Multi-agent incentive communication via decentralized teammate modeling [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(9): 9466-9474.

[21] Mao H, Zhang Z, Xiao Z, et al. Learning agent communication under limited bandwidth by message pruning [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(4): 5142-5149.

[22] Kim D, Moon S, Hostallero D, et al. Learning to schedule communication in multi-agent reinforcement learning [C]//International Conference on Learning Representations, 2019.

[23] Hu G, Zhu Y, Zhao D, et al. Event-triggered communication network with limited-bandwidth constraint for multi-agent reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 46: 1478-1486.

[24] Wang R, He X, Yu R, et al. Learning efficient multi-agent communication: An information bottleneck approach [C]//International Conference on Machine Learning. PMLR, 2020: 9908-9918.

[25] Saxe A M, Bansal Y, Dapello J, et al. On the information bottleneck theory of deep learning[J]. Journal of Statistical Mechanics: Theory and Experiment, 2019, 2019(12): 1240-1251.

[26] Wang T, Wang J, Zheng C, et al. Learning nearly decomposable value functions via communication minimization[C]//International Conference on Learning Representations, 2020.