杨桂松, 魏满舟. 通信受限下基于变长自动编码器的多智能体协同算法[J]. 智能计算机与应用,2026,16(2):35-43. DOI:10.20169/j.issn.2095-2163.24040307

通信受限下基于变长自动编码器的多智能体协同算法

杨桂松, 魏满舟
(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘 要: 现实场景的通信条件往往无法满足现有多智能体协同算法的需求。为了提高协同算法在通信资源受限环境下的性能,本文提出了一种创新的多智能体协同算法 AE-CommNet,从通信对象、通信时间以及通信内容三个角度进行优化:通信对象方面,智能体仅仅与其邻居智能体进行通信;通信时间方面,提出了一种基于事件触发的门控通信机制,能够自适应地选择通信时间;通信内容方面,提出了一种创新的变长自动编码器,能够根据智能体所处的状态自适应地生成变长的信息。优化后提高了在复杂多变通信条件环境下的多智能体协同性能,并有效减少了通信资源的消耗。实验结果表明,与 IQL、CommNet 算法相比,本文提出的算法 AE-CommNet 在协同导航和协同巡检等场景下均展现出了更出色的协同性能。此外,AE-CommNet 算法还显著减少了对通信资源的需求量,进一步提高了该算法在复杂环境中的适应性。

关键词: 多智能体系统; 多智能体协同; 多智能体强化学习; 通信资源受限; 自动编码器
中图分类号: TP393 **文献标志码**: A **文章编号**: 2095-2163(2026)02-0035-09

Variable-length autoencoder-based multi-agent cooperation with limited communication

YANG Guisong, WEI Manzhou
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The communication conditions of real-world scenarios often cannot meet the needs of existing multi-agent cooperation algorithms. To improve the performance of the cooperative algorithm in a communication-limited environment, this paper proposes an innovative multi-agent cooperative algorithm, which optimizes the communication object, communication time, and communication content. In terms of communication objects, agents only communicate with their neighbors. In terms of communication time, the paper proposes an event-triggered gated communication mechanism, which could adaptively select the communication time. In terms of communication content, this paper proposes an innovative variable-length autoencoder, which can adaptively generate variable-length messages according to the state of the agent. The optimization of the three levels improves the performance of multi-agent cooperation in complex and variable communication conditions and effectively reduces the consumption of communication resources. Experimental results show that compared with IQL、CommNet algorithms, the proposed algorithm AE-CommNet shows better cooperative performance in scenarios such as cooperative navigation and cooperative inspection. In addition, the proposed algorithm AE-CommNet significantly reduces the demand for communication resources, which further improves the adaptability of the algorithm in complex environments.

Key words: multi-agent system; multi-agent cooperation; multi-agent reinforcement learning; limited communication resources; autoencoder

0 引 言

随着多智能体强化学习技术的不断发展,多智能体系统已被广泛应用于各种协同场景如自动驾驶[1]、多机器人协作[2-3]等。这些协同任务要求多个智能体在复杂多变的动态环境中协同工作,达成共同目标。然而,受限于现实世界的约束,多智能体系统常常面临着计算资源有限、通信带宽受限[4-5]

基金项目: 国家自然科学基金(61602305,61802257); 上海市自然科学基金(18ZR1426000,19ZR1477600)。

作者简介: 杨桂松(1982—),男,博士,副教授,主要研究方向:物联网,普适计算。Email:gs_yang@aliyun.com; 魏满舟(1999—),男,硕士研究生,主要研究方向:多智能体强化学习,多智能体协同算法。

收稿日期: 2024-04-03

等挑战。这些限制不仅影响了智能体之间的协作效率,导致整体协同性能的下降,甚至无法完成协同任务。

为了解决通信受限问题,目前已有大量工作尝试缓解通信受限情况下智能体协作效率降低的问题,现有的工作主要从以下个角度解决:

(1)通信对象。即哪些智能体之间应该建立通信连接。在多智能体系统中,智能体间的通信是协同决策的关键。通过通信,智能体能获取更多环境状态信息,从而增强决策能力。但在通信受限的情境下,选择合适的通信对象变得尤为重要。现有的解决方案主要围绕 2 种策略:广播和邻近智能体通信。广播策略,如 CommNet[6] 智能体,允许将信息广播给所有其他智能体,并通过均值聚合方式融合信息。该方法虽然能提供丰富的信息,但可能导致连续时间和邻近智能体信息的冗余,造成通信资源的浪费。为了减少通信资源的消耗,另一种策略是只与网络邻近智能体进行通信[7-9]。GAXNet[10] 仅允许智能体与可观测范围内的智能体进行通信,DGC[11] 限制智能体只能与距离上最近的 3 个智能体进行通信,LSC[12] 通过学习的方式构建智能体之间的关系,低级智能体向高级智能体传递信息,高级智能体构建全局信息来实现多智能体之间的协作。仅与邻居智能体进行通信能够降低通信消耗,但仍未完全解决邻近智能体信息冗余的问题。因此,本文引入门控机制,根据通信信息的价值过滤信息并选择通信对象,选择性地与邻居智能体进行通信,避免了智能体之间传输冗余信息、浪费通信资源。

(2)通信内容。即如何提取关键信息或采用高效的编码方式来减少通信数据量。确定智能体间的通信关系后,如何选择通信内容成为关键。智能体的局部观测信息和历史信息,在部分可观测的环境下,显得尤为重要。通信内容直接影响智能体间的协同效率。原始的通信内容虽然信息丰富,有助于提高协同效率,但过大的信息量会增加通信时间和成本,尤其在通信受限的场景中,可能导致智能体的协同性能下降。另一方面,经过编码的通信内容虽然减少了信息量,提高了信息密度,但也可能丢失大量有用信息,从而影响智能体的决策和整体协同性能。因此,设计一种能够在降低信息量的同时尽可能保留原始信息的机制至关重要。现有的相关工作在通信内容的设计上主要分为 2 个方向:原始信息传输和编码特征传输。原始信息传输能够提供最全面、最丰富的内容,但可能导致通信负担增加。例如,文献[13-14]在传输信息时使用了原始的观测信息。而编码特征传输则通过编码器对原始信息进行压缩和提炼,以减少通信量。相关工作如文献[15]使用简单的前馈神经网络对局部观察进行编码,文献[16]则采用 MLP 进行编码。此外,还有 CNN[11]、RNN[6]、以及 GNN[17] 等编码器被应用于此目的。本文提出了一种创新的变长自动编码器模块,其通过训练多层的编码器,能够根据原始信息的信息价值最大化地压缩信息,生成变长的特征信息,增强信息的密度并减少通信内容长度。这种方法旨在降低通信成本的同时,尽可能保留原始信息,并提高传输信息的信息密度,从而优化智能体间的协同效率。

(3)通信时间。即选择何时与其他智能体进行通信。在智能体协同作业中,通信的时机选择至关重要。频繁的信息交换能够提升作业效率,但同时也会导致通信资源的过度消耗。相对地,若通信频率过低,则可能导致信息的时效性问题,并限制智能体间的协作能力。选择适合的通信间隔,是提高多智能体系统效率的关键挑战之一。当前的通信时间策略主要分为周期性通信和基于事件触发的通信。周期性通信方法如 CommNet[6],在训练阶段利用连续的信息传递来辅助学习,通过反向传播优化模型。然而,这一策略忽略了信息在时间上的冗余性,导致通讯负担加重并降低了效率,进而制约了其在通信受限环境下的广泛应用。类似地,LCC-UCB[18] 在每轮游戏结束的时候进行通信,但其需要所有智能体的信息,无法应用于大规模通信受限场景。LCC-UCB-GRAPH 适用于稀疏网络,扩展性好,但其算法性能依赖于智能体网络结构。与此相对,基于事件触发的方法则能更加精准地根据信息的实际价值判断是否进行传输。例如,IC3Net[19] 则采用门控机制,指导智能体学习何时通信以优化通信效果。类似地,文献[20]计算测量数据的新颖性,通过只分享新颖的数据来减少通信量。本研究提出的传输控制模块基于门控策略,其不同之处在于使用融合了多维感知数据后的特征作为输入,为智能体提供了丰富的信息以评估通信内容的价值。这种方法不仅有助于减少对通信资源的依赖,同时也能在通信资源受限的环境中促进更有效的信息交流,增强智能体间的协作性能。

现有的多智能体协同算法性能提升工作大多从通信对象、通信内容或通信时间某一方面出发。少部分研究从 2 个或多个角度考虑,但往往未全面整合这 3 个关键要素来解决通信受限下的协同算法性

能下降问题。因此本文提出了一种通信受限下端对端的多智能体协同算法，整合通信对象、通信内容和通信时间三个维度，旨在提高多智能体在通信受限环境下的协同能力和通信效率。具体来说，通信对象方面，本文结合门控机制来决定当前与哪些邻居智能体进行通信；通信内容方面，本文提出了一种创新的变长自动编码器模块，其在尽可能保留原始观测信息的关键特征的基础上，实现信息的有效压缩；通信时间方面，本文利用上下文信息设计了一种通信门控模块，来决定何时进行通信，减少冗余通信。本文的贡献主要如下：

(1)本文设计并实现了一种多智能体协同算法 AE-CommNet，其从通信对象、通信内容和通信时间 3 个维度出发，分别设计了 3 个子模块，针对通信受限场景进行优化，显著提高了多智能体协同的协同效率。

(2)本文提出并实现了一种创新的变长自动编码器，通过设计多层编码器模块，来实现对原始信息的变长编码，在降低通信内容的同时，保留了原始信息的关键特征。

(3)在经典多智能体协同场景(协同导航和协同巡检)下进行实验。实验结果验证了本文算法的协同效率，对比其他算法，该算法在保持协同能力的同时，显著降低了通信消耗。

1　前置知识

1.1　分布式部分可观测马尔可夫决策过程

分布式部分可观测马尔可夫决策过程(Dec-POMDP)是一个用于建模多智能体协同问题的框架，对经典的单智能体马尔科夫决策工程(MDPs)进行了扩展。Dec-POMDP[21]可以被形式化为一个元组 $\langle N,S,\boldsymbol{A},P,r,Z,O,s_0,\rho,\gamma\rangle$，其中 N 是智能体数量，S 是状态空间，$\boldsymbol{A}$ 是联合动作空间，P 是状态转移概率函数，r 是奖励函数，Z 是单个智能体的观测空间，O 是观测函数，$s_0 \sim \rho$ 是环境的初始状态、遵循初始分布 $\rho,\gamma \in [0,1]$ 是折扣因子。基于此，构成了如下交互逻辑：智能体 i 观测到环境状态 $s \in S$ 的一部分观测信息 $z_i = O(s): S \to Z$，其中 O 是观测函数，并做出决策 a_i，形成联合动作 $\boldsymbol{a} \in \boldsymbol{A}$，然后作用于环境，环境将根据状态转移概率 $P(s' \mid s,\boldsymbol{a}): S \times \boldsymbol{A} \times S \to [0,1]$ 转移到状态 s'，并返回奖励 $r = r(s,\boldsymbol{a}): S \times \boldsymbol{A} \to \mathbb{R}$。

1.2　深度 Q 网络

深度 Q 网络(DQN[22])是一种将强化学习与深度神经网络结合的算法，已被证明能在多种游戏环境中达到甚至超越专业游戏玩家的性能。该方法的目标是最大化总期望折扣回报 $R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i$。DQN 通过学习动作值函数来求值：

$$Q^{\pi}(s,a) = \mathbb{E}[R \mid s_t = s, a_t = a] \tag{1}$$

并可通过贝尔曼方程递归表示为：

$$Q^{\pi}(s,a) = \mathbb{E}_{s'}[r + \gamma \mathbb{E}_{a' \sim \pi}[Q^{\pi}(s',a')]] \tag{2}$$

训练过程中，DQN 通过最小化损失函数 $L(\theta)$ 更新网络参数 θ，损失函数被定义为：

$$L(\theta) = \mathbb{E}_{s,a,r,s'}[(Q^*(s,a \mid \theta) - (r + \gamma \max(a') Q(s',a')))^2] \tag{3}$$

其中，智能体以概率 $1-\epsilon$ 选择最大化 Q 值的动作，以概率 ϵ 随机选择动作。

1.3　注意力机制

注意力机制(Attention)[23]为神经网络模型提供了一种先进的方式来动态地聚焦于最重要的信息片段。与传统的聚集方法(比如简单的均值、最大值或拼接)相比，注意力机制通过对关键信息的软选择，实现了更为灵活和有效的信息融合。使用注意力机制时，通常包括如下步骤：特征提取、注意力分数计算和特征融合。在特征提取阶段，模型学习如何从输入数据中提取查询(Query)、键(Key)和值(Value)三种特征。然后，计算查询特征与键特征的相似度，得到注意力分数，并使用 Softmax 函数对注意力分数进行归一化得到每个值特征的权重。最后，使用注意力权重，对值特征进行加权求和得到聚合信息。注意力机制使模型能够根据给定查询相关性为不同的特征动态地分配权重，这使得信息聚集更具表达力和上下文感知性。

1.4　GAT

大多现有的图神经网络架构(如 GNN[24])在聚合信息时平等地看待所有邻居 $j \in N_i$ 信息(例如，使用均值或最大池化作为聚合操作)。不同于上述方法，GAT[25]通过引入注意力机制，对节点邻居赋予不同的重要性。具体来说，首先计算节点表示 h_i 和邻居节点表示 h_j 的注意力分数：

$$e(h_i,h_j) = \mathrm{LeakyRELU}(\boldsymbol{a}^{\mathrm{T}} \cdot [\boldsymbol{W}h_i \parallel \boldsymbol{W}h_j]) \tag{4}$$

其中 $\boldsymbol{a} \in \mathbb{R}^{2d'}$，$\boldsymbol{W} \in \mathbb{R}^{d' \times d}$ 是可学习的，“$\parallel$”表示向量的连接。得到所有邻居节点的注意力分数后，使用 Softmax 进行归一化，公式如下：

$$e_{ij} = \mathrm{Softmax}_j(e(h_i,h_j)) = \frac{\exp(e(h_i,h_j))}{\sum_{j' \in N_i} \exp(e(h_i,h_{j'}))} \tag{5}$$

最后,GAT 使用归一化的注意力系数计算邻居节点特征的加权平均值得到节点 i 的新表示:

$$\boldsymbol{h}^{'} = \sigma\left(\sum_{j \in N_i} \alpha_{ij} \cdot \boldsymbol{W} h_j\right) \tag{6}$$

1.5 Autoencoder

自动编码器(Autoencoder)[26]是一种广泛应用于无监督学习的神经网络结构,目标是通过训练数据进行有效的特征表示或编码。自动编码器通常包括 2 个主要部分:编码器(Encoder)和解码器(Decoder)。

(1)编码器(Encoder):编码器部分负责将输入数据(例如图像、文本或声音)压缩成一个较低维度的潜在表示或编码。提取输入数据中最重要的特征,并将其压缩成一个紧凑的形式。

(2)解码器(Decoder):解码器是将潜在表示重新构建或解码成与原始输入数据尽可能相似的输出。通过尝试重构输入数据,解码器学习到了如何从潜在空间中恢复数据的详细信息。

自动编码器的训练目标是最小化重构误差,即输入数据和输出数据之间的差异。

2 系统模型

多智能体场景如图 1 所示,由图 1 可知,环境中存在若干个智能体和若干个障碍物,其中每个智能体具备感知、计算、通信和决策的能力,智能体通过内部携带的各种传感器能够感知到环境感知半径以内的状态信息以及障碍物等信息,并通过通信模块与通信范围内的其他智能体进行通信交换信息,利用感知和通信阶段获取到的信息,智能体对信息进行编码融合并通过自身的决策模块进行决策,再将决策动作作用于环境。此外,门控通信机制使用融合特征信息决定是否与其他智能体进行通信并解码得到通信内容。

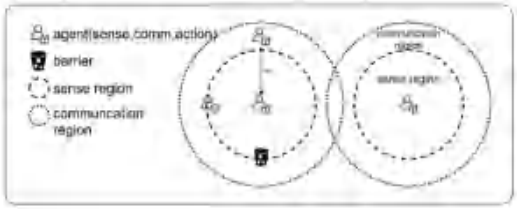


图 1 多智能体场景图

Fig. 1 Multi-agent scenario

由于环境的通信条件会限制到智能体之间信息交换的效率,导致智能体在同一时刻发送的信息量受到限制,进而导致智能体之间协同性能下降。为了解决通信受限场景的多智能体协同问题,本文提出了一个分布式多智能体强化学习模型框架 AE-CommNet,其中每个智能体都配备了决策模型,详细模型结构如图 2 所示。智能体能够综合利用历史轨迹信息、历史通信信息以及观测信息来进行决策。在此过程中,决策模型主要包含了 2 个模块,分别是特征提取及特征融合模块和决策通信编码模块。

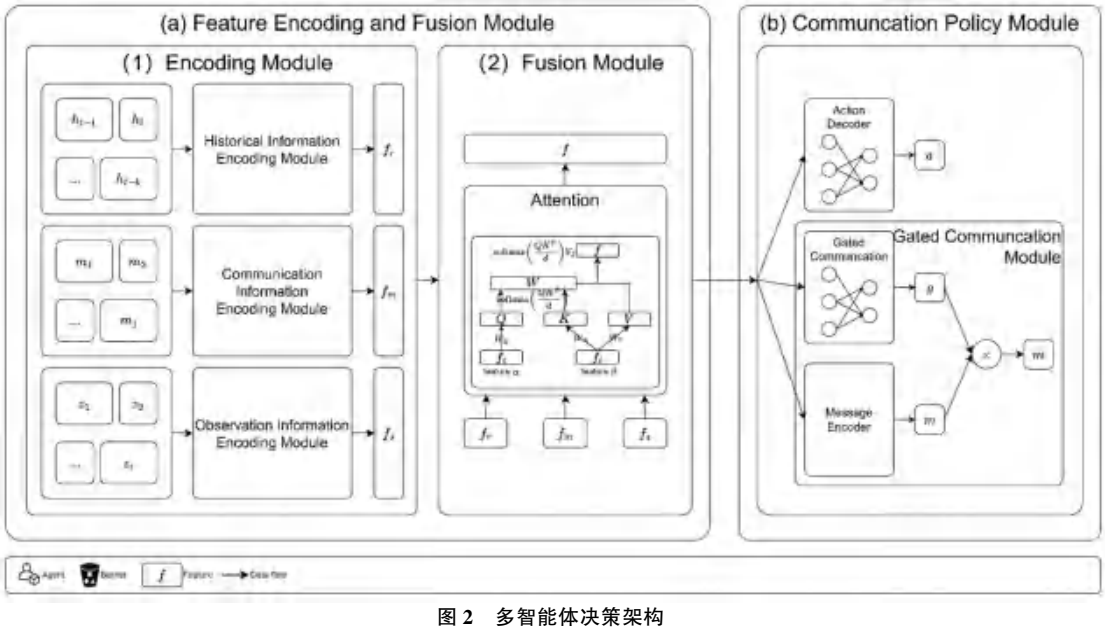


图 2 多智能体决策架构

Fig. 2 Overview of the proposed architecture for multi-agent decision-making

2.1　特征提取与融合模块

特征提取与特征融合模块(见图 2(a))可划分利用获取的多维信息提取出关键特征信息并融合，从信息维度可划分为时间维度信息和空间维度信息。

时间维度上，本文设计了一种多模块编码策略。在处理时间维度信息时，本文利用了长短期记忆网络(LSTM)，该网络能够捕捉智能体在时间序列上的依赖关系。LSTM 计算了智能体在过去 k 步内的隐状态序列 $\{h_{t-k},\cdots,h_t\}$，这些隐状态包含了智能体随时间变化的轨迹信息。为了凸显不同时间步隐状态的重要性，本文引入了注意力机制。该机制通过对 k 个隐状态 h_k 进行加权融合，赋予了关键时间步更大的权重，从而得到了时间维度的特征表示 h_{time}。这一特征不仅综合了智能体在时间维度上的轨迹信息，还强调了重要时间步的信息。通过这一编码策略，就能够更加准确地捕捉智能体的动态行为，为后续的协同决策提供有力支持。

空间维度上，本文采取了 2 种不同的编码策略。首先，针对观测信息 z，采用了多层感知机(MLP)网络，将原始观测数据从高维空间映射到低维的特征空间，实现了数据的降维和特征提取，在降低了数据处理的复杂性同时提取出了对智能体决策至关重要的特征信息。其次，针对邻近智能体的通信信息 m，本文结合了图注意力网络(GAT)。首先，计算上一时间步智能体自身的消息 m_{t-1} 与邻近智能体消息的相似性 e_{ij}。然后，通过应用 Softmax 函数，得到归一化后的权重。最后，根据这些权重，对邻近智能体的通信信息进行加权融合，得到融合后的消息特征。

通过上述时空编码模块，分别得到了时间维度和空间维度的特征。为了进一步提升智能体的感知能力，特征融合模块将不同维度的特征进行融合，形成统一的特征表示 f，为智能体的协同决策提供更为全面和准确的信息。这一融合策略不仅提高了智能体对环境的感知能力，还增强了其在复杂环境下的协同效率和鲁棒性。

2.2　通信决策模块

经过多维特征融合模块，感知和通信信息被整合为全局统一特征。这一特征作为通信决策模块的共同输入，为智能体的决策和通信提供了全面的信息基础。

在决策模块中，本文采用 MLP 网络，将融合后的特征映射到相应的动作空间 $\boldsymbol{A}$ 中，生成决策动作 $\boldsymbol{a}$。

门控通信模块如图 3 所示。本文受门控机制启发，利用一个门控网络对融合特征信息进行打分，当判断融合特征信息价值大于阈值时，消息编码模块产生的消息才会被发送给邻居智能体。具体来说，融合特征分数模块使用 MLP 生成特征分数，消息自动编码器编码生成通信信息。

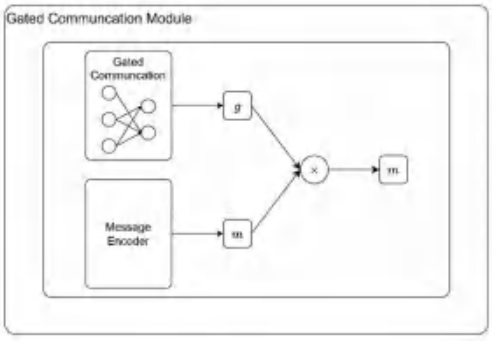


图 3　门控通信模块

Fig. 3　Gated Communication Module

消息编码模块包含了一个创新的变长自动编码器网络，由一个编码器、一个解码器和一个消息打分网络组成(见图 4)。编码器和解码器被设计成多级，实现特征变长编码及解码。消息打分网络对编码器产生的特征进行打分。详细地说，编码器首先对原始信息进行编码得到 k 个不同长度的特征向量，然后对特征向量进行零填充，得到统一维度的特征向量。接着，打分网络对特征向量进行打分，得到每一个特征的分数，得分最高的特征被送入解码器得到解码后的原始特征。最后，通过均方差损失函数训练模型。

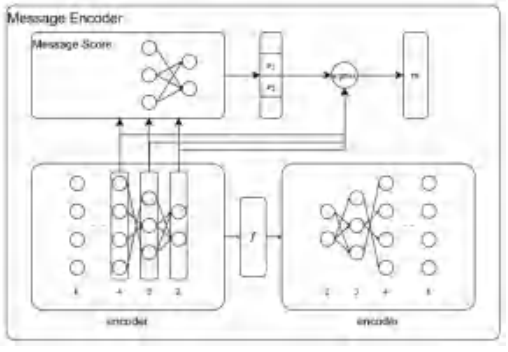


图 4　消息编码模块

Fig. 4　Message encoding module

在此基础上，智能体的决策被应用到环境中，获得奖励并观察到下一个状态。门控通信模块输出的通信信息被发送给周围邻近智能体进行后续决策和消息编码。

2.3　分布式训练和计算

在本文中，采用了分布式训练策略，每个智能体

基于自身的状态、邻近智能体的通信消息以及从环境中观测到的信息和获得的奖励,以分布式方式独立更新其策略,以最大化团队的累积折扣奖励。对此可以表示为:

$$J(\theta)=\mathbb{E}_{s\sim\rho,a\sim\pi_\theta}[R] \tag{7}$$

其中,R 表示团队的累积折扣奖励。状态信息可以包括环境状态信息 s 或单个智能体观测 z。环境以集体奖励或个体智能体奖励的加权和形式提供反馈。在实际训练时,智能体根据自身的观测和来自邻近智能体的通信信息,设定自己的优化目标,定义为:

$$J(\theta_i)=\mathbb{E}_{z\sim\rho,m\sim\rho_m,a\sim\pi_{\theta_i}}[Q_i^{\pi_i}(z,m,a)] \tag{8}$$

在算法实现上,本文借鉴了 DQN 的思想,利用时间差分方法定义损失函数,具体公式如下:

$$L(\theta_i)=\mathbb{E}_{z,m,a}[(Q_i^{\pi_i}(z,m,a)-y)^2] \tag{9}$$

其中,y 的定义公式为:

$$y=r_i+\gamma Q_i^{\pi_i}(z',m',a') \tag{10}$$

其中,r_i 表示智能体 i 获取的奖励;γ 表示折扣因子;Q' 表示下一时刻的 Q 值。通过这种方式,每个智能体独立地更新其策略,优化自身目标,以实现团队累积奖励的最大化。多智能体训练过程详见算法 1。

算法 1 多智能体训练过程

1. initialize agent network Q and the target network $\bar{Q}$
2. initialize the replay buffer D
3. for episode = 1 to max_episode do
4. reset environment and get initial state s
5. while not done do

receiver message m from neighbor agent

with probability ϵ select a random action a base on observation z and message m

execute action a in environment and observe reward r, next observation z' and terminate status done

store transition (z,a,r,z',m,done) in D

generate next message m' base on message encoder

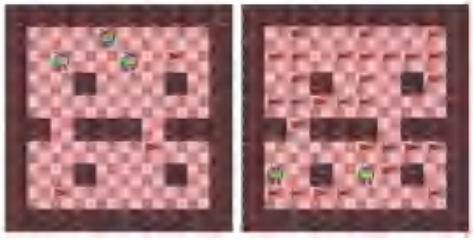
send message m' to neighbor agent when gate mechanism permits

6. end while
7. sample random minibatch of transitions $(z,a,r,z',m,m',\text{done})$ from D
8. update agent network Q
9. update agent target network $\bar{Q}$
10. end for

通过实施分布式训练,该算法允许各个智能体在综合考虑自身观测数据、与邻近智能体的通信信息以及从环境中观测的信息和接收的奖励信号的基础上,独立学习更新自身策略。这种分布式学习策略不仅加速了智能体个体策略的学习,还提高了团队协同性能。

3 实验与结果分析

在实验中,为了验证本文所提方法在多智能体合作场景中的优越性,设计了对比实验,并采用了多种评估指标进行性能分析。实验中,本文选择了 2 个具有代表性的合作场景:协同导航和协同巡检,其场景细节如图 5 所示。关于基准算法,本文选择了 IQL[27] 和 CommNet[6] 进行比较。



(a) 协同导航 (b) 协同巡检

图 5 实验场景

Fig. 5 Experimental scenarios

在协同导航场景中,存在 N 个智能体和 N 个目标点,智能体具备移动和感知环境状态的能力,并能够与通信范围内的其他智能体进行信息交换。通过协作,智能体需要在尽可能短的步数内,同时避免与其他智能体发送冲突,覆盖所有目标点。

在协同巡检场景中,存在 N 个智能体和 M 个目标点,其中 M 往往远大于 N, 智能体除了具备移动和感知能力,同时具备执行任务的能力,相比于协同导航,智能体除了到达目标点,还需要完成指定任务。通过协作,智能体需要在尽可能短的步数内,同时避免与其他智能体发送冲突,到达所有目标点并完成指定任务。

3.1 协同导航

在协同导航场景中,本文设置了 3 个智能体和 3 个目标,当智能体靠近并到达目标点时会获得奖励,离开目标点时会受到惩罚,发生碰撞时同样会受到惩罚。旨在让智能体在协同到达各自目标点的同时,最小化行动步数并完全避免碰撞。为全面评估

算法性能，本文定义了几个度量指标来衡量不同算法之间的性能，具体包括：平均任务完成率、平均任务完成步数和平均奖励。

在相同的实验设置下训练了 AE-CommNet 算法和基准算法（IQL、CommNet），并获得了训练结果。图 6 展示了不同算法在 40 000 轮游戏中获得的平均奖励。可以观察到，除了 CommNet 未完全收敛，其他算法最后均收敛到了最优附近。

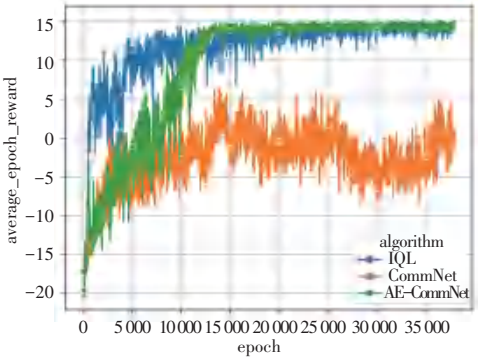


图 6　协同导航场景下的 AE-CommNet 算法与基线算法在 40 000 轮的训练奖励曲线

Fig. 6　The reward of AE-CommNet against baseline approaches on cooperative navigation after 40 000 episodes

相比于 IQL 算法，文中发现 AE-CommNet 算法虽在初始阶段收敛速度较慢，但最终均能达到最优性能。这是由于 AE-CommNet 算法在训练阶段，智能体之间需要进行通信，学习智能体之间的信息交互策略，因此会导致前期收敛较慢，而独立强化学习 IQL，在训练时不需要考虑与其他智能体的交互，智能体独立学习自身的策略，因此表现出较高的学习效率。

相比于 CommNet，AE-CommNet 算法在收敛速度和效果上均表现出优势，这是由于 AE-CommNet 算法在信息融合阶段结合注意力机制，能够区分不同信息的重要性，从而更有效地利用所接收的信息。相比之下，CommNet 的简单均值融合方式在处理信息时无法区分不同信息的重要性，影响了其学习效率。

为了验证本文提出的算法相比基线算法具备更好的协同性能，本文研究在相同的测试环境下执行 1 000 轮测试并得到了不同算法在不同指标的实验结果，具体细节见表 1。

从上述实验结果来看，本文提出的 AE-CommNet 算法在各项指标上均展现出优于 IQL 和 CommNet 算法的性能。

由于 IQL 算法是分布式的，智能体之间不需要进行信息交换，从而不消耗通信资源，但其固有的局限性也导致了协同性能的不足。由于智能体独立决策，缺乏与其他智能体的协同，容易出现贪婪选择最近目标点的情况，进而引发目标点竞争和冲突增多，任务完成率下降，平均完成任务步数增加，获取的任务奖励变少。

表 1　AE-CommNet 算法与基线算法在不同指标下的性能对比

Table 1　Performance comparison between AE-CommNet algorithm and baseline algorithms under different metrics

算法	平均任务完成率	平均完成步数	平均奖励
AE-CommNet	**0.987**	**18.531**	**14.009**
IQL	0.715	268.863	2.957
CommNet	0.765	263.239	2.105

CommNet 算法虽然引入了智能体间的信息交换，但其简单的平均融合策略在处理不同智能体的信息时存在明显不足，未能平等地对待来自不同智能体的信息。该策略未能充分考虑到不同信息在决策过程中的独特价值和重要性，同时也忽略了信息间可能存在的冗余性。因此，尽管 CommNet 算法引入了通信机制，但其协作能力的提升仍然受限，未能充分挖掘和利用多智能体协同的潜力。

为进一步验证本文提出的变长消息编码模块在协同性能与通信消耗之间的平衡优势，本文测试了不同算法在每一步的通信消耗，得到智能体在测试环境中的通信信息，并绘制了每个智能体的通信情况，如图 7 所示。

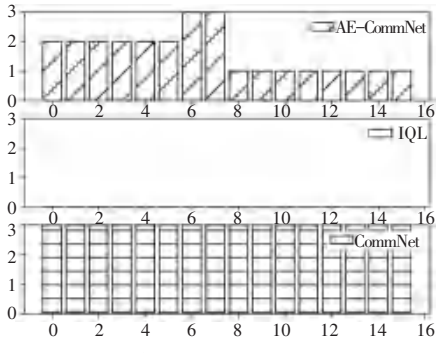


图 7　3 种算法在每个时刻平均通信消耗

Fig. 7　Average communication consumption at each time instant of three algorithms

通过对比分析实验结果，发现本文的变长消息编码模块在保持较高协同性能的同时，显著降低了通信消耗需求。相较于 CommNet，变长消息编码模块通过智能地编码信息得到变长消息，大幅减少了通信数据量；与 IQL 相比，通信信息的加入在满足通信约束的同时能够显著提高智能体协同性能。

3.2 协同巡检

在协同巡检场景中,设置了 2 个智能体和 30 个目标点,当智能体到达目标点并完成任务时会获得奖励,离开目标点时会受到惩罚。旨在让智能体以最小的冲突协同访问每个目标点并完成相应的任务。为全面评估算法性能,本文定义了如下度量指标来衡量不同算法之间的性能,包括:平均任务完成率(衡量智能体的平均协作能力)、平均完成目标数、平均完成步数(衡量智能体之间的协作效率)和平均奖励。

为进一步验证本文提出的 AE-CommNet 算法在更为复杂场景下的协同性能优势,在相同的实验设置下训练 AE-CommNet 算法和基准算法(IQL、CommNet),获得了训练结果。图 8 展示了不同算法在 40 000 轮游戏中获得的平均奖励。

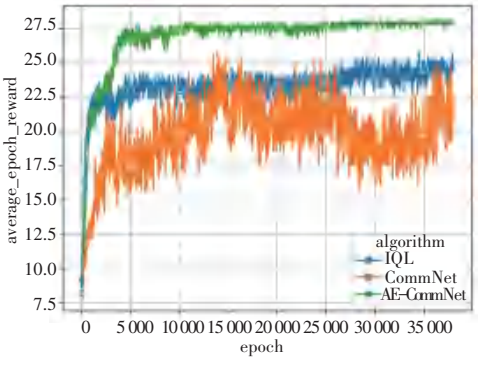


图 8 协同巡检场景下的 AE-CommNet 算法与基线算法在 40 000 轮的训练奖励曲线

Fig. 8 The reward of AE-CommNet against baseline approaches on cooperative inspection after 40 000 episodes

在这一复杂场景中,智能体不仅需要高效协作,还需面对更多的决策挑战。实验结果显示,所有算法在收敛速度上表现相当,但 IQL 因其独立强化学习的特性,在训练时仅更新自身策略,学习效率较高。然而,IQL 的最大协同奖励低于 AE-CommNet 算法,这归因于独立智能体在训练过程中因其他智能体的策略变化而导致冲突增多,协同能力下降。相比之下,AE-CommNet 算法通过智能体间的信息交流,拥有更丰富的信息量,有效减少了冲突,提升了协同水平。

CommNet 算法虽然在训练时也融合了其他智能体的信息,然而其聚合通信信息时,相比于 AE-CommNet 算法使用的注意力机制,仅使用了均值融合,不区分通信信息的重要性,无法充分利用接收到的通信信息。因此,在更加复杂的环境中,该算法在提升智能体协同效率方面的效果有限。

值得注意的是,相比协同导航场景,协同巡检场景 AE-CommNet 算法在训练过程中前期训练速度跟 IQL 相当,这显示了该算法在复杂环境的适应性,仍然能够保持不错的学习效率。

同样地,为了验证本文提出的算法相比基线算法具备更好的协同性能,分别在相同的测试环境下执行 1 000 轮测试并得到了不同算法在不同指标的实验结果,具体细节见表 2。

表 2 AE-CommNet 算法与基线算法在不同指标下的性能对比

Table 2 Performance comparison between AE - CommNet algorithm and baseline algorithm under different metrics

算法	平均任完成率	平均完成目标数	平均完成步数	平均奖励
AE-CommNet	**0.973**	**29.973**	**33.968**	**28.275**
IQL	0.871	26.132	100.000	21.132
CommNet	0.731	25.032	100.000	20.057

从上述实验结果来看,本文提出的 AE-CommNet 算法在各项指标上均展现出优于对比算法的性能。

具体而言,IQL 算法虽然在不需要通信资源方面具有优势,但在复杂环境中,其分布式特性导致智能体更倾向于贪婪地选择最近目标点,进而加剧了智能体间的竞争和冲突,使得任务完成率大幅下降,完成任务的平均步数增加,任务奖励减少。

CommNet 算法虽然实现了智能体间的信息交换,但其简单的平均信息融合在处理不同智能体信息时存在不足,在较为复杂的环境中,尤为突出,无法区分不同智能体信息的价值和冗余性。因此,简单的信息融合并不能显著提升性能,协作能力的整体提升有限。

从上述 2 个场景的实验结果及分析来看,验证了本文提出的 AE-CommNet 算法在协同场景中的卓越性能。该算法不仅在不显著降低协同性能的条件下适应通信受限的环境,而且在更加复杂的场景中表现出强大的适应性和稳定性。

4 结束语

本文深入研究了通讯受限条件下多智能体协同问题,提出了一种多智能体协同算法,该算法的核心在于智能体根据多维信息形成决策和通信信息,本文的关键创新是提出了一个变长消息编码模块,压缩通信信息的同时保留关键特征信息。实验结果表明,该算法在经典的多智能体任务中优于基线算法,

验证了算法的有效性和优越性。此外，实验结果还表明了该算法在通信受限的环境中能够在显著降低通信信息量的同时维持优秀的协同性能。

然而，尽管该算法在多智能体协同领域取得了一定成果，但仍存在一定的局限性。目前，该算法主要适用于同构多智能体协作任务，对于异构多智能体协作场景中的任务处理尚显不足。为了解决这一问题，未来的研究可以通过设计一系列的编码器将特征空间投影到统一特征空间，从而实现异构智能体之间的有效协作。

参考文献

[1] CAO Yongcan, YU Wenwu, REN Wei, et al. An overview of recent progress in the study of distributed multi-agent coordination [J]. IEEE Transactions on Industrial informatics, 2012, 9(1): 427-438.

[2] HOLLINGER G A, SINGH S. Multirobot coordination with periodic connectivity[J]. IEEE Transactions on Robotics, 2012, 28(4): 967-973.

[3] FIROOZI R, FERRANTI L, ZHANG X, et al. A distributed multi-robot coordination algorithm for navigation in tight environments[J]. arXiv preprint arXiv, 2006.11492, 2020.

[4] ZHANG Chongjie, LESSER V. Coordinating multi-agent reinforcement learning with limited communication [C]// Proceedings of 2013 International Conference on Autonomous Agents and Multi-agent Systems. Sain Paul, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2013: 1101-1108.

[5] MAO Hangyu, GONG Zhibo, ZHANG Zhengchao, et al. Learning multi-agent communication under limited-bandwidth restriction for internet packet routing[J]. arXiv preprint arXiv, 1903.05561, 2019.

[6] SUKHBAATAR S, FERGUS R. Learning multiagent communication with backpropagation[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems(NIPS'16). Barcelona, Spain: NIPS Foundation, 2016:2252-2260.

[7] QIN Jiahu, ZHENG Weixing, GAO Huijun. Sampled-data consensus for multiple agents with discrete second-order dynamics [C]//Proceedings of the 49th IEEE Conference on Decision and Control (CDC). Piscataway, NJ: IEEE, 2010: 1391-1396.

[8] ZHANG Wenbing, TANG Yang, HAN Qinglong, et al. Sampled-data consensus of linear time-varying multiagent networks with time-varying topologies[J]. IEEE Transactions on Cybernetics, 2020, 52(1): 128-137.

[9] SUN Jian, WANG Zhanshan, RONG Nannan. Sampled-data consensus of multiagent systems with switching jointly connected topologies via time-varying Lyapunov function approach[J]. International Journal of Robust and Nonlinear Control, 2020, 30(14): 5369-5385.

[10] YUN W J, LIM B, JUNG S, et al. Attention-based reinforcement learning for real-time UAV semantic communication [C]// Proceedings of the 17th International Symposium on Wireless Communication Systems (ISWCS). Piscataway, NJ: IEEE, 2021: 1-6.

[11] JIANG Jiechuan, DUN Chen, HUANG Tiejun, et al. Graph convolutional reinforcement learning[J]. arXiv preprint arXiv, 1810.09202, 2018.

[12] SHENG Junjie, WANG Xiangfeng, JIN Bo, et al. Learning structured communication for multi-agent reinforcement learning [J]. Autonomous Agents and Multi-Agent Systems, 2022, 36: 50.

[13] KONG Xiangyu, XIN Bo, LIU Fangchen, et al. Revisiting the master-slave architecture in multi-agent deep reinforcement learning[J]. arXiv preprint arXiv, 1712.07305, 2017.

[14] KILINC O, MONTANA G. Multi-agent deep reinforcement learning with extremely noisy observations[J]. arXiv preprint arXiv, 1812.00922, 2018.

[15] KIM D, MOON S, HOSTALLERO D, et al. Learning to schedule communication in multi-agent reinforcement learning [J]. arXiv preprint arXiv, 1902.01554, 2019.

[16] ZHANG Saiqian, ZHANG Qi, LIN Jieyu. Efficient communication in multi-agent reinforcement learning via variance based control [J]. arXiv preprint arXiv, 1909.02682, 2019.

[17] AGARWAL A, KUMAR S, SYCARA K. Learning transferable cooperative behavior in multi-agent teams[J]. arXiv preprint arXiv, 1906.01202, 2019.

[18] AGARWAL M, AGGARWAL V, AZIZZADENESHELI K. Multi-agent multi-armed bandits with limited communication[J]. Journal of Machine Learning Research, 2022, 23(1): 24.

[19] SINGH A, JAIN T, SUKHBAATAR S. Learning when to communicate at scale in multiagent cooperative and competitive tasks [J]. arXiv preprint arXiv, 1812.09755, 2018.

[20] KEPLER M E, STILWELL D J. An approach to reduce communication for multi-agent mapping applications[C]//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2020: 4814-4820.

[21] OLIEHOEK F A, AMATO C. A concise introduction to decentralized POMDPs[M]. Cham: Springer, 2016.

[22] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv, 1312.5602, 2013.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS Foundation, 2017: 5998-6008.

[24] XU Keyulu, HU Weihua, LESKOVEC J, et al. How powerful are graph neural networks? [J]. arXiv preprint arXiv, 1810.00826, 2018.

[25] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint arXiv, 1710.10903, 2017.

[26] MASCI J, MEIER U, CIRE A D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction[M]//HONKELE T, DUCH W, GIROLAMI M, et al. Artificial Neural Networks and Machine Learning. Lecture Notes in Computer Science. Cham: Springer, 2011, 6397: 52-59.

[27] ABED-ALGUNI B H K. Cooperative reinforcement learning for independent learners[EB/OL]. (2014-01-01). https://api.semanticscholar.org/CorpusID:113560486.