基于价值函数分解和通信学习机制的异构多智能体强化学习方法

杜 威[1] 丁世飞[1),2)] 郭丽丽[1),2)] 张 健[1),2)] 丁 玲[3)]

[1)](中国矿业大学计算机科学与技术学院 江苏 徐州 221116)

[2)](矿山数字化教育部工程研究中心(中国矿业大学) 江苏 徐州 221116)

[3)](天津大学智能与计算学部 天津 300350)

摘 要 许多现实世界的系统可以被建模为多智能体系统,多智能体强化学习为开发这些系统提供了一种有效的方法,其中基于集中训练与分散执行范式的价值函数分解方法得到了广泛的研究.然而现有的价值分解方法一般缺乏通信机制,在处理需要通信学习的多智能体任务时表现不佳.同时,目前大多数通信机制都是针对同构多智能体环境设计的,没有考虑异构多智能体场景.在异构场景中,由于智能体动作空间或观测空间的异构性,智能体之间的信息共享并不直接.如果不能对智能体的异构性进行有效地建模处理,通信机制将变得无效,甚至会影响多智能体的协作性能.为了应对这些挑战,本文提出一个融合价值函数分解和通信学习机制的异构多智能体强化学习框架.具体地:(1)与采用同构图卷积网络的方法不同,该框架利用异构图卷积网络融合智能体的异构特征信息得到有效的嵌入;(2)利用通信学习模块获得的嵌入信息和局部观测历史计算每个智能体的动作价值,以选择和协调智能体的动作;(3)通过设计的互信息损失函数和价值函数分解模块的损失函数联合训练,能够有效地训练整个方法.本文首先在两个异构多智能体平台上进行实验,实验结果表明该方法能学到比基线方法更有效的策略,在两个平台上相比基线方法分别提高了 13%的平均奖励值和 24%的平均胜率.此外,在交通信号控制场景中验证了该方法在现实系统中的可行性.

关键词 价值函数分解;异构多智能体强化学习;通信机制;图神经网络;互信息;交通信号控制

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2024.01304

Heterogeneous Multi-Agent Reinforcement Learning Method Based on Value Function Decomposition and Communication Learning Mechanism

DU Wei[1] DING Shi-Fei[1),2)] GUO Li-Li[1),2)] ZHANG Jian[1),2)] DING Ling[3)]

[1)](*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116)

[2)](*Mine Digitization Engineering Research Center of Ministry of Education (China University of Mining and Technology), Xuzhou, Jiangsu* 221116)

[3)](*College of Intelligence and Computing, Tianjin University, Tianjin* 300350)

Abstract Many real-world systems can be modeled as multi-agent systems in which multiple agents interact with the environment to learn and make decisions. Reinforcement learning has received wide attention recently and has achieved remarkable success in various fields. As practical tasks usually involve multiple agents interacting with the environment, multi-agent reinforcement learning has gradually become a research focus. Multi-agent reinforcement learning provides an effective way to develop these multi-agent systems and has achieved remarkable results in various complex sequential decision-making tasks. However, multi-agent reinforcement learning faces

收稿日期:2023-04-13;在线发布日期:2024-04-03.本课题得到国家自然科学基金项目(62276265,61976216)资助.**杜 威**,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为深度学习、强化学习.E-mail:1394471165@qq.com.**丁世飞**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为人工智能、模式识别.E-mail:dingsf@cumt.edu.cn.**郭丽丽**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为深度学习、情感计算.**张 健**,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为机器学习、模式识别.**丁 玲**,博士研究生,主要研究方向为深度学习、图机器学习.

many challenges such as non-stationarity and dimensional curse. The value function decomposition method is one of the most popular MARL methods. By decomposing the global value function into the local individual value function, the value function decomposition method reduces the dimension of the action space to a great extent and alleviates the dimensional curse problem. In addition, agents can select actions only according to individual value functions, which solves the non-stationarity problem caused by the interaction between agents. Value function decomposition method based on centralized training and decentralized execution paradigm has been widely studied. However, the existing value decomposition methods generally lack communication mechanisms and perform poorly when dealing with multi-agent tasks requiring communication learning. At the same time, most of the current communication learning mechanisms are designed for homogeneous multi-agent environments, without considering heterogeneous multi-agent scenarios. In heterogeneous scenarios, information sharing between agents is not direct because of the heterogeneity of the agent's action space or observation space. If the heterogeneity of agents cannot be modeled effectively, the communication mechanism will become ineffective and even affect the performance of multi-agent cooperation. To address these challenges, this paper proposes a heterogeneous multi-agent reinforcement learning framework that integrates value function decomposition and communication learning mechanisms. Specifically, (1) Different from the method using the homogeneous graph convolutional network, the framework utilizes the heterogeneous graph convolutional network to integrate the heterogeneous feature information of the agent to get effective embedding. (2) The embedding information and local observation history obtained by the communication learning module are used to calculate the action value of each agent to select and coordinate the actions of the agents. (3) Through the joint training of loss function of mutual information and value function decomposition, the proposed method can be effectively trained. The proposed method maintains the advantages of scalability and stability of value function decomposition and promotes better collaboration and decision-making of agents by utilizing diverse information interactions between heterogeneous agents. To the best of our knowledge, our work is the first attempt to combine the communication learning method based on graph convolution network and the value function learning method to develop the heterogeneous multi-agent system. The proposed framework provides a new idea for the field of heterogeneous multi-agent reinforcement learning. This paper first conducts experiments on two heterogeneous multi-agent platforms, and the experimental results show that the proposed method can learn more effective strategies than the baseline method, and the average reward value and average win rate of 13% and 24% respectively on the two platforms compared with the baseline method. In addition, the feasibility of this method in the real system is verified in the traffic signal control scenario.

Keywords　value function decomposition; heterogeneous multi-agent reinforcement learning; communication mechanism; graph neural network; mutual information; traffic signal control

1　引　言

多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)作为一种解决许多现实世界问题的关键工具受到了广泛的关注[1-3].集中训练与分散执行框架(Centralized Training with Decentralized Execution,CTDE)由于其处理局部可观测性约束的能力被广泛应用于 MARL 领域.价值函数分解方法为进一步解决该框架的可扩展性问题提供了有效的解决方案.不断涌现的价值函数分解方法[4-7]逐渐提高了价值分解的表示能力.然而,大多数价值函数分解方法侧重于完全分解的方式,即每个智能体仅根据其局部观测计算局部价值函数.在现实世界中,多智能体任务通常需要智能体之间的通信,以有效地协调智能体的动作.在缺乏通信机制

的情况下,智能体受到局部可观测性和随机性的限制.这种限制会加剧分散执行阶段智能体的状态和动作对其他智能体的不确定性,导致严重的智能体动作失调.

虽然通信机制在 MARL 中的应用已经非常普遍,但以往的大多数工作并没有考虑把通信学习机制引入至价值函数分解方法.同时,大多数通信学习方法只关注同构智能体环境下的智能体的交互,而未考虑异构智能体环境.异构智能体环境下的智能体可能有不同的状态空间和动作空间,以捕食者-捕获者-猎物环境为例,该环境中友方智能体包括捕食者和捕获者,其目标是合作捕获敌方猎物.捕食者能够观测环境但是捕获者不能观测环境,同时捕食者只能确定猎物的位置而不能捕获猎物,而捕获者的动作空间拥有捕获猎物这一动作设定.感知智能体(捕食者仅具有感知能力)和行动智能体(捕获者具有附加的动作能力)必须协作才能完成任务.然而目前的多智能体通信学习方法没有明确地对这种异构智能体的异构性进行建模处理.如果没有特定的异构通信机制,不同类型的智能体可能无法利用异构性信息,进而无法提取有价值的信息用于价值估计和决策.因此通信可能变得毫无用处,甚至可能降低性能.

为解决这些挑战,本文提出一个融合价值函数分解和通信学习机制的异构多智能体强化学习框架.该方法既保持了价值函数分解的可扩展性和稳定性,又利用异构智能体之间的信息交互促进了智能体之间更好的协作和决策.该框架可以很有效地将提出的异构通信机制与现有的各种价值函数分解方法融合.关于异构通信学习机制,本文设计一个基于异构图卷积网络的通信学习模型,构造考虑智能体异构性的通信信道.此外,本文展示如何将通信模块与现有的价值函数分解方法融合,并通过最大化输出信息和局部价值函数之间互信息,使智能体可以学到最有效和最有表达性的信息.本文的主要贡献总结如下:

(1)提出第一个融合价值分解与通信学习的异构多智能体强化学习方法(Heterogeneous multi-agent reinforcement learning method based on Value Decomposition and Communication learning,HVDC),可以同时解决异构智能体的通信问题和价值分解方法的智能体动作不协调问题.

(2)利用异构图建模多智能体系统,并针对不同类型的智能体,设计不同的通信信道.利用异构图卷积网络实现了异构信息的融合和传递,建立了一种有效的异构多智能体通信机制.

(3)首次引入通信模块输出信息和局部价值函数之间的互信息,并利用互信息最大化优化,使智能体可以从不同类型的信息中学到最有价值和有表达性的信息,以更好地协调动作.

(4)在多个异构智能体平台环境上验证所提出的方法的有效性.此外,构建在现实的异构多智能体系统上的实验,以证明所提出的方法在现实场景中的可行性和优越性.

本文在第 2 节中,介绍了价值函数分解、通信学习和图卷积神经网络,这是 HVDC 的必要背景;在第 3 节对 HVDC 展开详细描述,包括方法总体框架、通信学习模块、价值分解模块以及模型的算法细节;第 4 节展示该方法在各种异构多智能体环境下的实验,结果证明了提出的方法的优越性;最后在第 5 节总结全文.

2 背景知识

2.1 异构多智能体问题

在本节中考虑异构多智能体问题,它可以被建模为异构多智能体-局部可观测马尔可夫决策过程(Heterogeneous Multi Agent Partially Observable Markov Decision Process,HMA-POMDP),由多元组 $G=\langle C,I,S,A,R,P,\Omega,O,\gamma\rangle$ 表示.C 是异构多智能体设置中所有智能体类别的有限集,索引 $j\in C$ 表示智能体属于哪个类别.$I=\sum_{j\in C}I^j$ 表示环境中相互作用的智能体的有限集,其中 I^j 表示属于类 j 的智能体.$S=\{S^j\}_{j\in C}$ 是联合状态空间,其中 S^j 表示属于类 j 的智能体的状态空间.$A=\{A^j\}_{j\in C}$ 表示联合动作空间,其中 A^j 表示属于类 j 的智能体的动作空间.对于每个 A^j 有 $A^j=[a_i^j]|_{i=1}^{I^j}$,其中 a_i^j 属于类 j 的智能体 i 的动作.$\Omega=\{\Omega^j\}_{j\in C}$ 是观测空间,其中 Ω^j 表示属于类 j 的智能体的观测空间.

在每个时间步中,如果环境观测是可用的,则每个智能体 i 根据类 j 的特定观测函数 $O^j(s,i)$ 接收到一个局部观测 $o_i^j\in\Omega$.在每个时间步中,类别为 j 的智能体 i 选择一个动作 $a_i^j\in A$,形成一个联合动作 $a\in A$,并获得即时共享奖励 $R(s,a)$.这种共享奖励[13]鼓励智能体之间的合作和团队行为,并基于状态转移函数 $P(s'|s,a)$ 导致联合状态转变为下一个状态.智能体的目标是学习最优联合策略 $\pi^*(s):S\rightarrow A$.由联合策略可得联合动作价值函数 $Q_{tot}^{\pi}(\tau,a)=\mathbb{E}_{\pi(s)}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$,其中 τ 为联合动作观测历史,$\gamma\in[0,1]$

为时间折扣因子.

2.2　价值函数分解

在多智能体场景下，学习最优价值函数通常很困难. 为更好地协调合作智能体的动作，学习一个集中的联合价值函数似乎是一个很理想的解决方案. 然而，由于智能体的联合动作空间随着智能体数量的增加呈指数增长，中心化的全局价值函数难以学习. 相反，直接学习每个智能体的分散的局部价值函数可以缓解可扩展性问题. 然而，分散学习方法通常忽略了智能体之间的相互作用，这往往导致智能体之间的动作协调失调和策略次优.

为解决这一困境，价值分解方法将全局价值函数 Q_{tot} 表示为局部价值函数的组合，在复杂任务中表现出了有效性，如 VDN(Value Decomposition Network)[4]、QMIX[5]. 这些价值分解方法都遵循 IGM(Individual-Global-Max)原则使用联合全局价值函数 $Q_{tot}(\tau,a)$ 和个体局部价值函数 $[Q_i(\tau_i,a_i)]_{i=1}^n$ 来保证联合最优动作和局部最优动作选择之间的一致性：

$$\arg\max_{a\in A} Q_{tot}(\tau,a)=\left[\arg\max_{a_i\in A} Q_i(\tau_i,a_i)\right]_{i=1}^n \tag{1}$$

VDN 利用可加性对全局价值函数进行分解：

$$Q_{tot}^{\mathrm{VDN}}(\tau,a)=\sum_{i=1}^{n} Q_i(\tau_i,a_i) \tag{2}$$

QMIX 通过单调性条件约束全局价值函数：

$$\forall i\in N,\ \frac{\partial Q_{tot}^{\mathrm{QMIX}}(\tau,a)}{\partial Q_i(\tau_i,a_i)}>0 \tag{3}$$

价值函数分解方法之间的差异在于混合网络，目前的价值函数分解方法其混合网络的表征复杂性不断增加. VDN 将全局价值函数分解为局部价值的累加. 然而，VDN 设置了严格的累加假设，忽略了在整个训练阶段中可访问的任何附加信息. 针对 VDN 的局限性，QMIX 利用一个单调混合网络结构来估计全局价值，该结构是基于局部观测的局部价值函数的单调非线性聚合. QTRAN(Q-learning with Transformation)[6] 使用一种更通用的方式分解价值函数，避免了累加性或单调性的结构约束. QPLEX (duplex dueling multi-agent Q-learning)[7] 提出了一种新的网络结构来分解价值函数，将 IGM 原理转化为易于实现的优势函数约束，实现对价值函数的高效学习.

然而，目前的价值函数分解方法主要集中在完全分解的方式上，通过先学习分散的局部价值函数，再借助混合网络完成智能体协调动作的任务. 对于具有许多异构智能体和局部可观测的场景，尽管混合网络具有一定的表示能力，但完成智能体之间的动作协调任务是不够的. 完全分解操作切断了去中心化的局部价值函数之间的依赖关系，智能体容易对其他智能体的行为和状态产生不确定性. 随着时间的推移，这种不确定性会增加，导致分散执行中严重的智能体动作失调和性能下降. 本文提出的 HVDC 框架采用价值分解学习范式，但侧重于通过通信机制增强个体局部网络的学习能力. 各种价值分解方法如 VDN、QMIX 的混合网络可灵活应用于 HVDC 框架.

2.3　通信学习机制

通信学习是多智能体强化学习中的一个研究热点问题. 具有可区分通信信道的端到端学习是目前比较流行的一种学习方法. Sukhbaatar 等人[8] 以及 Das 等人[9] 专注于学习去中心化的通信机制，并解决了何时以及和谁进行通信的问题. Foerster 等人[10] 和 Das 等人[11] 研究多智能体学习背景下自然语言的出现. Singh 等人[12] 学习门控机制来控制智能体间的持续通信. Kim 等人[13] 研究了有限通信机制下的智能体的动作协调问题.

与本文最相关的工作是 TarMAC(Targeted Multi-Agent Communication)[9] 和 NDQ(Nearly Decomposable Q-function)[14]. TarMAC 使用注意力机制来区分传入信息的重要性. 与之相比，本文引入注意力机制来区分异构智能体信息的重要程度. NDQ[14] 提出了一种新的 MARL 框架，通过所提出的通信最小化机制辅助价值函数分解. 在这个框架中，智能体在大多数情况下独立选择动作，但偶尔会和其他智能体进行有效的通信. 与之相比，本文的方法利用互信息融合了价值函数分解和通信学习机制，通过最大化互信息使智能体可以隐式地学得智能体在何时得到哪个或哪类智能体的信息更有价值，使智能体能够从不同类型的信息中学到最有表达性的信息，以更好地协调动作.

2.4　图卷积神经网络

图卷积神经网络(Graph Convolutional Network, GCN)通常利用节点之间的信息传递来获取图节点之间的结构依赖关系，每个节点聚合相邻节点的特征向量来计算一个新的特征向量. 一般特征更新过程如式(4)所示.

$$\bar{\boldsymbol{h}}_i'=\sigma\Big(\sum_{k\in N(i)}\omega\bar{\boldsymbol{h}}_b/c_{ib}\Big) \tag{4}$$

式中 $\bar{\boldsymbol{h}}_i'$ 表示节点 i 更新后的特征向量，$\sigma(\cdot)$ 表示激活函数，ω 表示可学习的权值. $b\in N(i)$ 由节点 i 的邻居节点组成，其中 b 表示邻居的索引. c_{ib} 表示基于图结构的归一化项. c_{ib} 有很多选项，一个常见的选项

是$\sqrt{|N(i)N(b)|}$.

通过L层聚合后,节点i的特征表示聚合关于节点i可到达的节点的结构信息.然而,c_{ib}的结构依赖性会影响GCN对不同规模图的泛化能力.因此,对式(4)的有效改进是用式(5)计算的注意系数α_{ib}来代替c_{ib}.

$$\alpha_{ib}=\mathrm{softmax}_b(\sigma'(\bar{a}^{\mathrm{T}}[\omega\bar{\boldsymbol{h}}_i\|\omega\bar{\boldsymbol{h}}_b])) \tag{5}$$

式中$\bar{a}$表示可学习权值,σ'表示LeakyReLU非线性函数,"‖"表示拼接操作.利用softmax操作对所有邻居b的系数进行归一化.

对于具有不同节点和边类型的异构图,异构GCN通过学习不同边类型的信息传递和不同节点类型的特征约简机制,可以直接对异构图进行操作[15].与同构GCN相比,异构GCN具有更好的可解释性和表达性[16-17].有些工作尝试使用异构图神经网络来学习异构环境下的智能体之间的通信.HMAGQ-Net[18]采用有向标记图来处理和表示环境状态,进而使用关系图卷积层实现异构智能体之间的通信学习.HetNet[19]使用异构图注意力网络学习高效和多样化的通信,以协调异构智能体完成协作任务.但这些方法没有充分利用价值函数分解和互信息优化的优势,导致通信学习和价值学习不够有效.与之相比,本文的工作融合价值函数分解、通信学习机制和互信息最大化优化,可以实现异构智能体场景下有效的价值学习和通信学习.

3 方 法

在本节中,提出一个结合价值函数分解和通信学习机制的异构多智能体强化学习框架,以解决异构智能体的通信问题和价值分解方法的动作协调问题.首先阐述所提出的框架如何建模异构智能体场景以及框架各部分的组成.然后,介绍所提出的基于异构图卷积网络的通信机制,以及相应的互信息优化机制和全局优化目标.最后展示了算法的细节.

3.1 方法框架

如图1所示,在所提出的框架中,对于属于类别j的智能体i,首先得到局部观测o_i,利用多层感知机(Multilayer Perceptron,MLP)和门控循环神经网络(Gated Recurrent Unit,GRU)对局部观测o_i进行编码处理得到特征$\boldsymbol{h}_i$,然后将特征$\boldsymbol{h}_i$输入到异构图卷积网络通信模块.通信模块采用异构图建模异构智能体,并根据智能体的不同类别学习不同的通信策略.

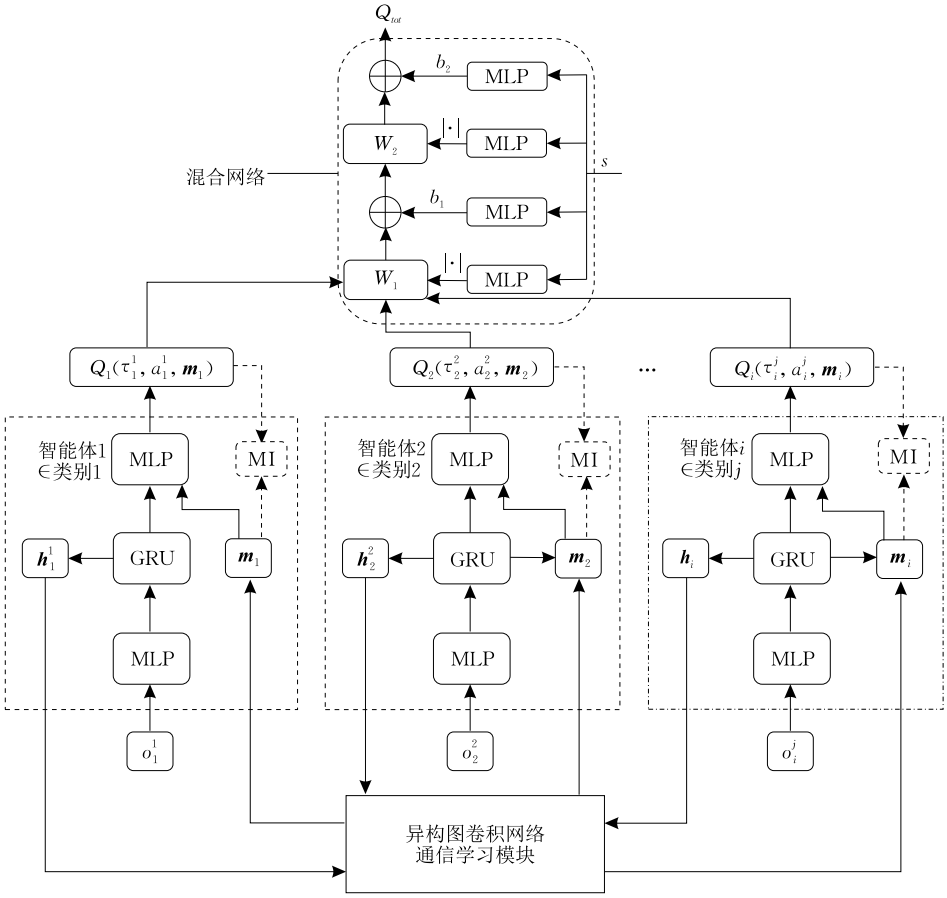


图1 基于价值函数分解和通信学习机制的异构多智能体强化学习方法的框架

具体地，利用在 2.1 节中定义的 HMA-POMDP，将异构智能体环境中的每一类智能体建模为异构图中一个独特类型的节点. 与同构图相比，异构图具有不同类型的节点和边. 这些不同类型的节点和边通常具有不同类型的属性，这些属性可以捕获不同节点和边的特征. 这一优势提高了异构图的图表达能力，使复杂的异构多智能体场景可以被直接建模.

异构图建模完成后，通过异构图神经网络融合异构智能体的特征信息得到信息嵌入. 然后，智能体 i 的个体局部价值函数 $Q_i(\tau_i, a_i, \boldsymbol{m}_i)$ 基于局部观测历史 τ_i 和从通信模块接收到的信息嵌入 $\boldsymbol{m}_i$ 得到个体局部 Q 值. 最终将不同智能体得到的局部 Q 值输入到混合网络中，得到全局价值函数的估计. 其中混合网络可以采用各种价值函数分解方法中的混合网络，图 1 的混合网络模块使用的是 QMIX 中的混合网络，混合网络只包含 MLP，但混合网络的权值和偏置是由超网络[5]获得的，以满足单调性约束. 本文方法采用集中培训和分散执行的范式. 在集中训练阶段，假设该方法可以获得个体的局部观测历史. 在分散执行阶段，各智能体根据通信信道进行去中心化的通信和执行动作.

3.2　通信机制

本小节详细介绍如何利用异构图卷积网络建立智能体之间的通信学习机制，该通信学习模块的框架如图 2 所示.

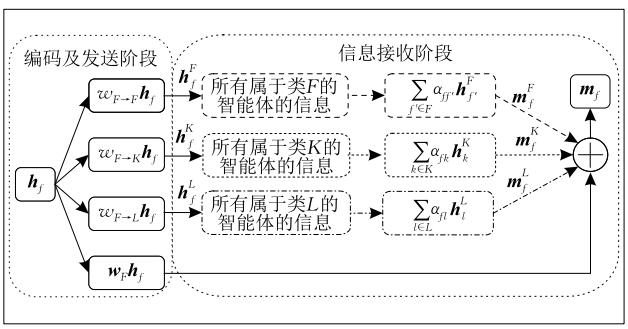


图 2　异构图卷积网络通信模块

为了符号的简单性和不失普遍性，考虑一个有三种类型智能体的场景 $C=\langle F, K, L\rangle$. 图 2 展示了信息编码和发送阶段以及信息接收阶段中信息的计算过程，以 F 类智能体为例. 对于 F 类的智能体 f，在信息编码和发送阶段中对其输入 $\boldsymbol{h}_f$ 进行不同的权值矩阵处理，如图 2 所示. 一方面，利用类特定的权值矩阵 $\boldsymbol{w}_F \in \mathbb{R}^{d' \times d}$ 对 $\boldsymbol{h}_f$ 进行处理，其中 d 表示输入特征的维数，d' 表示输出特征的维数. 另一方面，每种类型的异构边（异构信道）都利用特定的权值矩阵 $\boldsymbol{w}_{Edgetype} \in \mathbb{R}^{d'' \times d}$ 来处理输入特征 $\boldsymbol{h}_f$.

如图 2 所示，$w_{F\to K}$ 表示类 K 的智能体到类 F 的智能体之间的通信信道类型，然后将计算得到的向量 $\boldsymbol{h}_f^K = w_{F\to K}\boldsymbol{h}_f$ 发送到目标智能体 N_d，其中 d'' 表示输出特征维数. 在信息接收阶段，对于 F 类智能体 f 所连接的每一类通信信道，利用异构图卷积网络计算每一类信道的聚合特征. 它是通过对相同信道类型的邻居智能体接收到的信息进行加权来计算的，$\alpha_{Edgetype}$ 表示归一化注意力系数. 然后，将聚合结果与特征 $\boldsymbol{w}_F\boldsymbol{h}_f$ 相结合，计算融合信息表示.

因此对于智能体 f，最终得到的信息表示为

$$\boldsymbol{m}_f = \sigma\Big(\boldsymbol{w}_F\boldsymbol{h}_f + \sum_{f'\in N_f(F)} \alpha_{ff'}\boldsymbol{h}_{f'}^F + \sum_{k\in N_f(K)} \alpha_{fk}\boldsymbol{h}_k^F + \sum_{l\in N_f(L)} \alpha_{fl}\boldsymbol{h}_l^F\Big) \tag{6}$$

其中，$N_f(F)$，$N_f(K)$ 和 $N_f(L)$ 分别表示类 F，K 和 L 的邻居智能体.

需要注意的是，由于考虑了异构通信信道，在计算异构图的注意力系数时，将式(5)改写成式(7)～(9).

$$\alpha_{ff'} = \mathrm{softmax}_{f'}(\sigma'(\bar{a}^{\mathrm{T}}[\boldsymbol{\omega}_F\bar{\boldsymbol{h}}_f \| \boldsymbol{\omega}_{F\to F}\bar{\boldsymbol{h}}_{f'}])) \tag{7}$$

$$\alpha_{fk} = \mathrm{softmax}_k(\sigma'(\bar{a}^{\mathrm{T}}[\boldsymbol{\omega}_F\bar{\boldsymbol{h}}_f \| \boldsymbol{\omega}_{K\to F}\bar{\boldsymbol{h}}_k])) \tag{8}$$

$$\alpha_{fl} = \mathrm{softmax}_l(\sigma'(\bar{a}^{\mathrm{T}}[\boldsymbol{\omega}_F\bar{\boldsymbol{h}}_f \| \boldsymbol{\omega}_{L\to F}\bar{\boldsymbol{h}}_l])) \tag{9}$$

使用 $\boldsymbol{m}_f^F = \sum_{f'\in N(F)} \alpha_{ff'}\boldsymbol{h}_{f'}^F$ 来表示智能体 f 从类 F 的邻居智能体获得的聚合信息，则式(6)可以改写为

$$\boldsymbol{m}_f = \sigma(\boldsymbol{w}_F\boldsymbol{h}_f + \boldsymbol{m}_f^F + \boldsymbol{m}_f^K + \boldsymbol{m}_f^L) \tag{10}$$

对属于类 K 的智能体 k 和属于类 L 的智能体 l 进行类似的计算过程，更新方程如下：

$$\begin{aligned}\boldsymbol{m}_k &= \sigma\Big(w_K\boldsymbol{h}_k + \sum_{k'\in N_k(K)} \alpha_{kk'}\boldsymbol{m}_{k'}^K + \sum_{f\in N_k(F)} \alpha_{kf}\boldsymbol{m}_f^K + \sum_{l\in N_k(L)} \alpha_{kl}\boldsymbol{m}_l^K\Big)\\ &= \sigma(w_K\boldsymbol{h}_k + \boldsymbol{m}_k^K + \boldsymbol{m}_k^F + \boldsymbol{m}_k^L)\end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{m}_l &= \sigma\Big(w_L\boldsymbol{h}_l + \sum_{l'\in N_l(L)} \alpha_{ll'}\boldsymbol{m}_{l'}^L + \sum_{f\in N_l(F)} \alpha_{lf}\boldsymbol{m}_f^L + \sum_{k\in N_l(K)} \alpha_{lk}\boldsymbol{m}_k^L\Big)\\ &= \sigma(w_L\boldsymbol{h}_l + \boldsymbol{m}_l^L + \boldsymbol{m}_l^F + \boldsymbol{m}_l^K)\end{aligned} \tag{12}$$

由此，智能体 f，k 和 l 分别得到融合其他各类邻居智能体特征信息的信息表示 $\boldsymbol{m}_f$，$\boldsymbol{m}_k$ 和 $\boldsymbol{m}_l$. 此外，式(11)和式(12)中注意力系数的计算方法与式(7)～(9)相似. 在大规模异构多智能体场景中，考虑所有其他智能体成本很高，而且效果不好. 这是因为接收所有智能体的信息需要较高的计算复杂度，且智能体无法从大量全局共享的信息中区分有价值的信息. 由于图卷积运算可以逐渐扩大智能体的通信范围，因此，只考虑能够共享信息的邻居智能体是高效且有效的. 这背后的逻辑和直觉是相邻的智能体更有可能相互作用和影响.

在本文的方法中，邻居智能体的定义取决于实

验场景.不同的异构场景有不同的邻居定义.在 4.1 节中的捕食者-捕获者-猎物环境和 4.2 节中的星际争霸环境中,在每个时间步中,本文将智能体的邻居智能体定义为此智能体视野范围内的其他智能体.在 4.3 节中的交通信号控制环境中,本文将智能体的邻居智能体定义为此智能体的特定距离范围内的其他智能体.

在每个时间步中,异构图卷积层对应着智能体与邻居智能体之间的一轮信息交换.通过叠加多个异构图卷积层进行多轮通信,可以获取高水平的信息表示.最后,为了稳定学习过程,利用多头注意力机制[20-21]使其适应异构环境.本文利用 M 个独立的异构注意力层并行计算智能体的特征,然后将计算结果利用拼接操作得到异构图卷积网络中每个层的输出.此操作可以将每一种类型的通信信道再划分为 M 个独立的子信道.

3.3 互信息优化

本节将互信息(Mutual Information,MI)引入到提出的框架中,以实现智能体之间更有效的通信学习.对于智能体 i,它接收来自各类的邻居智能体的信息$(\boldsymbol{m}_i^1,\boldsymbol{m}_i^2,\cdots,\boldsymbol{m}_i^c)$,并将它们融合以获得最终的信息表示 $\boldsymbol{m}_i$.信息表示 $\boldsymbol{m}_i$ 与局部观测历史 τ_i 拼接,作为局部价值函数的输入,并输出局部价值.然后将局部价值输入混合网络(如 QMIX 的混合网络),估计全局价值.对于智能体 i,引入互信息来衡量不同类别的邻居智能体的信息$(\boldsymbol{m}_i^1,\boldsymbol{m}_i^2,\cdots,\boldsymbol{m}_i^c)$在不同时刻对智能体局部价值函数 Q_i 的影响,以学得最有表达性和最有效的信息表示 $\boldsymbol{m}_i$.

接下来定义和计算互信息.首先,利用重启随机游走方法(Random Walk with Restart,RWR)[22]对智能体的邻居智能体(视野范围内或物理范围的智能体)再次进行固定数量的采样处理.具体采样过程如下:(1)从智能体 i 开始随机游走,采用概率 p 选择智能体,然后放入集合 Z_i 中.它包含固定数量的智能体,同时限制不同类别的邻居智能体数量,以保证初始邻居智能体中所有类型的智能体都包含在新的集合 Z_i 中;(2)对 Z_i 中的智能体通过类进行分组.对于每个智能体类 c,根据频率从 Z_i 中选取前 k_c 个智能体,并将其视为智能体 i 的 c-类邻居集合,表示为 $N_i(c)$.

互信息可以通过学习判别器 $\mathcal{D}$ 来估计.判别器被用来区分正样本$(\boldsymbol{m}_i^c,Q_i)$和负样本$(\widetilde{\boldsymbol{m}}_i^c,Q_i)$.最大化智能体 i 接收的 c 类信息表示 $\boldsymbol{m}_i^c$ 与智能体局部价值 Q_i 之间的互信息.智能体 i 的判别器 $\mathcal{D}_i$ 用于对"信息-Q 值对"进行评分,利用一个双线性函数作为评分函数,其定义如下:

$$\mathcal{D}_i(\boldsymbol{m}_i^c,Q_i)=\sigma[(\boldsymbol{m}_i^c)^{\mathrm{T}}\boldsymbol{M}_i^cQ_i] \tag{13}$$

σ 是激活函数,$Q_i\in\mathrm{R}^{N\times1}$,$\boldsymbol{M}_i^c\in\mathrm{R}^{d''\times1}$ 是可学习的评分向量.

$$L_{\mathrm{MI}}=\sum_{i\in I^+}\log\mathcal{D}_i(\boldsymbol{m}_i,Q_i)+\sum_{i\in I^-}[1-\log\mathcal{D}_i(\widetilde{\boldsymbol{m}}_i,Q_i)] \tag{14}$$

因此,根据式(14),通过 N 个智能体的信息-Q 值损失和来最大化互信息,损失函数如下:

$$L_{\mathrm{MI}}=\sum_{i=1}^{N}\sum_{c=1}^{C}L_i^c \tag{15}$$

$$L_i^c=\sum_{\langle c,j\rangle\in N_i^+}\log\mathcal{D}_i(\boldsymbol{m}_i^c,Q_i)+\sum_{\langle c,j\rangle\in N_i^-}\log[1-\mathcal{D}_i(\widetilde{\boldsymbol{m}}_i^c,Q_i)] \tag{16}$$

其中 L_i^c 为类 c 的信息-Q 值损失,集合 N_i^+ 包含智能体的部分邻居智能体,它是从 Z_i 中使用 RWR 方法采样得到的.

具体地,从集合 Z_i 中选择一个抽样集合 U_i.对于集合 U_i 中的智能体 j,如果满足 $dist(i,j)\leqslant\delta$,则将智能体 j 及其类别添加到集合 N_i^+ 中,直到集合 N_i^+ 的数量达到批量大小.其中 $dist(i,j)$ 为两个智能体之间的距离,δ 是可调参数,可以根据不同实验场景设置.N_i^- 是 N_i^+ 的关于 Z_i 的补集.基于这个集合 N_i^-,利用 3.2 节设计的通信机制来产生负样本信息 $\widetilde{\boldsymbol{m}}_i^c$.由此,利用式(16)设计的互信息损失函数可以使互信息最大化.

3.4 全局优化目标

除了对通信模块中的信息的 MI 约束外,其他部分的所有参数都通过最小化 TD 损失进行更新.TD 损失目标函数如式(17)所示:

$$L_{\mathrm{TD}}=[r+\gamma\max_{a'}Q_{tot}(\tau',a';\theta^-)-Q_{tot}(\tau,a;\theta)]^2 \tag{17}$$

其中,θ^- 为目标网络的参数,θ 表示模型中的所有参数.

全局优化目标可以由式(18)表示:

$$L=L_{\mathrm{TD}}+\lambda L_{\mathrm{MI}} \tag{18}$$

其中 λ 表示可调的超参数,以权衡 TD 损失和所有智能体的总 MI 损失.

算法 1. HVDC 方法

输入:智能体 i 的局部观测 $o_i\in O_i$

输出:全局动作价值函数 Q_{tot}

1. 初始化网络参数,网络更新频率,经验回放池
2. FOR 每一回合
3. FOR 时间步 $t=1$ to n
4. FOR 每一智能体 $i\in N$ DO
5. 接收本地观测 o_i
6. 通过 MLP 和 GRU 获得特征 $\boldsymbol{h}_i$

7. 传输特征 $\boldsymbol{h}_i$ 到通信模块
8. 编码特征 $\boldsymbol{h}_i$ 为类特征($\boldsymbol{h}_i^1,\boldsymbol{h}_i^2,\cdots,\boldsymbol{h}_i^c$)
9. 发送($\boldsymbol{h}_i^1,\boldsymbol{h}_i^2,\cdots,\boldsymbol{h}_i^c$)到信息存储模块
10. 通过 RWR 构造邻居智能体集合 Z_i
11. 从 Z_i 中采样正样本邻居集合 N_i^+
12. 从信息存储模块获得 N_i^+ 中智能体的信息并通过 GNN 融合各类信息得到($\boldsymbol{m}_i^1,\boldsymbol{m}_i^2,\cdots,\boldsymbol{m}_i^c$)
13. 根据式(6)计算最终融合信息 $\boldsymbol{m}_i$
14. 基于信息 $\boldsymbol{m}_i$ 和观测历史 τ_i 得到局部价值 Q_i
15. 通过判别器 $\mathcal{D}_i$ 为“信息-Q 值对”($\boldsymbol{m}_i^c,Q_i$)评分
16. 根据 Z_i 和 N_i^+ 得到负样本邻居集 N_i^-
17. 从信息存储模块获得 N_i^- 中智能体的信息并通过 GNN 融合各类信息得到($\tilde{\boldsymbol{m}}_i^1,\tilde{\boldsymbol{m}}_i^2,\cdots,\tilde{\boldsymbol{m}}_i^c$)
18. 根据式(6)计算最终融合信息 $\tilde{\boldsymbol{m}}_i$
19. 通过判别器 $\mathcal{D}_i$ 为“信息-Q 值对”($\tilde{\boldsymbol{m}}_i^c,Q_i$)评分
20. 根据 $a_i=\pi(Q_i)$(ε-greed)选择动作
21. 存储观测历史 τ_i 和动作 a_i 至经验回放池
22. 输出局部价值函数 $Q_i(\tau_i,a_i,\boldsymbol{m}_i)$
23. 传输 $Q_i(\tau_i,a_i,\boldsymbol{m}_i)$ 到混合网络
24. 从经验回放池采样历史数据
25. 基于式(16)构造 MI 损失函数
26. 基于式(17)构造 QMIX 损失函数
27. 基于 QMIX 损失和 MI 损失最小化损失函数
28. 更新网络参数
29. 输出全局价值函数值 Q_{tot}
30. END FOR
31. END FOR
32. END FOR

算法 1 展示了HVDC方法的伪代码. 第 5～22 行描述了分散执行阶段. 在分散执行阶段,各智能体可以通过通信信道与其他智能体进行通信,并通过最大化互信息学习有效的信息. 智能体根据局部观测和信息表示以去中心化的方式选择动作. 可训练的异构图卷积网络包含一组对应每个智能体类的可学习的权值. 由于图卷积网络的信息传递特性,这些权值可以在分散执行阶段传递给各个智能体. 接下来,第 23～29 行描述了集中训练过程. 在集中训练过程中,假设 HVDC 可以从经验回放池接收各个局部观测-动作历史. 混合网络与 QMIX 使用的网络相同,采用多个局部价值作为输入进行单调性混合并生成全局价值 Q_{tot}. 此外,利用 QMIX 中定义的损失函数和互信息损失共同来训练网络和更新网络参数.

4 实　验

本文在不同环境中进行实验,并与几个基线方法进行比较,以此评估 HVDC 的有效性. 首先,为验证算法的收敛性,在相对简单的捕食者-捕获者-猎物(Predator-Capture-Prey)异构环境上进行实验. 接下来,在复杂的星际争霸平台(The StarCraft Multi-Agent Challenge,SMAC)上的一系列异构场景下评估了 HVDC 的性能. 此外,在真实交通信号控制环境(Traffic Signal Control Environment)上对 HVDC 方法进行评估,以证明方法在现实场景中应用的可行性. 所有实验都在 GPU Nvidia RTX 2080 上使用 Pytorch 框架构建.

4.1 捕食者-捕获者-猎物环境

如图 3 所示,捕食者-捕获者-猎物环境包含两种类型的友方智能体:捕食者和捕获者,以及一种敌方智能体:猎物. 捕食者的目标是找到猎物. 在每一个时间步中,所有捕食者的状态空间是一个联合特征向量,它包括捕食者的位置和其他智能体存在的信息. 捕食者能够观测环境,其观测是由其视野范围内所有状态向量连接起来的数组. 捕食者的动作空间维度为 5,包括动作下移、上移、左移、右移和停留.

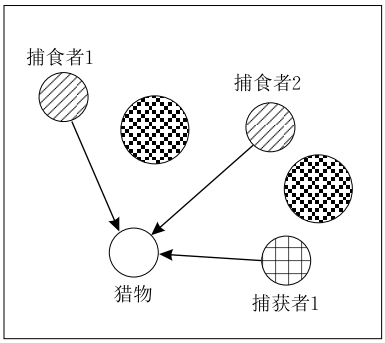


图 3　捕食者-捕获者-猎物环境示意图

第二种类型的友方智能体:捕获者,目标是定位和捕获猎物. 捕获者和捕食者在动作空间和观测空间上的区别在于捕获者不能从环境中获得任何观测. 此外,捕获者在它们的动作空间中有一个附加的动作,捕获:当它们移动到猎物的位置时,它们采取捕获动作来捕获位于相应位置的猎物. 因此,合作智能体的目标是,所有的捕食者找到固定的猎物,所有的捕获者到达该位置,然后捕获猎物. 每个合作智能体在每个时间步会受到－0.05 的惩罚,直到完成它的任务.

在这个环境中,将提出的方法 HVDC 与几个最先进的学习基线进行比较. 基线包括两种价值函数分解方法 QMIX[5] 和 NDQ[14],三种通信学习方法 CommNet[8]、TarMAC[9] 和 IC3Net[12]. QMIX 是一种代表性的价值函数分解方法,在多智能体任务中

表现出优异的性能. QMIX 利用混合网络表示联合动作的最优价值函数,提高了表示能力. CommNet 是 MARL 中最早的通信学习方法之一,展示了学习连续通信向量的能力. TarMAC 利用引入注意力机制的定向通信架构来实现智能体之间的通信. IC3Net 可以通过门控机制学习何时与其他智能体进行通信. NDQ 利用最小化通信机制来学习价值函数分解,智能体偶尔向其他智能体发送信息,以有效地协调动作.

实验使用包括多头注意力的异构图卷积网络构建通信学习模块. 使用三层图卷积网络,前两个图卷积层每层使用 4 个注意力头,每个注意力头计算 16 个特征. 最终 64 个特征通过拼接操作合并. 最后一层图卷积网络也使用 4 个注意力头,但将输出维度设置为与每个智能体的动作空间相同的大小,相应的特征通过平均操作合并. 由于场景相对简单,没有使用互信息优化. 表 1 给出了在异构捕食者-捕获者-猎物上不同方法的性能比较,包括学习策略的平均奖励和平均步数. 这些值表示使用三种不同随机种子初始化试验的平均值(±标准差). 表中每一列的最高性能结果以粗体显示. 如表 1 所示,HVDC 的性能优于其他基线方法.

表 1 不同方法在 Predator-Capture-Prey 中的性能比较

方法	平均奖励值	平均时间步
CommNet	−0.40±0.01	11.03±0.08
IC3Net	−0.40±0.02	11.23±0.63
QMIX	−0.42±0.03	11.36±0.17
NDQ	−0.39±0.01	10.58±0.24
TarMAC	−0.48±0.03	16.01±0.80
HVDC	**−0.35±0.02**	**9.36±0.58**

上述实验使用的捕食者数量为 3,捕获者数量为 1,猎物的数量为 1,智能体数量总数为 5. 为了验证 HVDC 可以扩展到更多的智能体场景,比较不同智能体数量的场景中 HVDC 和基线方法的性能. 所有的设置都是和之前的场景一样,除了各种智能体的数量按比例增加. 如表 2 所示,随着智能体数量的增加,HVDC 的性能也总是最佳的. 结果表明,该方法可扩展到大规模场景.

表 2 不同数量智能体的泛化性测试实验

	CommNet	IC3Net	QMIX	TarMAC	HVDC
5	−0.40±0.01	−0.40±0.02	−0.42±0.03	−0.48±0.03	**−0.35±0.02**
10	−0.36±0.03	−0.37±0.03	−0.39±0.02	−0.45±0.04	**−0.32±0.01**
15	−0.35±0.02	−0.38±0.03	−0.41±0.03	−0.43±0.03	**−0.30±0.01**
20	−0.38±0.03	−0.35±0.02	−0.42±0.03	−0.46±0.04	**−0.27±0.02**
40	−0.34±0.02	−0.32±0.01	−0.37±0.02	−0.42±0.03	**−0.28±0.01**

4.2 星际争霸平台

为进一步验证提出方法的有效性,在更复杂的 SMAC[23] 平台上对 HVDC 的性能进行了进一步评估. 在 SMAC 中,智能体的动作空间维度是 4,包括移动、无操作、攻击和停止. 在这些动作的控制下,智能体在这些连续空间的地图中攻击和移动. 在每个时间步中,智能体将获得与敌方单位受到的累积伤害相等的奖励. 智能体每杀死一个敌方单位可以获得额外奖励 10,每赢得一场战斗获得奖励 200.

本文调高了动作协调的难度,一方面,将智能体的视野范围从 9 缩小到 6,另一方面,选择具有挑战性和复杂地形的地图,如图 4 所示. 实验使用了多种不同的异构智能体场景,其中友方单位都由强化学习智能体控制,敌方单位由内置的 AI 控制. 友方单位和地方单位可以是不对称的,它们的初始位置是随机的. 本文使用在 9 个异构场景上评估所提出的方法,下面对其中主要场景进行简要描述.

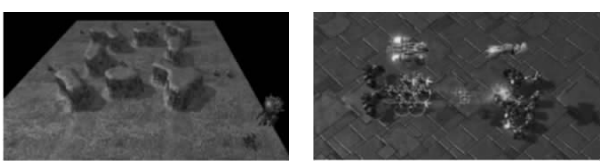


图 4 星际争霸平台异构场景示意图

1o10b_vs_1r 场景的友方单位包括一个监察者(Overseer)和 10 个毒爆虫(Baneling). 友方队伍需要杀死敌方蟑螂(Roach)才能赢得比赛. 在通信模块中,监察者需要将它的位置信息发送给毒爆虫. MMM 场景的友方单位包括 7 个海军陆战队员(Marine),2 个掠夺者(Marauder)和 1 个医疗兵(Medic). 为赢得战斗,智能体必须学会向友方医疗兵发送它们的生命值等信息. MMM2、MMM3 场景和 MMM 场景类似,拥有同样的智能体类型,区别是敌方单位多了一个或两个海军陆战队员. 1o2r_vs_4r 场景友方队伍包括 1 个监查者(Overseer)和 2 个蟑螂(Roach). 监查者的目标是找到敌方队伍的 4 个收割者(Reaper). 2 个蟑螂的目标是到达收割者的位置,并试图击败收割者. 考虑到只有监察者才能获得敌

人的位置信息，HVDC 必须学会将监察者的信息传递给蟑螂，以便有效地赢得战斗. 1c3s5z、2c3s5z 以及 3c5s7z 场景中友方单位和敌方单位都包括巨像(Colossus)、追猎者(Stalkers)和狂热者(Zealots)，区别是智能体的数量不同. HVDC 必须学会很多策略和技巧，比如用友方的狂热者拦截敌方狂热者，从而保护友方的追猎者免受严重的破坏.

实验基于 PyMARL 框架[23]，并利用其预设的网络结构和 QMIX 模块的超参数设置. 在 SMAC 实验环境中，HVDC 使用了互信息优化. 通信模块的异构图卷积网络部分的设置和捕食者-捕获者-猎物环境中的设置相同. 互信息优化模块超参数设置如下：p 设置为 0.6，Z_i、N_i^+ 和 N_i^- 可根据不同场景进行调整，δ 设置为 5. 图 5 展示了 HVDC 和基线的性能比较，使用的基线包括 QMIX[5]、QMIX+TarMAC、NDQ[14] 以及两个基于异构图卷积网络的方法 HMAGQ-Net[18] 和 HetNet[19]. QMIX+TarMAC：将 TarMAC 的通信模块融合到 QMIX 中，将注意力通信机制应用到价值函数分解框架. 其他基线方法在前文中已经介绍. 表 3 展示了所有场景的胜率比较，其中基线方法选择了 QMIX、NDQ 和 HetNet.

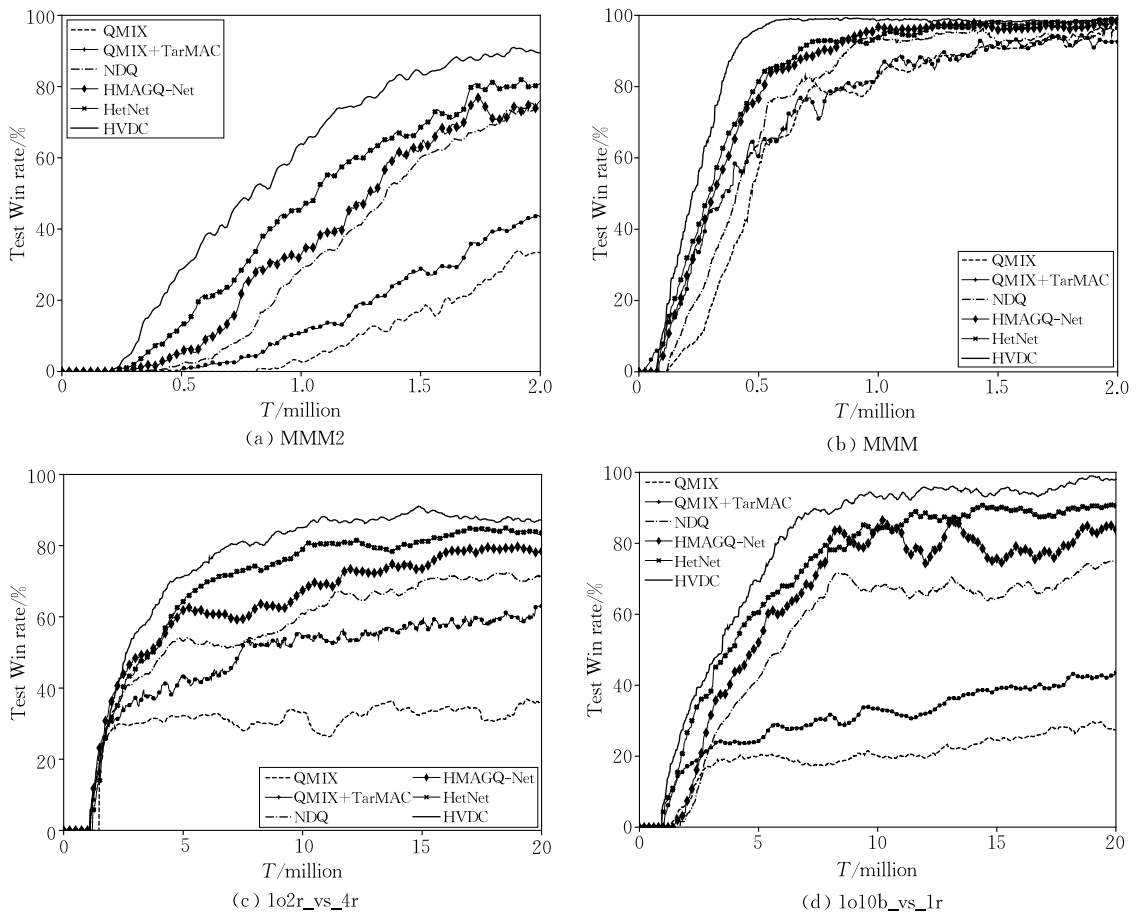


图 5　SMAC 下学习曲线对比

表 3　不同方法在异构场景中的胜率比较

场景	QMIX	NDQ	HetNet	HVDC
MMM	94.16±1.37	97.21±1.43	98.15±0.89	**98.81±0.93**
MMM2	33.06±8.92	74.39±3.93	81.36±4.03	**90.12±2.43**
MMM3	46.32±7.64	55.32±6.14	62.25±5.27	**85.25±2.93**
MMM4	42.13±8.14	46.74±7.23	58.63±5.35	**81.49±3.23**
1o2r_vs_4r	36.37±9.10	71.32±4.93	84.37±3.26	**86.83±2.27**
1o10b_vs_1r	28.31±9.27	75.41±3.76	90.73±1.38	**97.89±1.02**
1c3s5z	87.24±1.52	95.03±0.88	94.27±0.75	**97.03±0.88**
2c3s5z	60.85±6.17	76.38±4.19	80.21±4.15	**93.15±1.37**
3c5s7z	56.12±6.84	67.93±5.82	79.04±3.81	**90.21±2.13**

如图 5 和表 3 所示，在多个异构场景中 HVDC 方法的性能优于 QMIX 方法，说明 QMIX 以及其他分解方法的动作失调问题普遍存在，特别是在具有高随机性的地图中，如 1o10b_vs_1r 和 1o2r_vs_4r. 值得注意的是，HVDC 的性能大幅度超过引入注意力通信机制的 QMIX＋TarMAC. 由于智能体在 HVDC 和 QMIX＋TarMAC 两种方法中都使用通信学习机制且使用相同的 QMIX 模块，因此 HVDC 性能的优越性表明 HVDC 的异构通信学习机制相比 QMIX＋TarMAC 的通信机制更有效. 同时，该方法的性能也优于 NDQ，特别是在复杂场景中，如 1o10b_vs_1r 和 1o2r_vs_4r.

为评估 HVDC 中每个部分的效果，设计消融实验来验证 HVDC 性能的提升是否来源于所提出的异构通信学习模块、互信息优化和价值分解函数模块. 设计了 HVDC 的三个变体：(1) HVDC-H 是没有异构通信学习模块的 HVDC，直接使用同构图卷积网络进行通信学习；(2) HVDC-V 是没有价值分解模块的 HVDC；(3) HVDC-I 是没有进行互信息优化的 HVDC. 表 4 展示了 HVDC 在不同异构场景上的消融实验. 与 HVDC 相比，这三个变体的性能都有所下降，说明了每个模块的有效性. 这些实验结果表明，通信学习模块可以加强智能体之间的动作协调，互信息优化可以提高通信学习的质量，价值函数分解模块可以进一步促进智能体的策略学习.

表 4 HVDC 在不同异构场景上的消融实验

场景	QMIX-L	HVDC-V	HVDC-H	HVDC
MMM	97.25±1.34	98.25±1.27	94.56±2.30	**98.81±0.93**
MMM2	62.69±3.81	79.17±3.16	47.62±5.27	**90.12±2.43**
MMM3	52.74±2.43	82.27±2.54	50.14±4.76	**85.25±2.93**
MMM4	49.38±5.04	78.02±2.85	42.46±5.93	**81.49±3.23**
1o2r	65.29±3.47	83.78±2.02	46.13±5.51	**86.83±2.27**
1o10b	75.87±47.51	94.50±1.38	62.87±3.74	**97.89±1.02**
1c3s5z	90.36±1.73	93.71±1.74	89.43±1.63	**97.03±0.88**
2c3s5z	79.43±3.32	91.30±1.23	77.32±2.98	**93.15±1.37**
3c5s7z	64.87±4.02	87.56±1.80	68.10±3.92	**90.21±2.13**

HVDC 相比 QMIX 使用了通信学习模块，进而在决策时有了更多的共享信息. 于是我们设计了与 HVDC 相似的参数数量的 QMIX-LARGE，研究 HVDC 优于的 QMIX 是否源于参数数量的增加. 在 MMM2 场景上进行实验. 如图 6 所示，结果表明，具有更复杂网络的 QMIX 并不能从根本上提高性能.

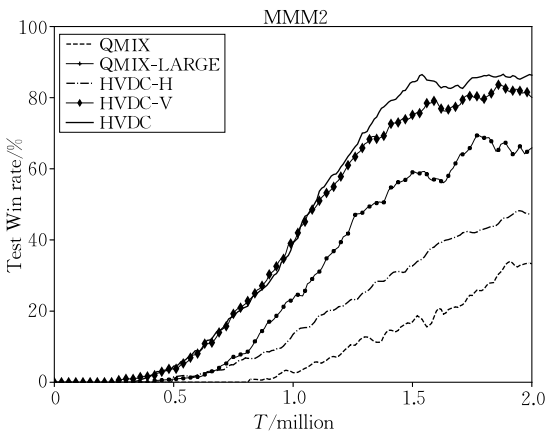


图 6 HVDC 在 MMM2 场景上的消融实验

HVDC 可以与各种价值函数分解方法集成，本文将其与目前主流的价值分解方法 VDN、QMIX 和 QPLEX 结合来进行消融实验. 集成方法分别被称为 HVDC(VDN)、HVDC(QMIX)和 HVDC(QPLEX)，在场景 3c5s7z 上进行实验. 如图 7 所示，所有集成方法的性能都优于原有的价值函数分解方法，说明 HVDC 的异构通信学习模块能够显著增强智能体的动作协调.

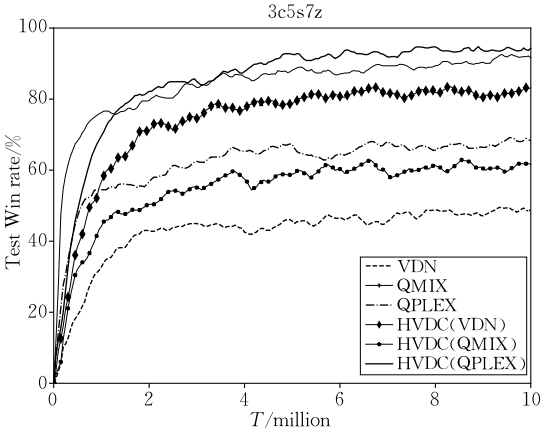


图 7 HVDC 在 3c5s7z 场景上的消融实验

为直观地验证和说明所提出的方法可以轻松地扩展到包含更多数量智能体的场景，我们比较了 HVDC 和基线方法在场景 1c3s5z、2c3s5z 和 3c5s7z 中的性能，三个场景中智能体的数量分别为 18、20

以及 30. 总的来说，随着智能体数量的增加，不同类型智能体之间的交互和关系变得越来越复杂，导致策略学习变得越来越困难. 如图 8 所示，随着智能体数量的增加，智能体学习策略的难度增加，基线方法的胜率出现明显的下降趋势. 尽管如此，HVDC 算法仍然保持很高的胜率. 随着智能体数量增加，HVDC 相比其他基线方法的胜率优势更加显著.

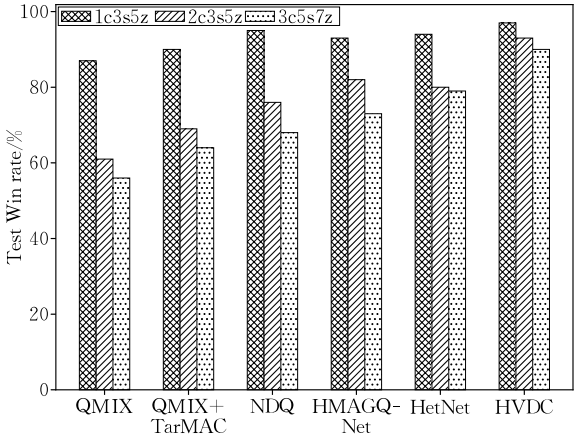


图 8　HVDC 在不同数量智能体场景上的消融实验

值得注意的是，与价值分解方法如 QMIX 相比，HVDC 的网络结构多了一个异构通信学习模块，这使得 HVDC 相比价值函数分解方法，需要更多的计算时间. 在相同的场景和参数设置下，比较 HVDC(QMIX)和 QMIX 在几个场景下所需的训练时间. 为便于比较，基于 HVDC(QMIX)对训练时间进行了归一化. 如图 9 所示，随着场景复杂性的增加，HVDC 比 QMIX 需要更多的训练时间. 然而，即使在最复杂和困难的场景 MMM-4 中，HVDC(QMIX)只需要比 QMIX 多大约 10%的训练时间. 与胜率的显著提高(40%到 82%)相比，这种程度的训练时间的增加是可以接受的.

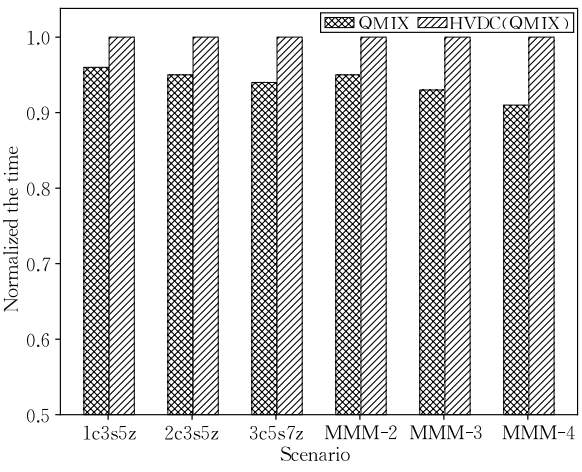


图 9　HVDC 和 QMIX 在不同场景上的训练时间对比

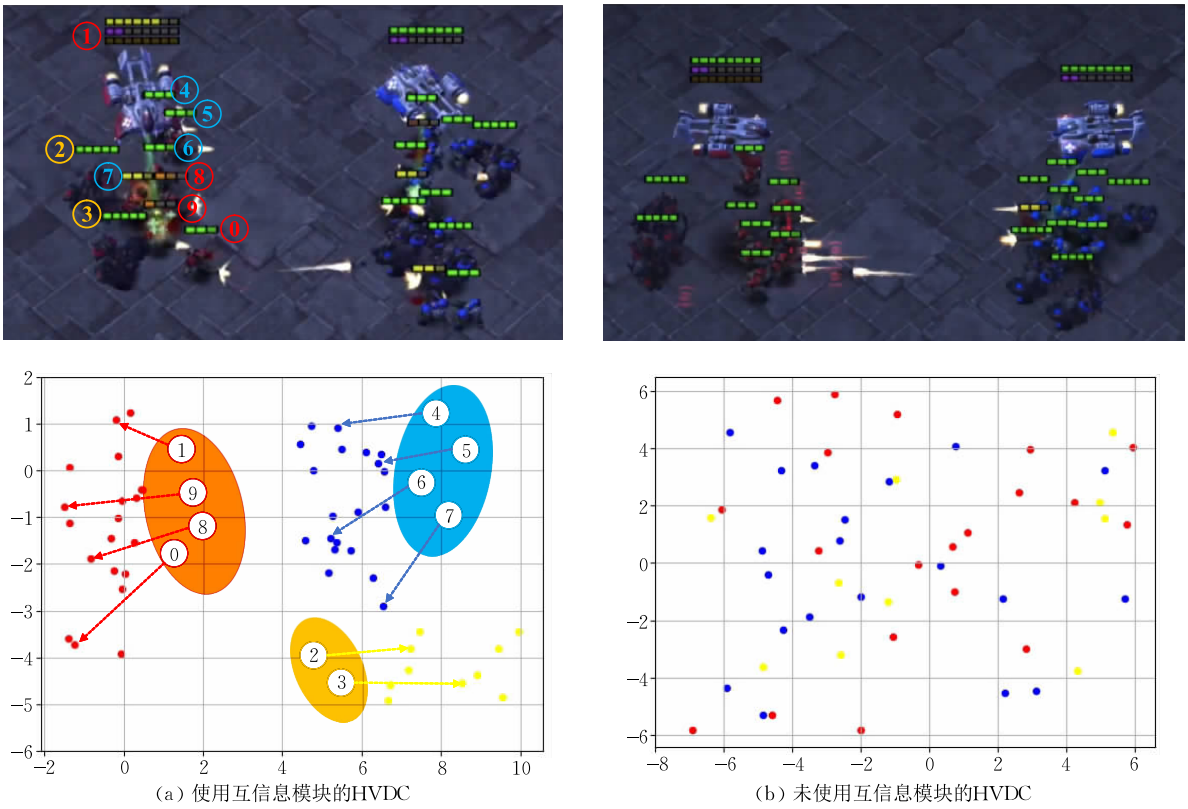
为进一步验证互信息优化模块对动作选择的影响，进行消融实验. 如上文提到的，HVDC-I 表示没有进行互信息优化的 HVDC. 对 HVDC 和 HVDC-I 在 SMAC 的 MMM2 场景下不同时间步的动作选择进行展示和分析，并将对应时间步的多个智能体学得的信息表示映射到表示空间，进行相应的可视化分析. 具体地，首先选择两个不同的时间步，获得两个时间步的环境状态. 然后，收集两个时间步的所有存活的友方智能体的信息表示. 由于实验是使用 5 个随机种子进行的，可以在一个特定的时间步得到每个存活的智能体的 5 个信息表示(为了方便，每个特定时间步的五次实验，只采用一个环境状态). 死亡的智能体不能接收信息没有信息表示. 使用 t-SNE[24] 降低了每个信息表示的维数，因此可以在二维平面上表示.

如图 10 所示，在第 10 时间步，使用互信息模块的 HVDC 训练的智能体之间已经出现了局部的动作协作，距离敌方智能体比较近的 8 号、9 号、0 号海军陆战队员形成了局部的协作，它们协调动作集中火力，共同攻击某个敌方海军陆战队员智能体，而 1 号医疗兵率先对这三个友方海军陆战队员进行治疗，而不去关注其他的友方智能体. 根据信息表示在表示空间中的分布位置，可以对智能体进行分组.

在第 10 时间步，可以看到在游戏中形成的智能体组与在信息表示空间中形成的智能体组之间的对应关系. 同一组中的智能体往往会收到相似的信息表示，并完成更多的动作协作. 与之相反，在第 10 时间步，未使用互信息模块的 HVDC 训练的智能体之间未出现明显的动作协作，同时，对应的信息表示在表示空间中几乎是随机分布的.

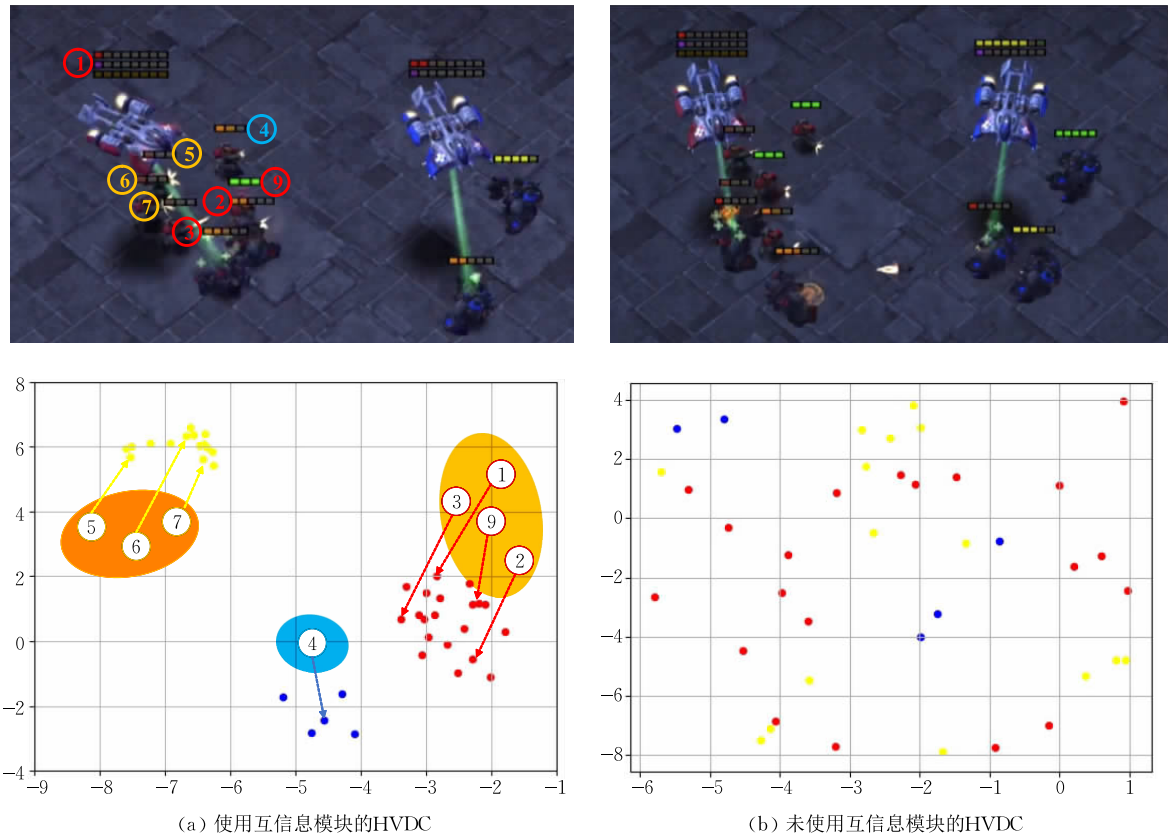
在第 30 时间步，如图 11 所示，使用互信息模块的 HVDC 训练的智能体之间相比 HVDC-I 训练的智能体，表现出更合理的局部的动作协作. HVDC 训练下，1 号医疗兵选择对距离敌方更近的友方智能体 2 号、3 号、9 号进行治疗，而在 HVDC-I 训练下，1 号医疗兵选择距离它最近的 6 号智能体进行治疗. 显然，6 号智能体目前不存在死亡的危险，所以使用互信息模块的 HVDC 训练下的智能体的动作选择更加合理. 同时，在信息表示的分布方面，展现了和第 10 时间步基本一致的分布规律.

表 5 总结了 4.1 节和 4.2 节场景中使用的所有基线方法和本文提出方法的各种变体的特征比较，比较维度主要包括是否使用通信模块，是否使用 GNN，是否包括价值函数分解模块，以及是否处理异构特征.



(a) 使用互信息模块的HVDC (b) 未使用互信息模块的HVDC

图 10 第 10 时间步的环境状态和信息表示的可视化



(a) 使用互信息模块的HVDC (b) 未使用互信息模块的HVDC

图 11 第 30 时间步的环境状态和信息表示的可视化

表 5　不同基线方法的特征比较

方法	通信模块	GNN	价值分解	异构特征
CommNet	是	是	否	否
IC3Net	是	是	否	否
QMIX	否	否	是	否
NDQ	是	否	是	否
TarMAC	是	是	否	否
HMAGQ-Net	是	是	否	是
HetNet	是	是	否	是
HVDC	是	是	是	是
HVDC-H	是	是	是	否
HVDC-V	是	是	否	是
HVDC-I	是	是	是	是

4.3　交通信号控制

为验证 HVDC 在真实系统中的可行性和有效性，在交通信号控制环境下对 HVDC 进行评估，并在真实交通网络上进行了实验[25]. 如图 12 所示，中间矩形表示的 SQ2 由主路和连接主路的 10 条支路和 20 个路口组成. 利用城市交通仿真(Simulation of Urban Mobility，SUMO)平台[26]来模拟真实的交通状况. 交通网络可以定义为异构图 $G=(V,E,O_V,R_E)$，V 和 E 代表节点和边的集合，O_V 和 R_E 代表节点类型和边的类型的集合. 交通网络中的对象如交通信号控制器和车辆被建模为节点，对象之间的关系被建模为边. 交通信号网络包括四类节点：交通信号控制器(Traffic-Signal Controller，TSC)、车辆、车道和路口. 节点的类别和异构特征如表 6 所示.

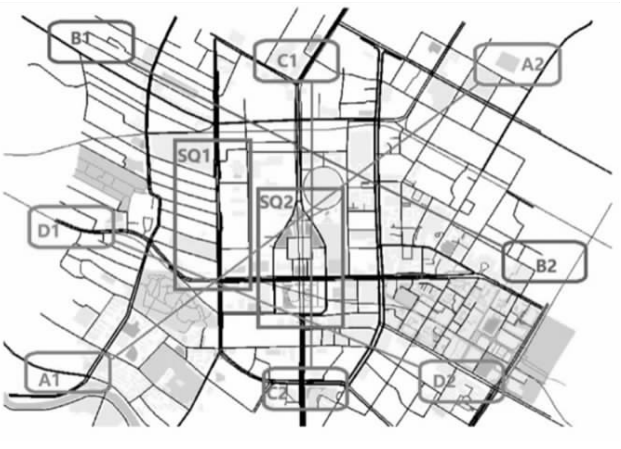


图 12　交通信号控制真实地图示意图[24]

表 6　交通网络中节点的异构特征

节点类型	异构特征
TSC	$A_i=\{(EWG,d_1^t),(EWLG,d_2^t),(NSG,d_3^t),(NSLG,d_4^t)\}$
路口	$\{IS_{EWG},IS_{EWLG},IS_{NSG},IS_{NSLG}\}$
车辆	等待时间、速度、位置
车道	队列长度、车道长度

将每个交通信号控制器建模为一个 TSC 智能体，将网络交通状态建模为全局状态，该状态由 TSC 智能体、路口、车道和车辆的异构特征组成. 例如，IS_{NSG} 表示与 TSC 智能体特征 (NSG,d_3^t) 同步的路口的可通过性，队列长度表示在路口等待的车辆数量，等待时间表示第一辆车辆在路口的累计等待时间. 采用 SUMO 平台的检测器来捕获特征信息.

具体地，交通信号控制任务用部分可观测的马尔可夫决策 $G=\langle C,I,S,A,R,P,\Omega,O,\gamma\rangle$ 建模.

(1) 全局状态和局部观测

I 是 TSC 智能体的集合，其中在每个时间步 t，智能体 i 观测到全局状态 s 的一部分作为其局部观测 o_i^t，局部观测包括交通场景中的路口、车道和车辆的异构特征，如车辆的数量. 路口节点的异构特征对应 TSC 节点的异构特征同步，车道和车辆节点的异构特征受到 TSC 智能体采取的动作影响.

(2) 动作空间

TSC 的动作空间与表 6 中所示特征相同：$A_i=\{(EWG,d_1^t),(EWLG,d_2^t),(NSG,d_3^t),(NSLG,d_4^t)\}$，其中 $d_c^t(c=1,2,3,4)$ 表示信号灯的持续时间，$a_i^t=(EWG,d_1^t)$ 表示 TSC 智能体 i 打开南北方向左转灯 d_1^t 秒，这些动作根据特定顺序循环改变，信号灯持续时间的范围是 $d_c^t[s]\in[10,50]$.

(3) 奖励函数

车道中车辆的等待时间和队列长度经常被 TSC 智能体的动作(即交通信号)改变，于是 TSC 智能体 i 的奖励函数由式(19)表示：

$$r_i^t=-\sum_{j\in\mathcal{N}_i}(\eta\times wait_{i,j}^{t+d_c^t}+queue_{i,j}^{t+d_c^t}) \tag{19}$$

其中 $\mathcal{N}_i$ 表示智能体 i 的邻居智能体集合，$wait_{i,j}^{t+d_c^t}$[s] 表示车辆在交叉路口的累计延迟. $queue_{i,j}^{t+d_c^t}$[veh] 表示在时间 $t+d_c^t$ 从路口 j 出发到路口 i 的每条车道上车辆的队列长度，η[veh/s] 表示权衡因子.

每个交通信号控制器智能体 i 选择动作，以最大化累积折扣奖励 $R_i=\sum_{t=\tau}^{T}\gamma^{t-\tau}r_i^t$. 其中 $\gamma\in[0,1]$ 为折现因子，T 为时间范围. 学习率设置为 0.001，批量大小设置为 200. 此外，η，γ 和 T 分别设置为 0.7，0.99 和 720.

(4) 问题定义

给定具有交通网络中节点异构特征集的异构图，首先对 TSC 节点 i 及其邻域节点的异构特征进行编码学习得到特征表示，然后通过提出的框架完成 Q 值的学习和全局目标的优化.

其中互信息最优化模块的超参数设置如表 7 所示：返回概率 p 设为 0.7，TSC 智能体 i 的采样邻居

节点数 Z_i 设为 76(包括 4 个 TSC 智能体、4 个路口、8 个车道和 60 辆车),正样本 N_i^+ 个数设为 40(包括 2 个 TSC 智能体、2 个路口、6 个车道和 30 辆车). 负样本集 N_i^- 由 $Z_i - N_i^+$ 得到,δ 根据场景设置. HVDC 其他模块的超参数设置与 SMAC 环境相同.

表 7 互信息优化模块的超参数设置

参数名	参数符号	参数值
返回概率	p	0.7
邻居节点数	Z_i	76
正样本个数	N_i^+	40
负样本个数	N_i^-	36
距离上界	δ	根据场景调整

在真实的交通信号控制场景中,为了展示 HVDC 的效率和有效性,选择最大压力控制(Max pressure control)[27]、Metalight[28]、Colight[29] 和 IG-RL(Inductive Graph Reinforcement Learning)[30]、IHG-MA(Inductive Heterogeneous Graph Multi-agent Actor-critic)[24] 等几种基线算法. 最大压力控制是交通信号控制领域最先进的方法之一. 通过选择信号相位,使进入路口的车辆数量最大化. Metalight 是 MARL 方法中的一种,采用周期性交替的个体适应和全局适应来处理交通信号控制任务. Metalight 利用从现有交通网络中学习到的知识,在新的交通网络中加快学习过程. Colight 是一种分散的 MARL 方法,利用图注意力网络促进多个智能体之间的信息传递,以实现协作完成交通信号控制任务. IG-RL 是另一种基于同构图神经网络的分散 MDRL 方法,利用学习到的可迁移策略来处理多路口信号控制. IHG-MA 是一种基于异构图神经网络的 MDRL 方法,使用多智能体行动者-评论家框架学习策略.

评价结果以各路口的拥堵状况和车辆通行效率为主要评价对象,包括三个主要指标:平均延迟、平均队列长度和平均行驶时间. 平均延迟表示交通路口的所有车辆的平均等待时间除以平均队列长度. 它的值越高,表示该方法的表现就越差. 平均行驶时间表示交通网络中的所有车辆从开始行驶到结束所花费的平均时间. 平均行驶时间越长,方法的表现就越差.

首先,随机选取采用 SUMO 中的合成道路数据集,合成交通网络对 HVDC 进行模拟训练. 训练过程包括 10 个并行模拟,每次模拟进行 1 百万时间步,基线也使用相同的实验设置进行训练. 本文设计一个合成交通网络集,其中包含 20 种类型的交通网络. 典型的路口状态如丁字路口、十字路口存在于上述网络中,且连接的每个方向有 3 个车道(长 500 m,宽 4 m).

然后,利用不同的合成交通网络对这些方法进行训练,以获取有效的信号控制策略. 交通网络中车辆的起点和终点被均匀地随机设置在任意车道. 利用 TSC 动作空间中采用不同的动作组合来控制路口的交通运动. 此外,单向和双向交通模式分别在合成交通网络集中实现. 具体地,将交通流设置为东西方向 520 辆/车道/h,南北方向 210 辆/车道/h. 其他实验参数见 4.4 节表 13.

由于本文采用的该场景下的基线方法所定义的奖励函数有所不同,为了统一,将平均延迟作为评估指标,采用它的负数值作为奖励值. 表 8 展示了不同方法的训练结果,即平均奖励值,结果显示 HVDC 算法显著优于所有的其他基线方法,当达到最终收敛状态时,HVDC 算法的奖励值大约为 −98.59.

表 8 不同方法在实际交通网络中的结果比较

方法	平均奖励值
Max Pressure	−1104.42±20.17
Metalight	−735.40±15.62
Colight	−596.73±12.83
IG-RL	−591.56±13.64
IHG-MA	−220.62±10.58
HVDC	**−98.59±8.17**

然后,对训练方法在真实交通网络 SQ2 中进行 1h 真实交通流测试. SQ2 中测试集有 $A_1 \leftrightarrow A_2$、$B_1 \leftrightarrow B_2$、$C_1 \leftrightarrow C_2$、$D_1 \leftrightarrow D_2$ 四种交通流. 时变交通流设置为 $[1,2,4,5,4,2,1] \times 330$ 辆/车道/h,每个交通流的间隔为 400 s. $A_1 \leftrightarrow A_2$ 和 $B_1 \leftrightarrow B_2$ 的开始时间为 0,$C_1 \leftrightarrow C_2$ 和 $D_1 \leftrightarrow D_2$ 的开始时间为 600 s,其他实验参数见 4.4 节表 10. 有些车辆可能需要很长时间才能完成计划的路线,为控制计算成本,在 1h 后停止评估. 为了获得稳健的结果,测试和训练过程都使用不同的随机种子重复实验 5 次,此外,每个过程包括 5 个并行模拟,即总共有 25 个模拟,最后的结果为 25 次模拟结果的平均值.

图 13 展示了各方法的平均延迟曲线,如图 13 所示,HVDC 方法是所有方法中性能最好的. 具体地,HVDC 在 2900 s 处达到约为 17 的最大值;IHG-MA 的最大值约 32,在 2950 s 处达到;IG-RL 在 3000 s 处达到最大值,约为 53;Metalight 的最大值约 120,在 3100 s 达到. 所有的平均延迟曲线都在早期稳定上升,最大压力控制和 Metalight 在后期仍然上升,而 Colight、IG-RL、IHG-MA 和 HVDC 都在不同的时间达到峰值,并在后期趋于下降.

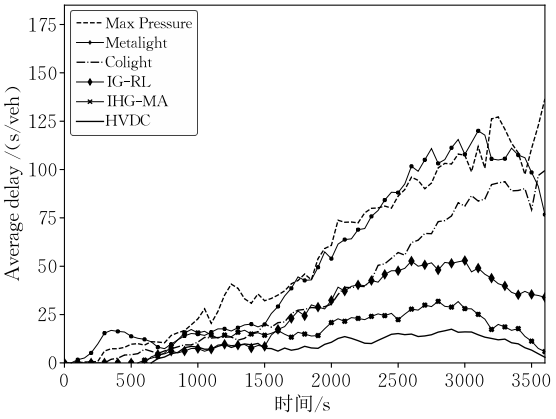


图 13　不同方法下的平均延迟对比

实验结果表明，最大压力控制和 Metalight 都不能依靠可持续的策略来快速恢复拥堵的交通网络. IG-RL 得益于同构 GNN 结构，学习的策略优于 Colight. HVDC 与 IG-RL 相比，学得了更稳定、更可持续的策略，实现了更低的路口拥堵水平和更快速的交通恢复，这得益于异构通信机制. IHG-MA 由于采用了异构 GNN 结构去建模交通网络，取得了比其他基线方法更好的效果. 由于 HVDC 相比 IHG-MA 还使用互信息优化以及价值函数分解，实现了更有效的通信学习和更高质量的策略学习，所以取得了相比 IHG-MA 更好的效果. HVDC 方法下的后期的平均延迟趋于零，说明 HVDC 方法对缓解路口拥堵，提高车辆行驶效率发挥了重要作用.

图 14 展示了不同方法下的平均队列长度对比，可以发现，不同方法的平均延迟曲线和平均队列长度曲线的变化趋势存在相似性. 它们大多在初期增加，在不同时间达到峰值，在后期趋于下降. 事实上，这两个指标是相关的. 如果平均队列长度增大，平均延迟也可能会相应增加. 同样值得注意的是，基线方法和 HVDC 都在学习了历史交通经验数据后，不同程度地减少了平均队列长度. 表 9 给出了在真实交通网络中不同方法的性能比较. 可以发现，HVDC 的性能优于其他基线，这说明通过异构通信学习机制、互信息优化和价值函数分解方法，HVDC 可以更高效、有效地处理真实的异构场景.

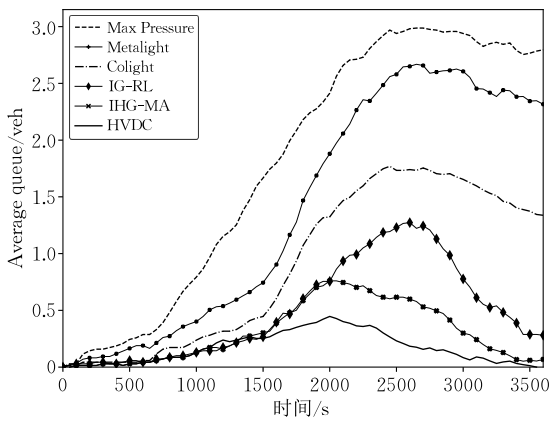


图 14　不同方法下的平均队列长度对比

表 9　不同方法在实际交通网络中的结果比较

方法	平均延迟/(s/veh)	平均队列/veh	平均行驶时间/s
Max Pressure	136.51±12.27	2.99±0.71	395.63±13.47
Metalight	119.79±10.64	2.67±0.65	342.57±11.82
Colight	99.47±8.17	1.74±0.49	369.21±12.16
IG-RL	52.71±5.40	1.27±0.35	312.17±9.27
IHG-MA	31.76±3.18	0.75±0.21	295.32±8.63
HVDC	**17.42±2.03**	**0.52±0.13**	**262.93±6.07**

4.4　实验参数

这一节我们给出实验相关参数，表 10 描述了所有实验场景中固定的参数. 表 11～表 13 给出不同实验场景的参数.

表 10　各实验环境固定的实验参数

超参数	值
GNN 层数	3
MLP 层数	2
GNN 激活函数	PReLU
判别器激活函数	Sigmod
折扣因子	0.99
学习率	1e−5
RNN 类型	GRU
优化函数	Adam
混合网络	QMIX
注意力头个数	4

表 11　捕食者-捕获者-猎物环境的实验参数

超参数	值
互信息优化	无
批量大小	50
训练步数	1e+5
随机种子	3
视野范围	1
友方智能体个数	4～32
敌方智能体个数	1～8
动作空间维度	[5,6,5]

表 12　星际争霸环境的实验参数

超参数	值
互信息优化	有
批量大小	100
训练步数	5e+6～2e+7
随机种子	5
视野范围	5
采样概率	0.6
友方智能体个数	3～15
敌方智能体个数	4～15
动作空间维度	4

表 13 交通信号控制环境的实验参数

超参数	值
互信息优化	有
批量大小	200
训练步数	1e+6
随机种子	5
视野范围	2
采样概率	0.7
邻居智能体个数	76
正样本邻居个数	40
动作空间维度	4
黄灯持续时间	3.3 s
最大车辆速度	50 km/h
车辆长度	4 m
最大加速度	1.2 m/s^2

表 14 展示参数 λ 的敏感度分析，如表 14 所示，λ 取值为 0.1 时，HVDC 在大部分异构场景上取得最高的胜率，所以在使用互信息优化的 SMAC 和交通信号控制场景中，统一将参数 λ 设置为 0.1.

表 14 HVDC 在不同异构场景上使用不同 λ 的胜率

场景	0.05	0.10	0.15
MMM	97.13±1.25	**98.81±0.93**	98.42±0.97
MMM3	81.14±3.52	85.25±2.93	**86.18±2.37**
1o2r_vs_4r	80.93±3.74	**86.83±2.27**	84.52±2.70
1o10b_vs_1r	92.69±1.38	**97.89±1.02**	97.04±1.13
1c3s5z	90.16±1.87	**97.03±0.88**	96.13±1.28
3c5s7z	82.95±4.02	**90.21±2.13**	89.36±2.58

5 结 论

本文提出了一种新的多智能体强化学习框架，该框架融合了价值函数分解和通信学习机制和互信息优化，解决了异构智能体之间的通信问题和价值分解方法的动作不协调问题. 通过引入异构图卷积网络来建模异构智能体间的信息交换，促进智能体之间的动作协调，并结合价值函数分解方法实现最优策略学习. 在各种任务中的实验结果表明，所提方法在异构环境中都明显优于基线方法. 通过在交通信号系统上的实验，验证了该方法在真实异构场景中应用的可行性.

本文所提出的框架为异构多智能体强化学习领域提供了一种新思路，据我们所知，本文的工作是第一次尝试结合基于图卷积网络的通信学习方法、互信息优化和价值函数学习方法解决异构多智能体问题. 通过结合这三类方法的优势，该框架同时解决了可扩展性问题、通信的有效性问题以及智能体动作协调等问题.

此外，我们计划在未来将提出的框架应用到更多的真实异构多智能体场景. 进一步实现异构智能体子任务的划分，以及在更复杂的场景中建模异构智能体之间的相互作用和关系，这些挑战和课题需要在未来解决，以构建有效的和可扩展的异构多智能体系统.

参 考 文 献

[1] Ling Xing-Hong, Li Jie, Zhu Fei, et al. Asynchronous advantage actor-critic with double attention mechanism. Chinese Journal of Computers, 2020, 43(1): 93-106(in Chinese)
(凌兴宏，李杰，朱斐等. 基于双重注意力机制的异步优势行动者评论家算法. 计算机学报，2020，43(1)：93-106)

[2] Li Jing-Chen, Shi Hao-Bin, Hwang Kao-Shing. A multi-agent reinforcement learning method based on self-attention mechanism and policy mapping recombination. Chinese Journal of Computers, 2022, 45(9): 1842-1858(in Chinese)
(李静晨，史豪斌，黄国胜. 基于自注意力机制和策略映射重组的多智能体强化学习算法. 计算机学报，2022，45(9)：1842-1858)

[3] Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839

[4] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Stockholm, Sweden, 2018: 2085-2087

[5] Rashid T, Samvelyan M, Schroeder C, et al. QMIX: Monotonic value function factorization for deep multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4295-4304

[6] Son K, Kim D, Kang W J, et al. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 5887-5896

[7] Wang J, Ren Z, Liu T, et al. QPLEX: Duplex dueling multi-agent Q-learning//Proceedings of the International Conference on Learning Representations. Virtual Event, Austria, 2021: 1-27

[8] Sukhbaatar S, Fergus R. Learning multi-agent communication with backpropagation//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 2252-2260

[9] Das A, Gervet T, Romoff J, et al. TarMAC: Targeted multi-agent communication//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 1538-1546

[10] Foerster J, Assael I A, De Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 1782-1791

[11] Das A, Kottur S, Moura J M F, et al. Learning cooperative visual dialog agents with deep reinforcement learning//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2951-2960

[12] Singh A, Jain T, Sukhbaatar S. Learning when to communicate at scale in multi agent cooperative and competitive tasks//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1236-1251

[13] Kim D, Moon S, Hostallero D, et al. Learning to schedule communication in multi-agent reinforcement learning//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1724-1732

[14] Wang T, Wang J, Zheng C, et al. Learning nearly decomposable value functions via communication minimization//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 2168-2182

[15] Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 2022-2032

[16] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 793-803

[17] Yang X, Deng C, Liu T, et al. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1992-2003

[18] Meneghetti D, Bianchi R. Towards heterogeneous multi-agent reinforcement learning with graph neural networks. arXiv preprint arXiv:2010.02663, 2020

[19] Seraj E, Wang Z, Paleja R, et al. Heterogeneous graph attention networks for learning diverse communication. arXiv preprint arXiv:2108.09568, 2021

[20] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing, 2021, 45(2): 48-62

[21] Song W, Xiao Z, Wang Y, et al. Session-based social recommendation via dynamic graph attention networks//Proceedings of the 12th ACM International Conference on Web Search and Data Mining. Naples, Italy, 2019: 555-563

[22] Tong H, Faloutsos C, Pan J Y. Fast random walk with restart and its applications//Proceedings of the 6th International Conference on Data Mining. Hong Kong, China, 2006: 613-622

[23] Samvelyan M, Rashid T, De Witt C S, et al. The StarCraft multi-agent challenge//Proceedings of the 18th International Conference on Autonomous Agents and Multi Agent Systems. Montreal, Canada, 2019: 2186-2188

[24] Maaten L. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

[25] Yang S, Yang B, Kang Z, et al. IHG-MA: Inductive heterogeneous graph multi-agent reinforcement learning for multi-intersection traffic signal control. Neural Networks, 2021, 139(7): 265-277

[26] Codeca L, Härri J. Monaco SUMO traffic (most) scenario: A 3D mobility scenario for cooperative ITS. EPiC Series in Engineering, 2018, 2(1): 43-55

[27] Yu C, Feng Y, Liu H X, et al. Corridor level cooperative trajectory optimization with connected and automated vehicles. Transportation Research Part C: Emerging Technologies, 2019, 105(8): 405-421

[28] Zang X, Yao H, Zheng G, et al. Metalight: Value-based meta-reinforcement learning for traffic signal control//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 1153-1160

[29] Wei H, Xu N, Zhang H, et al. Colight: Learning network-level cooperation for traffic signal control//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China, 2019: 1913-1922

[30] Devailly F X, Larocque D, Charlin L. IG-RL: Inductive graph reinforcement learning for massive-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 7496-7507

DU Wei, Ph. D. candidate. His research interests include deep learning, reinforcement learning.

DING Shi-Fei, Ph. D., professor, Ph. D. supervisor. His research interests include artificial intelligence and pattern recognition.

GUO Li-Li, Ph. D., lecturer. Her research interests include deep learning and emotional computing.

ZHANG Jian, Ph. D., lecturer. His research interests include machine learning and pattern recognition.

DING Ling, Ph. D. candidate. Her research interests include deep learning and graph machine learning.

Background

In the real world, there are many heterogeneous multi-agent scenarios, in which agents have different attributes or features and even the state space and action space are different. Multi-agent reinforcement learning(MARL) provides a promising way to model and develop such systems. However, the MARL encounter scalability issue and partial observability constraint. To tackle these challenges, the framework of centralized training with decentralized execution is proposed and then widely used in the MARL domain.

Based on the CTDE framework, the value decomposition methods and communication learning methods provide different solutions for further exploiting this framework. The value decomposition methods decompose the global value function into the set of the individual value function for each agent based on the Individual-Global-Max principle. Nevertheless, most of these methods focus on learning full decomposition, in which each agent generates local value functions only based on its local observation. In the real world, many heterogeneous multi-agent tasks necessitate information obtained from other agents to coordinate their actions effectively. In the absence of communication mechanisms, agents are limited by partial observability and randomness, which will intensify the uncertainty of agents' state and action to other agents in the decentralized execution period, leading to catastrophic discoordination.

Communication learning mechanisms in MARL have become very prevalent. However, in the heterogeneous environment, agents have different states and action spaces, and most existing multi-agent communication learning frameworks do not explicitly model such heterogeneity. Without a specific heterogeneous communication mechanism, agents of different types may not distinguish the heterogeneity in received messages and extract valuable information for value function estimation and decision-making. As a result, communication can become useless and may even degrade performance.

To tackle these challenges, we present a framework that combines the value decomposition and communication scheme for heterogeneous multi-agent settings. The proposed method maintains the advantages of scalability and stability of value function decomposition and promotes better collaboration and decision-making of agents by utilizing diverse information interactions between heterogeneous agents. The proposed framework provides a new idea for the field of heterogeneous multi-agent reinforcement learning. To the best of our knowledge, our work is the first attempt to combine the communication learning method based on graph convolution network and the value function learning method to develop the heterogeneous multi-agent system.

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62276265 and 61976216.