# 深度强化学习下的多智能体思考型半多轮通信网络

邹启杰, 汤　宇, 高　兵†, 赵锡玲, 张哲婕

(大连大学 信息工程学院, 辽宁 大连 116622)

**摘要**: 针对多智能体系统在合作环境中通信内容单一和信息稀疏问题, 本文提出一种基于多智能体深度强化学习的思考型通信网络(TMACN). 首先, 智能体在交互过程中考虑不同信息源的差异性, 智能体将接收到的通信信息与自身历史经验信息进行融合, 形成推理信息, 并将此信息作为新的发送消息, 从而达到提高通信内容多样化的目标; 然后, 该模型在软注意力机制的基础上设计了一种半多轮通信策略, 提高了信息饱和度, 从而提升系统的通信交互效率. 本文在合作导航、捕猎任务和交通路口3个模拟环境中证明, TMACN对比其他方法, 提高了系统的准确率与稳定性.

**关键词**: 多智能体系统; 合作环境; 深度强化学习; 通信网络

**引用格式**: 邹启杰, 汤宇, 高兵, 等. 深度强化学习下的多智能体思考型半多轮通信网络. 控制理论与应用, 2025, 42(3): 553 – 562

DOI: 10.7641/CTA.2023.30028

# The thinking communication network with semi-multiple communication cycles under the multi-agent deep reinforcement learning

ZOU Qi-jie, TANG Yu, GAO Bing†, ZHAO Xi-ling, ZHANG Zhe-jie

(School of Information Engineering, Dalian University, Dalian Liaoning 116622, China)

**Abstract:** To address the problem of single communication content and sparse information in multi-agent systems under a cooperative environment, this paper proposes a thinking multi-agent communication network (TMACN) based on deep reinforcement learning of multi-agent. Firstly, the agent considers the differences of different information sources in the interaction process, and the agent fuses the received communication information with their own historical experience information to form inference information, and use this information as a new sent message, so as to achieve the goal of improving the diversity of communication contents. Then, the model designs a semi-multi-round communication strategy based on the soft attention mechanism, which improves the information saturation and thus enhances the communication interaction efficiency of the system. This paper demonstrates that TMACN improves the accuracy and stability of the system compared to other methods in three simulated environments: cooperative navigation, hunting task and traffic junction.

**Key words:** multi-agent systems; cooperative environment; deep reinforcement learning; communication network

**Citation:** ZOU Qijie, TANG Yu, GAO Bing, et al. The thinking communication network with semi-multiple communication cycles under the multi-agent deep reinforcement learning. *Control Theory & Applications*, 2025, 42(3): 553 – 562

## 1　引言

多智能体系统 (multi-agent system, MAS) 是通过多个智能体之间的协同合作来完成特定任务的系统, 该系统在分布式控制、远程调度和建模分析等领域应用广泛[1−3]. 深度强化学习将深度学习的高感知能力和强化学习的序贯决策优势有效结合起来, 形成了从环境输入到控制输出的端对端学习模型. 多智能体深度强化学习将深度强化学习引入多智能体系统中, 进一步提高了多智能体之间的协同合作能力. 在即时游戏领域, AlphaStar[4]和Openai Five[5]分别在星际争霸2和Dota2中达到了顶尖人类玩家的水平, 展现了多智能体深度强化学习在游戏领域中的极高应用价值.

收稿日期: 2023−01−20; 录用日期: 2023−12−07.

†通信作者. E-mail: 2366180678@qq.com; Tel.: +86 18340857180.

本文责任编委: 高会军.

国家自然科学基金项目(61673084), 2021年辽宁省教育厅项目(LJKZ1180)资助.

Supported by the National Natural Science Foundation of China (61673084) and the Project of Liaoning Provincial Department of Education in 2021 (LJKZ1180).

多智能体之间的通信是实现协同合作的关键所在, 其通过信息交互自动学习协同合作行为, 以进一步提高多智能体系统的精确性与稳定性[6]. 多智能体通信可分为显示通信和隐式通信两种, 其中大部分的工作集中于前者. 显示通信和隐式通信本质上的区别在于是否具有实际的通信信道, 显示通信可以快速、高效地完成多智能体之间数据和信息的转移与交换, 从而实现许多在隐式通信下无法完成的先进协调协作策略[7]. 目前, 多智能体之间的通信优化可以聚焦于4个方向: 通信对象优化、通信内容优化、通信时机优化以及在限制条件下的通信优化.

在多智能体深度强化学习领域, RIAL&DIAL[8]对通信内容进行了优化, RIAL提出使用深度Q网络通过反向传播来学习通信模型. DIAL将智能体的动作信息和状态信息作为通信信息, 在通信内容方面实现了对通信信道的有效训练, 但是DIAL无法解决非静止环境中的问题. BicNet[9]通过引入一个双向协调网络对通信内容进行优化, 从而解决了动态环境中的复杂问题. 然而BicNet中的智能体共享全局信息, 这在现实场景下难以实现. MADDPG[10]将演员–评论(actor-critic, AC)[11]方法扩展到多智能体协调领域, 能够在部分可观察的情况下解决动态环境中的复杂问题. 由于MADDPG引入了其他智能体的观察和动作来解决协调问题, 这导致状态空间过大, 无法应用于大规模多智能体环境. 为了解决动态环境中的大规模问题, Sukhbaatar等[12]提出了CommNet模型, 该模型通过优化通信内容, 采用智能体隐层状态作为通信信息, 从而使得智能体在进行协同决策时更加准确和高效. 然而, CommNet并没有考虑来自不同智能体信息的重要性, 而是简单地平均接收来自其他智能体的信息, 这将导致智能体可能接收一些杂乱或干扰的消息, 降低了协同合作的效果.

为了在带宽受限的情况下实现信息的稳定传输, 近年来研究人员提出了一些解决方案. GACML[13]提出了一种门控机制以自适应性删除无用消息, 从而将通信消息限制在系统控制范围内. IMAC[14]通过协调调度直接压缩通信信息以达到节省带宽的目的. SchedNet[15]通过调度器进行消息编码操作和消息选择操作, 能够决定信息是否发送, 从根本上减少通信内容, 从而有效解决了带宽不足的问题. Message-Dropout MADDPG[16]采用集中训练分散执行框架, 在训练阶段以一定的概率丢弃所接收的消息, 并通过将丢弃的块单元的权重乘以校正概率来补偿这种影响, 具有较高的通信鲁棒性. 另一方面, IC3Net[17]在CommNet的基础上优化通信时机, 利用门控机制让智能体自主地决定通信交互时机, 降低了对通信带宽的要求, 进一步提升了通信的效率.

为了减少不必要的通信来实现高效的通信交流, 软注意力机制[18]给众多研究人员提供了一条新的路径. ATOC[19]提出用注意力单元来构建通信组, 并利用该机制筛选进行信息交互的智能体, 使得智能体可以自主选择通信时机和通信对象. TarMac[20]同样利用软注意力机制, 提出了向量查询方法, 即TarMac在通信信道中广播信息的同时附带一个密钥, 当接收方接收到该密钥后, 通过计算相关度来选择性接收该消息, 从而实现了高效的通信交流. 两者的不同点在于, ATOC在发送阶段选择性地发送信息, 而TarMac在接收阶段选择性地接收信息. 值得注意的是, TarMac和CommNet都采用了多轮通信的方法, 即智能体在执行动作之前进行多次通信交互. CommNet在执行动作之前进行两次通信, 而TarMac的通信次数是超参数. 同时, 两者的多轮通信内容只考虑智能体部分观察信息, 没有考虑其他信息. 另一方面, ATOC和TarMac都没有充分重视基于内存的智能体交互作用. Memory-Driven MADDPG[21] 提出一种基于内存的共享通信机制, 其让智能体通过内存设备建立通信协议, 从而提高了多智能体的协调能力. Niu等[22]提出多智能体图注意力通信MAGIC方法, 通过图注意力机制学习其他智能体对于当前智能体的重要程度, 并改进了基于卷积计算的软注意力机制, 从而有效解决通信时机和通信对象的问题. 表1总结了上述方法之间的主要对比.

表 1 多智能体显示通信方法对比

Table 1 Comparison of multi-agent display communication methods

| 算法 | 通信策略 | 通信优化 | 基础方法 |
|---|---|---|---|
| DIAL[8] | 一次通信 | 通信内容(观察信息+动作信息) | Q-Learning |
| CommNet[12] | 两次通信 | 通信内容(观察信息) | 策略梯度 |
| BiCNet[9] | 一次通信 | 通信内容(共享信息+观察信息) | AC |
| ATOC[19] | 一次通信 | 通信时机、通信对象(观察信息) | AC |
| GACML[13] | 一次通信 | 通信限制、通信内容(观察信息) | AC |
| TarMAC[20] | 多次通信 | 通信内容(观察信息) | AC |
| TMACN(此论文) | 半多轮通信 | 通信对象、通信内容(观察信息+历史信息) | AC |

上述研究方法从通信对象、通信内容、通信时机以及现实带宽与噪声限制等不同方面进行了探索和优化, 有效提升了多智能体系统的整体性能. 然而在真实世界中, 智能体在进行通信交互中需要考虑多样化的信息源, 而上述方法大多只考虑了智能体自身的观察信息, 存在一定的局限性. 因此, 本论文提出了一种思考型多智能体通信网络(thinking multi-agent communication network, TMACN), 旨在通过将智能体的历史经验和当前接收信息进行思考与推理, 提取出相关信息, 并采取一轮通信和多轮通信迭代交互的半多轮通信方式, 在减轻带宽资源消耗的同时提高信息饱和度, 优化通信内容, 增强系统的稳定性与精确性.

本文的主要贡献在于: 1) 提出一种新的通信算法TMACN, 该算法能够将智能体历史信息和当前接收信息进行融合, 提取相关信息, 从而达到信息多样化的目的; 2) 设计了一种基于软注意力机制的半多轮通信策略, 通过迭代执行两种不同的通信模式, 解决了信息稀疏问题; 3) 在软注意力机制中, 设计了两种评分函数, 并采用分时执行的方式来平衡智能体附近区域和偏远区域信息的重要程度, 进一步优化了通信效果.

## 2　问题描述

多智能体的通信优化是一种典型的解决多智能体协同合作问题的方法, 其主要思想在于将通信策略的学习纳入到多智能体系统中, 以提高智能体之间的交互效率, 从而优化系统的训练速度以及整体策略的稳定性. 图1为多智能体通信框架图.

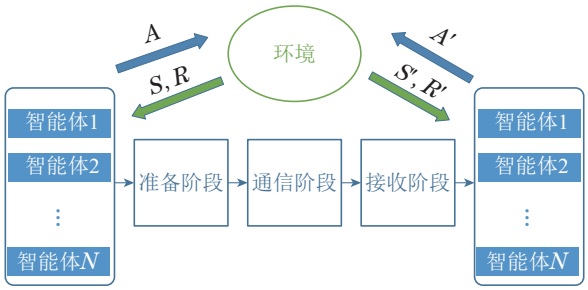



图 1　多智能体通信框架图

Fig. 1　Block diagram of the multi-agent communication

在现实生活中, 由于多智能体自身的观察范围受限且多智能体之间存在合作关系. 因此, 多智能体协作合作被视为一种分散式部分可观察马尔科夫过程(decentralized partially observable Markov decision processes, Dec-POMDPs)[23]. 在这种设定下, 多智能体通信问题可以基于Dec-POMDPs进行建模, 并通过九元组来表示: $\langle N, S, C, A, T, R, \Omega, O, \gamma\rangle$. 多智能体之间的通信分3个步骤, 分别是准备阶段、通信阶段和接收阶段.

1) 数量$N$.

$N$表示智能体的数量, 在部分情况下, 智能体的数量不固定, 可能会发生变化.

2) 联合状态$S$.

对于$N$个智能体来说, $S$表示为所有智能体的状态的组合$\{s_1^t, s_2^t, \cdots, s_N^t\}$, $s^t$为模拟环境中智能体的具体位置信息, 其位置可以用笛卡尔坐标系$(x^t, y^t)$表示.

3) 通信交互$C$.

$C$表示多智能体之间的通信, 包括信息的准备、发送以及接收.

4) 联合动作空间$A$.

$A = \{a_1, a_2, \cdots, a_N\}$表示所有智能体的一个联合动作, $a_i$是智能体$i$在状态$s_i$下可能采取的动作, 在模拟环境中可以表示为方向上的移动或停止.

5) 状态转移概率函数$T: S \times A \Rightarrow S'$.

$T = P(S'|S, A)$表示多智能体处于联合状态$S$并采取联合动作$A$的情况下, 下一时刻到达另一个新状态的概率. 在模拟环境中, 这可表示为智能体位置信息的变化概率.

6) 全局奖励函数$R$.

$R = \{R_1, R_2, \cdots, R_N\}$为所有智能体共享的一个奖励函数, 表示多智能体系统在联合状态$S$下执行联合动作$A$带来的环境回报. 其中$R_i$为智能体$i$的奖励函数, 其具体公式如下, $r_t$为智能体$t$时刻的立即奖励.

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+1} + \cdots = r_t + \gamma R_{t+1},$$

其中$\gamma$为折扣因子, $\gamma \in [0, 1]$, 其表示对未来奖励期望的折扣, $\gamma$数值越小, 表示对近期奖励越看重.

7) 联合观察空间$\Omega$.

$\Omega = \{\Omega_1, \Omega_2, \cdots, \Omega_N\}$为智能体的一个联合观察, $\Omega_i$表示智能体$i$在状态$s_i$下的局部观察, 该局部观察为模拟环境中智能体可观察到的附近信息, 例如障碍物信息、其他智能体位置信息等.

8) 联合观察值转移函数$O$.

$O = (o_1, o_2, \cdots, o_N)$, $o_i = P(\Omega_i|s', a)$定义为智能体$i$在执行动作$a$到达新状态$s'$后观察到$\Omega_i$的概率. 由于现实世界中存在传感器受外界干扰而失灵, 导致输出错误观察结果或观察不完全的问题. 因此, 在模拟环境中加入干扰因子来产生观察信息干扰, 从而更好模拟真实环境.

在决策过程的每一时刻, 首先$N$个智能体在给定状态$S$下, 通过联合观察值转移函数$O$获得局部观察$\Omega$; 其次, 智能体之间进行通信交互$C$, 然后根据策略做出联合动作$A$, 该联合动作$A$根据转移概率函数$T$将系统引向下一个状态$S'$; 最后, 环境基于此反馈一个共享全局奖励函数$R$. 整个系统的目的在于最大化

总回报$U=\sum_{i=1}^{N}R_i$.

在多智能体之间的通信中, 常出现通信内容单一、通信信息稀疏和无关信息干扰问题. 这些问题主要由以下3方面原因引起: 1) 在准备阶段, 智能体只使用自身部分环境观察作为通信消息, 缺乏将过去经验等信息作为参考, 导致通信内容单一且通信效率低下; 2) 多数通信只进行单轮通信便做出决策, 而缺乏迭代通信处理, 从而导致信息稀疏; 3) 在接收阶段, 智能体未评判通信消息的重要程度, 而是采取完全接收策略, 从而引发了无关信息的干扰问题.

针对上述问题, 在接收阶段, 可引入软注意力机制以筛选通信消息, 剔除无关信息. 同时, 智能体将新接收的消息和历史经验相融合, 提取出相关消息, 并通过半多轮通信方式发送. 通过上述措施, 在平衡探索和开发的基础上, 能显著提高智能体之间的交互效率, 从而提高多智能体协作能力.

## 3 TMACN算法

### 3.1 TMACN算法总体设计

本文设计的多智能体通信体系结构采用集中训练分散执行(centralized training and decentralized execution, CTDE)框架和AC算法, 同时采用时序差分(temporal-difference learning, TD)算法[24]对网络参数进行更新. 多智能体系统框架图如图2所示.

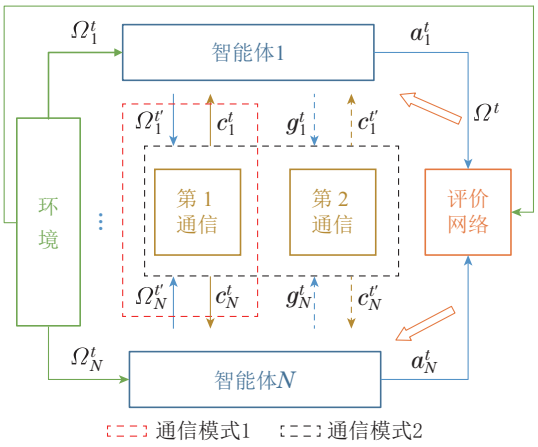



图 2 多智能体系统框架图

Fig. 2 Block diagram of the multi-agent system

本文所设计的半多轮通信方式是指, 在多智能体执行动作之前, 交互过程分为两种模式: 通信模式1和通信模式2. 其中, 通信模式1为单轮通信, 即只进行第1通信便执行动作; 通信模式2则为双轮通信, 先进行第1通信, 再进行第2通信, 最后执行动作. 半多轮通信框架图如图3所示.

对于多智能体系统而言, 在执行回合数为奇数时, 采用通信模式1. 智能体首先获取某状态下的观测信息, 随后进行第1通信以交换并筛选智能体的观测信息; 其次, 多智能体作出联合动作并与环境交互; 最后对评价网络和行为网络进行更新.

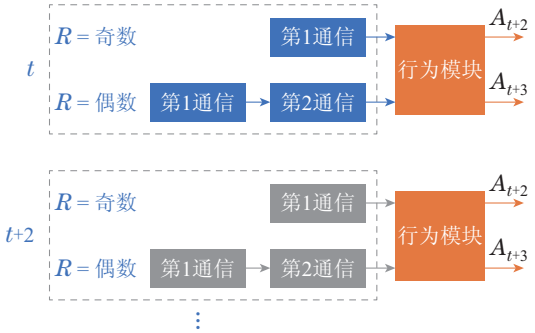



图 3 半多轮通信框架图

Fig. 3 Block diagram of the half-multi-round communication

在执行回合数为偶数时, 多智能体系统采用通信模式2. 智能体同样首先获取观测信息, 并进行第1通信. 其次, 智能体进行第2通信以交换推理后的新通信信息, 随后作出联合动作并与环境交互, 最后对两个网络进行更新. 在整个多智能体协同合作过程中, 智能体按照执行回合数的奇偶顺序交替执行上述操作.

单智能体设计图如图4和图5所示, 图4为通信模式1框架图, 图5为通信模式2框架图. 智能体由4个模块组成, 分别是编码器、记忆模块、思考模块以及行为模块. 编码器对局部观察信息进行编码, 以便后续计算; 记忆模块储存智能体历史经验信息; 思考模块将历史经验信息与通信信息融合, 并提取出相关信息, 该信息作为第2通信的输入信息; 行为模块用于控制智能体做出相应的动作. 训练阶段, 在每个时间步$t$, 智能体$i$首先通过环境获得其局部观察信息$\Omega_i^t$, 随后形成编码信息$\Omega_i^{t'}$, 然后交替采用不同的通信模式, 最后执行动作$a_i^t$.

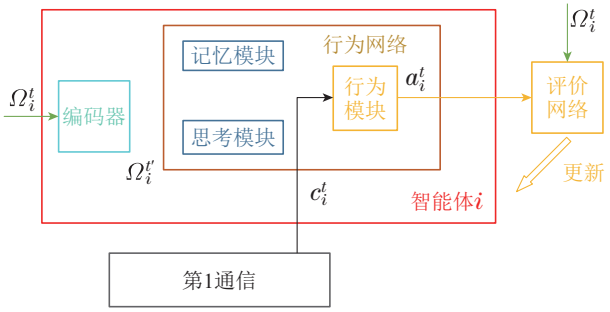



图 4 通信模式1框架图

Fig. 4 Block diagram of the communication mode I

在通信模式1中, 智能体$i$进行第1通信收到注意力信息$c_i^t$, 随后智能体根据策略$\pi_i$做出相应的动作$a_i^t$, 同时与环境交互获得奖励信息$r_i^t$和新的局部观察信息$\Omega_i^{t+1}$, 最后对评价网络和行为网络进行更新.

在通信模式2中, 智能体$i$在完成第1通信后, 智能体将注意力信息$c_i^t$和过去经验信息$m_i^t$进行思考推理,

从而形成新的通信消息$g_i^t$, 该消息作为第2通信的交互内容, 第2通信完毕后再做出动作与环境交互, 最后对两网络进行更新. 整个系统训练完毕之后, 智能体在执行阶段删除评价网络, 仅依靠行为网络做出决策.

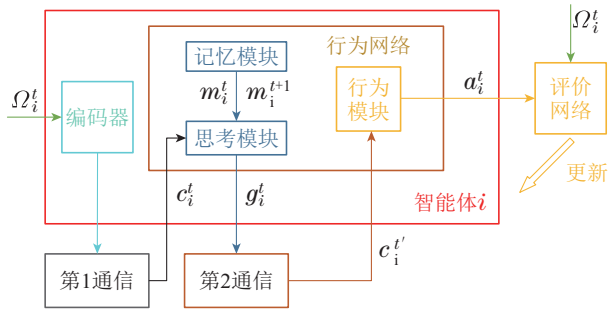



图 5　通信模式2框架图

Fig. 5　Block diagram of the communication mode II

## 3.2　半多轮通信机制

### 3.2.1　基于软注意力机制的第1通信模块

软注意力机制采用两种不同的评估函数来计算权重信息, 并在不同时间段计算每个时间步的权重. 此外, 基于第1种评判函数, 该机制使用门控机制过滤低权重信息, 以此平衡智能体附近区域和偏远区域信息重要性.

图3中第1通信的交互方式采用软注意力机制, 该机制具体设计如图6所示. 在当前时刻$t$, 智能体由门控循环单元(gate recurrent unit, GRU)单独控制[25], 每个GRU的输入是对应智能体的部分环境观察编码信息$\Omega_i^{t'}$, 其输出隐层信息$h_i^t$. 对于每个智能体$i$, 软注意力模块根据来自不同智能体的消息赋予不同权重$w_i^t$, 然后通过门机制筛除低权重的智能体信息, 最后将不同权重的信息进行加权求和即可得到注意力信息$c_i^t$.

在$t$时刻, GRU将自身观察编码信息转换成隐层信息$h_i^t$, 而软注意力模块对这些输入信息进行权重的分配, 从而让智能体获知不同信息对其的重要性. 具体而言, 该模块通过加权累加真实消息和权重信息得到注意力信息$c_i^t$, 随后该注意力信息作为行为模块的输入, 计算公式如式(1)所示:

$$c_i^t = \sum_{j=1}^{N} w_{ij}^t h_j^t, \tag{1}$$

其中$w_{ij}^t$表示智能体$j$的信息对于智能体$i$的重要程度, 其值范围在$[0, 1]$, 值越大, 即信息越重要, 其计算公式如下:

$$\{w_{i1}^t, w_{i2}^t, \cdots, w_{iN}^t\} = \mathrm{softmax}(e_{i1}^t, e_{i2}^t, \cdots, e_{iN}^t), \tag{2}$$

其中: $e_{ij}^t$表示发送端智能体$j$和接收端智能体$i$之间的相关程度, 值越大表示越相关, 其计算如式 (3) 所示; softmax()为归一化函数, 用来对输出值进行归一化处理, 其计算如式(4)所示.

$$e_{ij}^t = \mathrm{Score}(h_j^t, h_i^t), \tag{3}$$

$$\mathrm{softmax}(w_{ij}^t) = \frac{\exp(e_i^t)}{\sum_{j=1}^{N} \exp(e_j^t)}, \tag{4}$$

式(3)中Score()为评分函数, 用来计算来自不同智能体消息和自身的相关性, 这里采用两种评分函数, 第1种采用相关度设计, 即

$$\mathrm{Score}(h_j^t, h_i^t) = \frac{{h_i^t}^{\mathrm{T}} h_j^t}{\sqrt{d_k}}, \tag{5}$$

其中$d_k$为输入智能体的总数量; 第2种评分函数采用平均策略, 即

$$\mathrm{Score}(h_j^t, h_i^t) = \frac{1}{d_k}. \tag{6}$$

两种评分函数按时间顺序交替执行. 第2种评分函数主要为全局信息而设计, 避免智能体忽略偏远区域智能体的信息, 从而减少多智能体系统陷入局部最优解的概率.

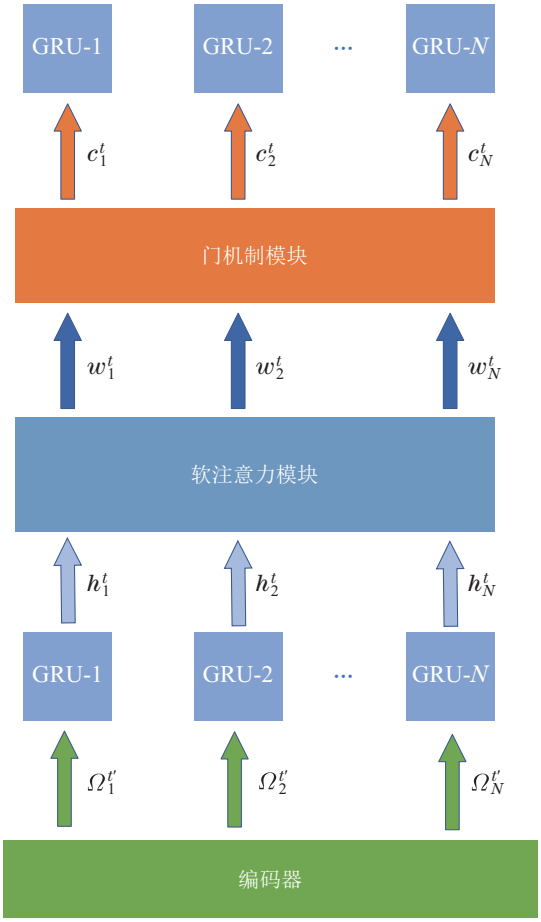



图 6　软注意力机制框架图

Fig. 6　Block diagram of the soft-attention mechanism

为了更好筛除不必要的信息, 在第1种评分函数的注意力机制的基础上引入门机制, 此处设定一个阈值$\tau$, 当满足式(7)时, 智能体$j$发送的消息则被筛出, 不予考虑, 反之则通过.

$$w_{ij}^t < \tau. \tag{7}$$

### 3.2.2 基于思考型交互的第2通信模块

本文设计了一种仿效人类思考的方法, 即智能体接收信息后不会立即采取行动, 而是将接收到的通信消息与自身历史经验消息相融合, 提取出相关信息, 然后进行通信交互. 此方法能够增强多智能体之间通信内容的多样性, 从而达到提高通信交互效率的目标.

如图5所示, 记忆模块和思考模块共同构成了信息融合机制. 该机制将新消息数据和历史经验数据进行融合, 进行思考和推理, 从而形成新的通信信息$g_i^t$. 该信息随后通过第2通信形成新的注意力信息$c_i^{t'}$, 最终输送至行为模块. 信息融合机制主要进行 3 个工作: 1) 融合第1通信信息和历史经验信息以提取相关信息; 2) 生成新的经验消息; 3) 对融合后的信息进行第2通信交互.

信息融合机制中思考模块充当了融合器的作用, 该模块在各个智能体之间进行共享, 其输入由智能体$i$选择性接收后的信息$c_i^t$和自身上一时刻记忆信息$m_i^{t-1}$组成, 输出为融合后的信息$g_i^t$, 计算公式如下:

$$g_i^t = c_i^t \otimes m_i^{t-1}, \tag{8}$$

其中$\otimes$表示哈达玛积, 该新消息$g_i^t$作为第2通信中的新通信消息发送给其他智能体.

在信息融合之后, 记忆模块将上一时刻的记忆信息和注意力信息进行叠加, 并通过线性变化生成新的记忆信息, 以此来更新记忆信息, 计算公式如下:

$$m^t = W(m_i^{t-1} + c_i^t), \tag{9}$$

其中$W$是一个线性变换的矩阵, 用于转换叠加数据的格式, 该记忆信息$m_i^t$将存储在记忆模块中.

随后进行第2通信, 智能体$i$根据第1通信计算出来的权重信息选择性接收新通信消息$g_i^t$, 形成新的注意力信息$c_i^{t'}$, 该注意力信息将作为行为模块的输入信息, 计算公式如下:

$$c_i^{t'} = \sum_{j=1}^{N} w_{ij}^t g_j^t. \tag{10}$$

综上所述, 通过上述机制, 信息融合能够有效解决智能体通信内容单一的问题. 此外, 半多轮通信能够缓解信息稀疏问题, 从而有助于优化通信交互的效果.

## 4 实验

### 4.1 实验环境设定

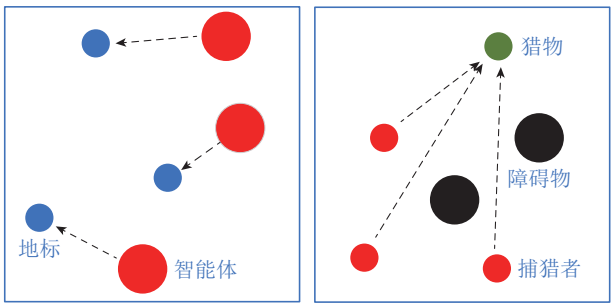
本文利用合作导航、捕猎任务和交通路口3个模拟环境来评估Tarmac[20], CommNet[12], MADDPG[10]和TMACN的性能. 其中, MADDPG是典型的隐式通信, 其缺乏具体的通信信道; CommNet建立了一种新的显式通信框架, 其通信信息是平均分配的隐层信息; Tarmac采用了基于向量查询的注意力机制通信模式. 此3个实验环境可以调整任务难度, 从而可以充分评估各基准方法的性能.

1) 合作导航环境.

在此环境中, $N$个智能体需要合作到达$N$个地标, 如图7(a)所示. 智能体能够观察其视野范围内其他智能体和地标的位置, 多智能体的集体奖励根据智能体与每个地标的接近程度判定. 实验中有3种事件会受到惩罚: 第一, 智能体在合作的过程中出现碰撞$C$; 第二, 智能体与地标的距离$D$超出预设范围; 第三, 随着系统时间流逝, 智能体也会受到一定的惩罚$T$. 因此, 智能体之间需要交流合作, 学会在避免碰撞的同时覆盖所有地标.

2) 捕猎任务环境.

在此环境中, 如图7(b)所示. 捕食者需要合作来捕捉猎物, 同时猎物也需要合作来对抗捕食者的抓捕, 捕食者和猎物之间构成竞争关系. 每个智能体可以观察视野范围内的猎物、捕食者和静态地标. 捕食者移动速度较慢, 视野较大; 猎物移动速度较快, 视野较小. 每次捕食者和猎物发生碰撞时, 捕食者会得到奖励, 而猎物则会受到惩罚, 同时随着时间的流逝, 捕食者也会受到一定的惩罚. 因此, 无论是捕食者还是猎物之间都需要交互合作, 以达到捕食或逃跑的目的.




(a) 合作导航环境 (b) 捕猎环境

图 7 合作导航环境与捕猎环境

Fig. 7 Cooperative navigation environment and hunting environment

3) 交通路口环境.

此环境由交叉路线和车辆(智能体)构成, 如图8所示. 为了安全有序通过十字路口, 车辆必须进行通信合作以避免碰撞. 车辆以固定概率进入环境, 其行走路线预先已经设定好. 环境中的最大车辆数是固定的, 车辆能够执行两个动作: 前行和停止. 当车辆发生碰撞时, 车辆会得到惩罚, 同时随着时间的流逝, 车辆也会收到一定的惩罚. 该环境可以调整道路的数量或最大智能体的数量来设定难度大小, 在不同难度下, 碰撞率或成功率是评价不同方法性能优越的标准.

实验配置中评价网络采用TD算法更新, 行为模块中的策略网络采用普通的策略梯度上升方法更新, 两者学习率都采用0.001, 优化器选择ADAM. 对于整体

数据部分, 实验的batch-size大小设置为256, 奖励折扣系数设定为0.95.

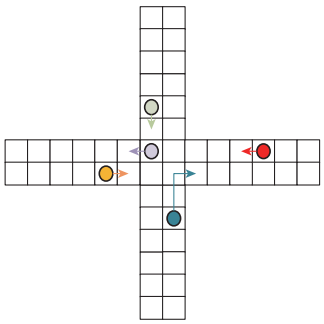


图 8　交通路口环境

Fig. 8　Traffic junction environment

本文利用交通路口实验对门机制阈值$\tau \in [0, 1]$进行测试选择, 实验将成功率作为性能指标, 实验结果如表2所示.

表 2　阈值实验表

Table 2　Threshold test table

| 实验方法 | $\tau = 0.55$ | $\tau = 0.65$ | $\tau = 0.75$ | $\tau = 0.85$ |
|---|---|---|---|---|
| TMACN | 81.3±0.5% | 86.7±0.6% | 90.8±0.6% | 88.1±0.6% |

根据实验结果, $\tau$为0.75能够取得较好结果, 后续的实验将$\tau$设置为0.75.

### 4.2　消融实验

本次实验旨在对通信结构进行消融, 即去除TMACN的信息融合机制, 得到其无信息融合机制的简化版(no information fusion, NIF). 为了验证信息融合模块对于复杂任务的影响, 此次实验在合作导航模拟环境中设置两种不同难度等级来测试信息融合模块的有效性. 本次实验将实验分数Score作为性能指标, 其最大值为0, 分数越高则代表性能越好, 其计算公式为

$$\text{Score} = k_1 C + k_2 D + k_3 T, \tag{11}$$

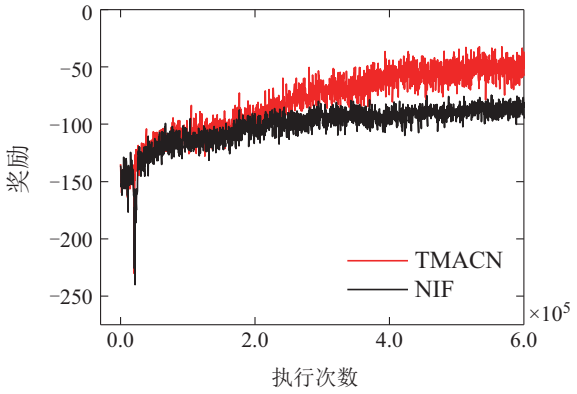
其中: $C$为碰撞次数, $D$为智能体与地标距离, $T$为系统时间, $k_1$, $k_2$以及$k_3$是权重因子.

在合作导航环境中, 第1种难度下智能体数量设置为3个, 障碍物设置为1个; 第2种任务难度将智能体数量设置为6个, 障碍物设置为2个. 具体实验对比如图9所示, 图9上图为第1种难度, 下图为第2种难度. 在第1种难度中, TMACN方法在训练后期总奖励平均约为−65, 而NIF总奖励平均约为−90. 在第2种难度下, 两种方法的总奖励都有所下降, 其中TMACN平均约为−80, 而NIF平均约为−110. 通过对比两种不同难度级别的实验结果, 能够证明信息融合模块可以有效提高增多智能体的性能.

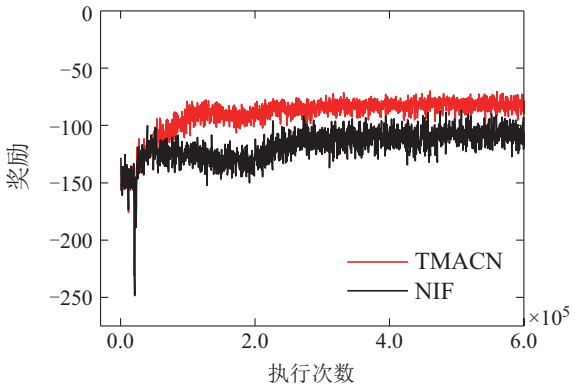
此外在不改变信息融合模块的同时, 设计了一种无半多轮通信方式的结构NHMC, 其只有一轮注意力机制通信. 本次实验在交通路口设置两种不同难度等级来测试半多轮通信方式的有效性, 实验使用成功率Suc作为评价标准, 定义如下:

$$\text{Suc} = \frac{\text{Num}_{\text{suc}}}{\text{Num}}, \tag{12}$$

其中: $\text{Num}_{\text{suc}}$表示成功事件, Num表示总事件. $\text{Num}_{\text{suc}}$定义为在达到允许步数后没有发生碰撞, 如果发生一个或多个碰撞, 则归类为失败实验, 结果如表3所示.




(a) $N = 3$




(b) $N = 6$

图 9　消融实验图

Fig. 9　Ablation diagram

表 3　消融实验表

Table 3　Ablation test table

| 实验方法 | $N = 4$ | $N = 8$ |
|---|---|---|
| TMACN | 90.8 ± 0.6% | 84.4 ± 0.8% |
| NHMC | 86.7 ± 0.5% | 80.7 ± 0.9% |

根据表3数据统计, 在智能体数量为4的情况下, TMACN相比无半多轮通信方结构NHMC的成功率提高了4.1%. 在智能体数量为8的情况下, TMACN相比无半多轮通信方式的结构NHMC的成功率提高了3.7%. 根据上述实两种不同难度的实验证明, 半多轮通信能够显著提高系统的性能.

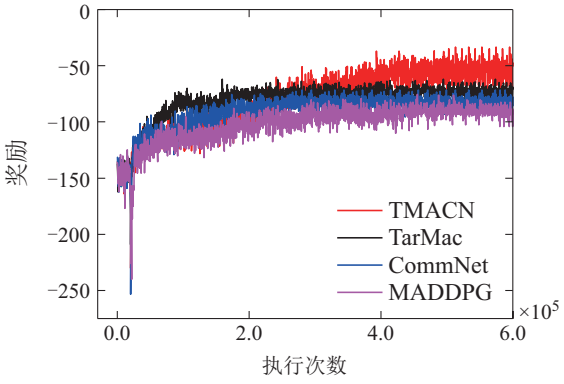
### 4.3 对比实验

#### 4.3.1 合作导航实验

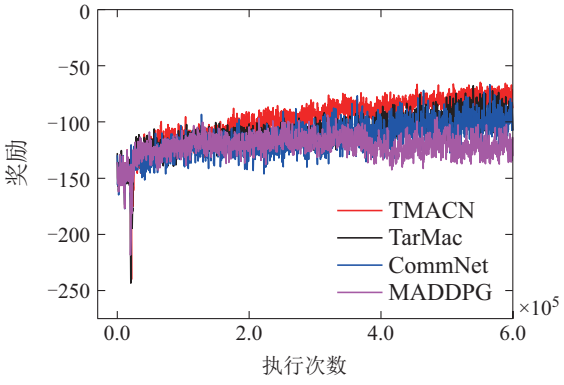
基线方法同样设置两种难度等级, 第1种难度智能体数量为3, 障碍物数量为1; 第2种难度智能体数量为6, 障碍物数量为2. 本次实验分别对TarMac, CommNet和MADDPG进行测试. MADDPG 无显示通信交流, 其和TMACN都采用了CTDE框架和AC算法. CommNet采用了简单的通信模式, 其通信内容是各个智能体隐层信息的平均值, 没有侧重性而言. 而TarMac相比于CommNet增加了一个基于向量查询的注意力机制, 使得智能体能够评估信息的重要性并选择性地接收通信消息, 并且TarMac通过多轮通信提高了信息的饱和度. 基线方法性能的对比通过奖励Score和碰撞率Col来体现, 碰撞率定义如下:

$$\mathrm{Col}=\frac{\mathrm{Num}_{\mathrm{col}}}{\mathrm{Num}}, \tag{13}$$

其中: $\mathrm{Num}_{\mathrm{col}}$表示发生碰撞的实验次数, Num则表示总实验次数. 实验结果如图10–11所示.




(a) $N=3$




(b) $N=6$

图 10 合作导航对比实验图

Fig. 10 Cooperative navigation comparison experiment diagram

如图10所示, 上下两图分别为智能体数量为3和6的两种不同难度实验结果. 图10上图第1种难度中, 对比其他算法, TMACN从40w步之后的奖励明显高于其他方法. 图10下图第2种难度, 其他算法在40w步后奖励明显低于TMACN, 并且后期TMACN收敛效果明显优于其他算法.

如图11所示, 左右两图分别为智能体数量分别为3和6的两种不同难度实验结果. 分析实验数据可以发现, 在$N=3$难度下, TMACN算法碰撞率比具有注意力机制的TarMac低3.8%, 比不带注意力机制的CommNet低8.9%, 但比隐式通信的MADDPG仅低0.5%. 在$N=6$难度下, 相比较$N=3$难度, TMACN的碰撞率提高了18.9%, TarMac提高了22.4%, CommNet提高了40.4%, MADDPG则提高了49.3%. 综上所述, 从碰撞率和稳定性两方面来看, 实验证明了TMACN具有有效性和高效性.

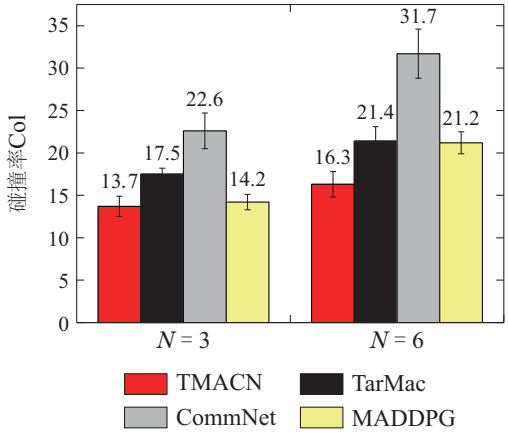



图 11 合作导航对比柱状误差图

Fig. 11 Bar error plot for cooperative navigation comparison

#### 4.3.2 捕猎任务实验

在此模型环境中, 捕猎者和猎物之间为竞争关系, 而双方内部是合作关系. 为了对比不同基线方法的性能, 本次实验将猎物的合作策略设置为MADDPG, 并保持不变. 同时, 将设置捕猎者的合作策略分别设置为TMACN, TarMac和CommNet.

本次实验将每个试验下的捕猎数量设定为性能标准, 其数量越大, 则该基线方法性能就越好. 实验设定了两种不同的场景, 第1种场景下, 捕猎者数量为5, 猎物为2; 第2种场景下, 捕猎者数量为10, 猎物为4. 实验结果如表4所示.

表 4 捕猎任务对比实验表

Table 4 Hunting task comparison experiment table

| 捕猎者 vs 猎物 | 5 vs 2 | 10 vs 4 |
|---|---|---|
| TMACN vs MADDPG | 2.54±0.07 | 3.71±0.15 |
| TarMac vs MADDPG | 2.32±0.12 | 3.56±0.27 |
| CommNet vs MADDPG | 1.92±0.09 | 2.73±0.21 |

根据表4数据统计结果, 在第1种场景下, TMACN性能超出CommNet 32.3%, 而与TarMac相比则略高9.5%. 在第2种场景下, TMACN性能超出CommN-

et35.9%, 略高于TarMac 6.9%. 综合而言, TMACN 对比上述方法, 系统性能分别提升了34.1%和8.2%.

在通信交互中, TMACN和TarMac算法都采用注意力机制来传递通信信息, 而CommNet算法则采用平均分配的方式. 这表明在一定时间内为了取得较好的结果, 需要集中智能体的精力来把握最近的机会, 而非将精力分散, 最后导致围捕性能下降. 同时, TMACN的捕猎数量略高于TarMac, 这证明了半多轮通信机制和信息融合机制的有效性.

### 4.3.3　交通路口实验

此模型环境仍然设置两种难度任务, 简单任务设置为2条道路, 智能体最大数量为4. 困难难度设置为4条道路, 智能体最大数量为8. 本次实验采用碰撞率Col(1-Suc)来评估基线方法的性能, 实验数据图如图12所示.

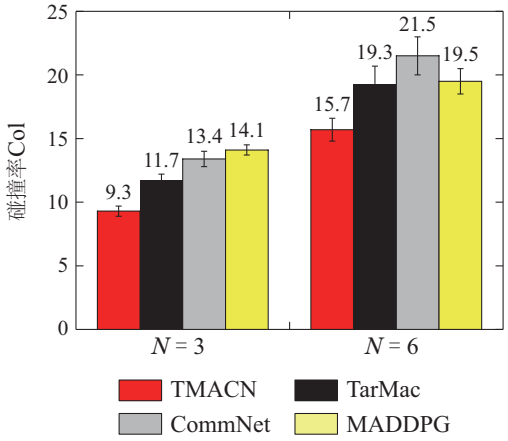



图 12　交通路口对比实验图

Fig. 12　Comparison experiment diagram of traffic junction

根据上述数据统计结果, 在智能体数量为4难度下, TMACN 碰撞率为 9.3%, 对比 TarMac, CommNet 以及MADDPG方法, 其成功率分别提高2.4%, 4.1%以及4.8%. 在智能体数量为8难度下, TMACN碰撞率为15.7%, 对比其他基准实验, 多智能体系统成功率分别提高3.6%, 5.8%以及3.8%. 两种不同难度实验证明TMACN方法失败次数低于其他基准实验, 从而证明了该方法的优越性.

## 5　总结

本文提出了一种新的通信算法TMACN, 在执行多智能体协同任务中, TMACN用于学习多智能体之间的通信策略. 该通信算法设计了两种通信模式, 并采用半多轮通信策略进行交互. 同时本文在通信交互中引入了信息融合机制, 用于结合新接受的消息和历史经验信息来形成推理信息. 通过上述方法, 能够在一定程度上解决通信内容优化问题, 从而提高多智能体之间通信交互的效率. 另外, 本算法在注意力机制中引入了门机制和两种不同的评分函数, 以实现有效信息提取与整体稳定性的平衡. 最后, 通过在模拟环境中进行了不同任务和不同智能体数量的实验, 实验结果表明本算法对比其他实验具有更好的性能. 期待该方法未来能够和基于效果评估的通信算法相结合, 比如ATOC和IC2.

## 参考文献:


[1] ZHANG K, YANG Z, BASAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, 2021, 13: 321 – 384.

[2] HU Y, FANG S, LEI Z, et al. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in Neural Information Processing Systems*, 2022, 35: 4874 – 4886.

[3] ZOU Qijie, LIU Shihui, ZHANG Yue, et al. Rapidly-exploring random tree algorithm for path replanning based on reinforcement learning under the peculiar environment. *Control Theory & Applications*, 2020, 37(8): 1737 – 1748.
(邹启杰, 刘世慧, 张跃, 等. 基于强化学习的快速探索随机树特殊环境中路径重规划算法. 控制理论与应用, 2020, 37(8): 1737 – 1748.)

[4] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, 575(7782): 350 – 354.

[5] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with large scale deep reinforcement learning. *ArXiv Preprint*, 2019, arXiv: 1912.06680.

[6] ZOU Q, HU Y, YI D, et al. Cooperative multiagent attentional communication for large-scale task space. *Wireless Communications and Mobile Computing*, 2022, 1: 1 – 13.

[7] ZOU Qijie, JIANG Yajun, GAO Bing, et all. An overview of cooperative multi-agent deep reinforcement learning. *Aero Weaponry*, 2022, 29(6): 78 – 88.
(邹启杰, 蒋亚军, 高兵, 等. 协作多智能体深度强化学习研究综述. 航空兵器, 2022, 29(6): 78 – 88.)

[8] FOERSTER J, ASSAEL I A, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 2016, 29: 2137 – 2145.

[9] PENG P, WEN Y, YANG Y, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *ArXiv Preprint*, 2017, arXiv: 1703.10069.

[10] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 2017, 30: 6379 – 6390.

[11] KONDA V R, TSITSIKLIS J N. Onactor-critic algorithms. *SIAM Journal on Control and Optimization*, 2003, 42(4): 1143 – 1166.

[12] SUKHBAATAR S, FERGUS R. Learning multiagent communication with backpropagation. *Advances in Neural Information Processing Systems*, 2016, 29: 2244 – 2252.

[13] MAO H, ZHANG Z, XIAO Z, et al. Learning agent communication under limited bandwidth by message pruning. *Assoc Advancement Artificial Intelligence*, 2020, 34: 5142 – 5149.

[14] WANG R, HE X, YU R, et al. Learning efficient multi-agent communication: An information bottleneck approach. *International Conference on Machine Learning*. New York: PMLR, 2020: 9908 – 9918.

[15] KIM D, MOON S, HOSTALLERO D, et al. Learning to schedule communication in multi-agent reinforcement learning. *In International Conference on Learning Representations (ICLR)*. New Orleans, USA: OpenReview.net, 2019: 1732 – 1746.

[16] KIM W, CHO M, SUNG Y. Message-dropout: An efficient training method for multi-agent deep reinforcement learning. *AAAI Conference on Artificial Intelligence*. Palo Alto, CA, USA: AAAI, 2019: 6079 – 6086.

[17] SINGH A, JAIN T, SUKHBAATAR S. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *International Conference on Learning Representations (ICLR)*. New Orleans, USA: OpenReview.net, 2019: 2173 – 2188.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, 30: 5998 – 6008.

[19] JIANG J, LU Z. Learning attentional communication for multi-agent cooperation. *Advances in Neural Information Processing Systems*, 2018, 31: 7254 – 7264.

[20] DAS A, GERVET T, ROMOFF J, et al. Tarmac: Targeted multi-agent communication. *International Conference on Machine Learning*. San Diego, CA: PMLR, 2019: 1538 – 1546.

[21] PESCE E, MONTANA G. Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Machine Learning*, 2020, 109(9): 1727 – 1747.

[22] NIU Y, PALEJA R R, GOMBOLAY M C. Multi-agent graph-attention communication and teaming. *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. United Kingdom: IFAAMAS, 2021: 964 – 973.

[23] AMATO C, CHOWDHARY G, GERAMIFARD A, et al. Decentralized control of partially observable Markov decision processes. *IEEE Conference on Decision and Control*. Firenze, Italy: IEEE, 2013: 2398 – 2405.

[24] SUTTON R S. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988, 3(1): 9 – 44.

[25] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*, 2014, DOI: 10.3115/v1/D14-1179.

## 作者简介:

**邹启杰** 副教授, 硕士生导师, 目前研究方向为智能驾驶、计算机视觉、智能规划与决策, E-mail: jessie_zou_zou@163.com;

**汤 宇** 硕士研究生, 目前研究方向为多智能体深度强化学习, E-mail: tang_yu_dlu@163.com;

**高 兵** 副教授, 硕士生导师, 目前研究方向为大数据分析、知识图谱, E-mail: 2366180678@qq.com;

**赵锡玲** 硕士研究生, 目前研究方向为分层强化学习, E-mail: 17861403295@163.com;

**张哲婕** 硕士研究生, 目前研究方向为多智能体深度强化学习, E-mail: zhangzhejie@s.dlu.edu.cn.