

工程化超越在 MARL 中的新兴通信以实现可扩展且 样本高效的协同任务分配部分可观测网格中

Brennen A. Hill

Department of Computer Science
University of Wisconsin-Madison
Madison, WI 53706
bahill14@wisc.edu

Mant Koh En Wei

Department of Computer Science
National University of Singapore
Singapore 119077
e0958776@u.nus.edu

Thangavel Jishnuanandh

Department of Computer Science
National University of Singapore
Singapore 119077
jishnuanandh@u.nus.edu

Abstract

我们将学习到的与工程化的通信策略在合作多智能体强化学习 (MARL) 环境中的有效性进行了比较。对于学习方法, 我们介绍了直接通信 (LDC), 其中代理通过神经网络同时生成消息和动作。我们的工程化方法——意图通信, 则采用了一个想象轨迹生成模块 (ITGM) 和一个消息生成网络 (MGN) 来根据预测的未来状态制定消息。这两种策略都在完全可观察和部分可观察条件下的合作任务中进行了成功率评估。研究结果表明, 虽然自生通信是可行的, 但工程化方法在性能和扩展性方面表现出更优效果, 尤其是在环境复杂度增加时尤为明显。

1 介绍

1.1 动机与理由

现实世界中的任务, 从群机器人技术到分布式决策制定, 都要求多个代理在部分可观察性和相互干扰的情况下进行协调 [Vincent, 2024, Anglen, 2024]。在这种情况下, 环境变得非平稳因为每个代理的动作改变了其他代理所观察到的动力学, 违反了经典单代理方法如 Q 学习或 DQN 的核心假设 [Kefan et al., 2024]。

允许智能体共享信息是一个有前景的解决方法。虽然先前的研究已经证明交流可以提升整体表现, 但不同交流方案的相对有效性仍然是一个开放问题 [Ming et al., 2024]。因此我们探讨两个主要问题: (i) 有效的交流协议是否可以通过学习而无需显式设计出现, 以及 (二) 工程化交流策略是否提供更优的表现?

1.2 问题定义

我们考虑一个在网格环境中具有两个自主代理的合作场景，任务是导航至两个不同的目标状态。最优结果要求每个代理占据一个独特的目标位置。代理可以独立行动或通过沟通共享有关其观察或意图的信息，从而避免冲突，例如两个代理都瞄准同一个目标。

2 方法论

2.1 实验设置与环境

2.1.1 环境

为了分离通信策略的影响，我们使用了一个简单的确定性网格世界，该世界具有离散的状态和动作空间。环境由一个包含两个相同代理和两个目标的 $x \times x$ 网格组成，每个占据一个单元格。我们使用了 PettingZoo 库 [Terry et al., 2021]，因为它提供了轻量级且可定制的框架。所有实验均在 Google Colab 上进行，这带来了显著的计算约束。这就需要开发出稳健并且训练需求最少的模型。

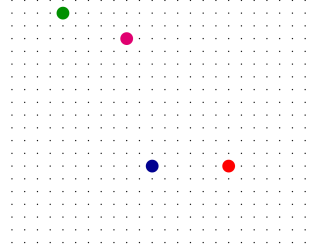


图 1: 实验设置。

2.1.2 观察和行动

观测空间：在完全可观察的设置下，每个代理都会收到两个目标的坐标。代理 i 在时间 t 的观测是一个 4 元组 $o_t^{(i)} = [x_1, y_1, x_2, y_2] \in [0, x)^4$ 。为了提供短期记忆，我们将最近四个帧的观测堆叠起来，为每个代理的策略生成一个 16 维输入向量。

在部分可观察的设置中，智能体仅在其位于指定的视野范围范围内时才能观测到目标的位置。

在每个时间步，智能体选择五个离散动作之一：保持静止、向上、向下、向左或向右。为了防止智能体采用静态的预分配目标策略，在每次试验中随机化一个智能体观察向量中目标呈现的顺序。

2.1.3 初始化和终止

在每个剧集的开始，智能体和目标会在不同的单元格中随机均匀放置。当满足成功条件（两个不同目标上的智能体）或经过最多 200 个周期后，剧集结束。

合作行为通过共享的对称奖励信号得到鼓励：

2.1.4 奖励方案

- +1.0 如果两个代理均占据不同的目标（成功条件）。

- **-0.10** 如果两个智能体都占据了相同目标。
每步
- **-0.01** 处罚以鼓励效率。

2.1.5 训练

我们采用了一种基于 REINFORCE 的在线策略、演员-评论家方法，在每个回合结束时更新策略和价值函数 [Sutton and Barto, 2018, Achiam and OpenAI, 2018, Konda and Tsitsiklis, 1999]。虽然在早期探索阶段固定的学习率是有效的，但它可能在后期微调阶段破坏学习的稳定性。因此，我们应用了一个线性衰减计划：

$$lr_{ep} = lr_0 \left(1 - \frac{episode}{N_{total}}\right),$$

最小值被剪辑为 1×10^{-5} 。此计划最初通过较大的更新促进探索，后来通过较小的更新来保持已建立的策略，从而减轻灾难性的遗忘。

3 学会直接沟通

在我们的学习直接通信 (LDC) 方法中，智能体学习端到端地编码和解码信息。在每个时间步长中，一个智能体会生成一个动作和一条消息。该消息会在后续的时间步长中被另一个智能体接收。

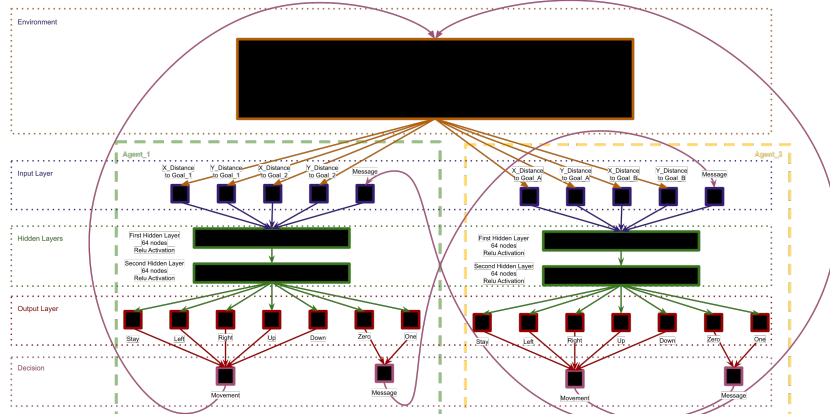


图 2: 两个代理之间的直接通信已学会（批评网络未显示）。

3.1 消息生成

消息由代理的策略网络与其动作一起生成，并直接输入到另一个代理的网络中。因此，通信协议是涌现出来的，在学习过程中没有设计奖励或结构来指导其内容。在我们的实验中，我们使用一个二进制消息空间 $\{0, 1\}$ 。策略网络输出在这个空间上的概率分布。在训练期间，从这个分布中抽样消息；在评估时，则贪婪地选择最有可能的消息。我们的目标是确定代理是否能学会编码有意义的信息，如目标位置或意图目标。

3.2 完全可观察的实验

在完全可观察的设置中，代理学会了有效地导航到不同的目标，这表明它们交换了有意义的信息以协调其选择。

3.2.1 条件概率研究

为了研究学习到的消息的作用，我们分析了代理接收到消息后其行动的条件概率。

表 1: 代理 1 的动作概率取决于代理 2 的消息

Message	动作				
	Stay	Left	Right	Up	Down
0	31.47%	16.98%	19.21%	13.54%	18.80%
1	33.95%	22.86%	14.76%	4.53%	23.90%

表 2: 代理 2 的动作概率取决于代理 1 的消息

Message	动作				
	Stay	Left	Right	Up	Down
0	13.04%	21.61%	38.08%	22.83%	4.44%
1	13.81%	24.55%	29.18%	24.75%	7.71%

如表 1 和表 2 所示，接收代理的动作分布取决于它接收到的消息，表明存在行为相关性。

3.2.2 消融研究

为了测试因果关系，我们进行了一项消融研究，在该研究中消息被替换为一个常数值 (0)，并测量了这对接收者成功率的影响。为了便于清晰比较，我们在本实验中将成功定义为两个代理在 6×6 网格中直接导航至不同目标而没有绕路。

表 3: LDC 性能比较（有通信与无通信，完全可观察）

Condition	Success Rate	Average Steps
With Messages	89.4%	4.39
Messages Ablated (Set to 0)	88.6%	4.43

成功率下降，达到目标的平均步骤增加，如表 3 所示。

3.2.3 全可观测实验结果的讨论

消息与接收代理的行为高度相关。这种关联程度足以让一个代理的消息用于预测另一个代理随后的动作。例如，当接收到 0 (13.54%) 时，代理 1 移动 *Up* 的可能性显著高于接收到 1 (4.53%) 时。这表明每个代理都发展了对另一个代理策略的隐含模型。消息的使用提高了成功率和收敛速度，证实了所学协议传达了协调所需的有效信息。

3.3 部分可观察实验

我们接下来在部分可观测的环境中评估了 LDC，其中视野范围=3 在一个 6×6 网格中。在这种设置下，沟通对于实现近最优策略变得至关重要。

如前所述，成功定义为代理直接移动到各自的目标。成功率低于完全可观察情况下的成功率，主要是因为代理必须首先寻找最初在其视野范围之外的目标。随着足够长的时间范围，代理最终通过找到目标实现接近完美的成功率。如表 4 所示，当消息被消除时，成功率会下降。

表 4: LDC 性能对比（有通信与无通信，部分可观察）

Condition	Success Rate
With Messages	31.89%
Messages Ablated (Set to 0)	30.26%

消息在一种代理拥有另一方缺乏的信息（例如，可以看到目标）的场景中变得尤为重要。因此，通信的影响比在完全可观察设置中更为显著。总结来说，代理学会了显著提高性能的通信协议，并且这种通信的价值在部分可观测性的情况下被放大了。

4 意图沟通

4.1 网络概述

意图共享是明确交换未来导向信息的过程，其中每个智能体广播其预期行动或目标偏好的摘要，使队友能够据此规划 [Qiu et al., 2024]。在合作型多智能体强化学习中，这可以带来更快的协调和更高的任务成功率 [Liu et al., 2021]。

我们设计了一种**意图沟通**架构来实现这一概念。首先，代理通过内部模拟基于当前观察的未来状态短序列生成想象中的轨迹。其次，一个轻量级注意力机制将每条轨迹压缩成一个紧凑的消息以实现高效传输。

4.2 架构组件

在我们的部分可观察设置中，代理很少能看到一个以上的目标。没有沟通，朴素的演员-评论家策略经常会停滞或振荡，因为它们无法推断队友的意图。我们的架构通过两个关键模块解决了这个问题。

想象轨迹生成模块 (ITGM) 该模块在潜在空间 $\tau = \langle z_{t+1}, \dots, z_{t+H} \rangle$ 中预测一个时间范围为 H 的展开过程，条件是基于代理的局部感知和最后接收到的信息 [Liu et al., 2024]。它消耗局部观察 o_t 和之前的消息 m_{t-1} ，并展开学习到的状态转移模型 H 步，为代理提供其下一步操作如何影响未来状态的“心理预览”。

消息生成网络 (MGN) 本模块对想象的轨迹 τ 应用多头自注意力机制，以生成一个固定长度的向量 m_t ，该向量捕捉了智能体计划 [Li et al., 2025] 的本质。此消息 m_t 然后与队友共享。通过交换前瞻性强、信息量大的消息，智能体可以比使用 LDC 或不进行通信更有效地协调。

4.3 积分

前向传递 完整的前向传递过程如下：(1) 堆叠的部分观测与之前的信令被拼接在一起 $[o_t, m_{t-1}]$ 。(2) ITGM 模拟 H 个潜在步骤以生成轨迹 τ 。(3) MGN 跨越 τ 进行注意力机制以创建信令 m_t 。(4) 当前观测与新信令 $[o_t, m_t]$ 拼接并输入共享的 MLP 中。(5) 输出分为策略对数（演员）和状态值估计（评论家）。

训练 整个模型，包括 ITGM、MGN、演员和批评家，使用来自单一 A2C 损失函数的梯度进行端到端训练。这使得框架能够联合学习什么通信和如何根据该通信采取行动。

5 结果

我们将意图交流与 LDC 和一个基本的非通信模型在不断增加规模的部分可观测环境中进行了比较。对于这些实验，智能体被给予最多 200 步来达到独特目标，其中视野范围=2。

表 5: 在不断增加规模的部分可观察环境中的成功率

Setting	Model	Success Rate
10×10	Baseline	0%
10×10	Learned Direct Communication	30.8%
10×10	Intention Communication	99.9%
15×15	Baseline	0%
15×15	Learned Direct Communication	12.2%
15×15	Intention Communication	96.5%

一个不进行通信的基础模型甚至在 10×10 环境中也未能学习到成功的策略，这突显了该任务中通信的必要性。LDC 的性能随着环境规模的扩大而显著下降。相比之下，意图通信在网格尺寸增加时仍保持高性能和鲁棒性。这表明结构化、前瞻性的工程信息使得在更大更复杂的环境中能够进行更有效的协调。值得注意的是，在显著的计算限制（Google Colab）下实现了这些结果，突显了意图通信方法的样本效率。

我们的工作表明，在复杂的协调任务中，经过工程设计的通信模块比单纯依赖自发产生的协议要有效得多。

6 结论

在这项工作中，我们系统地比较了在部分可观测性和计算约束下的合作多智能体强化学习（MARL）中涌现的和工程化的通信策略。我们介绍了 Learned Direct Communication（LDC），一种端到端学习的消息传递协议，以及 Intention Communication，一种通过想象规划模块共享前瞻性轨迹的结构化方法。

我们的结果显示，尽管 LDC 可以在更简单、完全可观察的环境中产生有意义的交流和改进的协调，但它难以在复杂性增加时扩展。相比之下，意图通信在更大且更具挑战性的设置中始终能够实现高成功率。其稳健性和样本效率证明了通过工程模块嵌入归纳偏见的价值。

这些发现表明，新兴的通信能力可能不足以独立完成复杂的协调任务。未来的多代理强化学习系统可能会从结合了学习行为与结构化先验知识的混合方法中获益，以实现灵活性和可扩展性。

Acknowledgments and Disclosure of Funding

我们感谢 Denon Chong Cheng Zong 和 Jeff Lee 对 MARL 通信研究的贡献。我们感谢 Matthew Berland 的指导 and 反馈。

本工作不需要资金支持。

贡献

布伦南山: 项目概念、项目领导、环境开发、基线模型开发、Learned Direct Communication 开发、扩展、优化、多智能体强化学习通信研究、代理训练。

陈国恩威: 意图通信发展, 基线模型开发, 优化, 扩展, 智能体训练。

Thangavel Jishnuanandh: 基线模型开发, 多智能体强化学习通信研究, 优化, 扩展, 智能体训练。

参考文献

Joshua Achiam and OpenAI. Spinning up in deep reinforcement learning, 2018. URL <https://spinningup.openai.com/en/latest/spinningup/spinningup.html>.

Jesse Anglen. The impact of multi-agent reinforcement learning (marl). *Rapid Innovation*, 2024.

Su Kefan, Zhou Siyan, Jiang Jiechuan, Gan Chuang, Wang Xiangjun, and Lu Zongqing. Multi-agent alternate q-learning. *IFAAMAS*, 2024.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Xuesi Li, Shuai Xue, Ziming He, and Haobin Shi. Tmac: a transformer-based partially observable multi-agent communication method. *PeerJ Computer Science*, 2025. doi: 10.7717/peerj-cs.2758. URL <https://peerj.com/articles/cs-2758.pdf>.

Zeyang Liu, Lipeng Wan, Xue Sui, Kewu Sun, and Xuguang Lan. Multi-agent intention sharing via leader-follower forest. *arXiv preprint arXiv:2112.01078*, 2021. doi: 10.48550/arXiv.2112.01078. URL <https://arxiv.org/abs/2112.01078>.

Zeyang Liu, Lipeng Wan, Xinrui Yang, Zhuoran Chen, Xingyu Chen, and Xuguang Lan. Imagine, initialize, and explore: An effective exploration method in multi-agent reinforcement learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*. AAAI Press, 2024.

Yang Ming, Zhao Kaiyan, Wang Yiming, Dong Renzhi, Du Yali, Liu Furui, Zhou Mingliang, and Hou Leong. Team-wise effective communication in multi-agent reinforcement learning. *Springer US*, 2024.

Xihe Qiu, Haoyu Wang, Xiaoyu Tan, Chao Qu, Yujie Xiong, Yuan Cheng, Yinghui Xu, Wei Chu, and Yuan Qi. Towards collaborative intelligence: Propagating intentions and reasoning for multi-agent coordination with large language models. *arXiv preprint arXiv:2407.12532*, 2024. URL <https://arxiv.org/abs/2407.12532>.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018. URL <http://incompleteideas.net/book/the-book-2nd.html>.

J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym

for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 15032–15043, 2021.

Caroline R Vincent. Multi-agent reinforcement learning for autonomous robotics. *DSpace@MIT*, 2024.

A 附录：额外的实验

我们进行了几个额外的实验以更好地理解 LDC 方法的局限性。

A.1 其他沟通风格和环境变化的探索

消息空间： 我们试验了通过扩大值的范围（例如，从 0 到 4 或 0 到 99 的整数）和每次时间步长发送多个值来增加消息容量的方法。

表 6: LDC 在不同消息容量下的收敛结果

Message Range	Message Count	Converged
0-1	1	Yes
0-1	2	No
0-1	4	No
0-4	1	No
0-4	2	No
0-9	1	No

值得注意的是，唯一收敛的策略是使用单一二进制消息（表 6）。我们假设更大的消息空间会增加智能体观察和行动空间的维度，可能会向学习信号中引入显著噪声并阻碍收敛。

环境观测变异： 我们调查了绝对坐标是否比相对坐标更有利于学习。然而，在这种设置下，策略未能与 LDC 收敛。我们怀疑这是因为从绝对坐标中提取方向信息是一项更为复杂的任务，并且各集之间增加的方差阻碍了学习。这一结论适用于完全可观测和部分可观测两种版本。

奖励有意义的信息： 我们试图通过引入发送“有意义”消息的显式奖励来引导通信协议的学习，该奖励基于消息对接收者价值函数估计的影响。然而，这种奖励塑造被证明难以调整：高奖励系数使学习不稳定，而低系数则对性能没有明显影响。

使沟通更加重要： 为了创建一个沟通不可或缺的场景，我们设计了一个环境，在这个环境中每个智能体只能看到分配给其队友的目标。即使在这种设置下，LDC 也未能产生收敛策略。

网络架构变体： 我们探索了各种网络架构。与我们的标准架构（两个隐藏层，每层有 64 个节点）相比，无论是更大的还是更小的架构，通常都会导致训练时间变长或无法收敛。

A.2 附加实验的结论

这些补充实验强调了在没有提供明确架构偏置的情况下，计算约束下促进涌现通信的困难。意图交流方法的成功突显了工程策略的价值，这些策略对通信问题进行结构化处理，从而实现更加稳健和样本高效的习得。