TARMAC: TARGETED MULTI-AGENT COMMUNICATION

Introduction：

作者提出一种部分观测环境多智能体合作的设定。在该设定下，有效的通信协议是关键，作者提出一种通信框架，可以允许定向通信，智能体可以唯一地通过不带有任何监督式通信下的下游特定任务奖励不仅学习发送什么信息，还学习对哪个智能体发送信息。另外，作者介绍了一种多阶段通信方法，在智能体行动之前，通过多回合的通信来协调智能体之间。

作者提出定向通信和多阶段通信的好处。

有效通信对于多智能体合作系统很关键，一种显著的人类通信属性是有针对性的（定向）互动，而不是将所有信息无差别发送给所有参与的智能体；直接将不同的信息发送给需要它的特定智能体是很有用的，在复杂环境中具有灵活性。文中，作者提出有针对性地通信，关键是，每一个智能体都会积极地选择发送信息的智能体。

有针对性的通信技巧是使用基于记号的软注意力机制：与消息一起，发送者广播一个密钥协议，该密钥被接收者用来测量信息的相关性。该种通信机制隐式学习，作为使用下游特定任务的 reward 的端到端训练结果。

由软注意力机制提供的归纳偏差，能使智能体足够（1）通信智能体目标特定信息，比如消防的火（2）适应不同的智能体规模大小（3）被（允许对发送的信息和对象检查的预测性注意力概率解释。通过消息 m 和最后生成的信息集合向量 c。

但作者表示只使用针对性通信没有效果。复杂环境，大规模任务和智能体总是需要多阶段合作和推理，涉及大量信息在内存中持久存在并通过高带宽通信通道交换。

作者提出的模型内部的通信框架每一步都有多阶段的有针对性通信构成，而智能体循环策略由连续相关的内部状态决定。作者使用连续的向量表达，最后能根据不同任务需求学习不同的通信协议。

Related Work：

紧急通信协议：最近的工作都是约束智能体使用带有目标的离散符号通信，以学习紧急语言。而作者是在连续动作下进行的。分散式执行同时带上注意力通信机制，来决定对谁通信，这个谁是决定领域的参数决定的。

Background：

Dec - POMDP：是协作智能体的局部马尔科夫决策拓展。MDP 下，reward 是全局 reward，目标是最大化期望回报。

中心化训练和分散式执行，即 AC 中的 actor 各自学习各自的策略，使用联合动作下的 Q 进行评估，即中心一个 critic。

Algorithm:

假设一组智能体，它们在做协作任务。局部观测是 Ω，有个连续通信信息 m_i，在下一个时间步，从其他智能体接收，以便最大化全局 reward，因为没有智能体能接受环境的完整状态，因此互相通信是一种激励，对于团队来说有帮助。

策略和分散化执行：

对于策略，每一个智能体的策略通过单层门闸循环单元来实现。在每一个时间戳，局部观测 Ω 和一个从其他智能体上一个时间戳得到的信息总和向量 c_i，被用于更新 **GRU** 的隐藏状态 h_i，它编码整个消息-行动-观察历史，直到时间 t。从这种内部状态表达中，智能体

的策略就可以预测出绝对动作分布，同时另一个输出头产生外部信息向量 m_i，注意到每个智能体同构，策略通过共享策略参数 θ ~n 来加速学习。

中心化 critic：

使用内部中间状态 h_i 还有联合动作，来估计 Q 函数。使用普通 TD 算法。

而策略梯度使用如下公式计算：

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}\left[\nabla_{\theta_i} \log \pi_{\theta_i}(a_i^t|h_i^t)\ \hat{Q}_t(h_1^t, ..., h_N^t, a_t^1, ..., a_t^N)\right].$$

注意到使用联合动作下的 Q 能减少方差。

有针对性、多阶段通信：

协作交流策略需要针对性通信。比如：特定的信息传递给特定的智能体。还有多阶段通信。作者使用基于标记的软注意力机制去实现针对性，其中 Actor 输出的 m_i 头有两个部分构成：其中 K 表示标记，是指向智能体参与者；values 是注意力值。

$$m_i^t = [\overbrace{k_i^t}^{\text{signature}} \quad \underbrace{v_i^t}_{\text{value}}].$$

而对于信息的接收端，即 m_i 去生成 c_i，使用每一个智能体从隐藏状态 h 去预测一个序列向量 q，然后点成标记 k，最后通过根号 d 归一化以形成注意力权重。

$$\alpha_j = \text{softmax}\left[\frac{q_j^{t+1^T} k_1^t}{\sqrt{d_k}} \cdots \underbrace{\frac{q_j^{t+1^T} k_i^t}{\sqrt{d_k}}}_{\alpha_{ji}} \cdots \frac{q_j^{t+1^T} k_N^t}{\sqrt{d_k}}\right] \tag{2}$$

$$c_j^{t+1} = \sum_{i=1}^{N} \alpha_{ji} v_i^t. \tag{3}$$

同时注意到公式 2 也含有 α ii 表示自我注意力，经验上发现它可以改善性能。

对于多阶段通信，c 和 h 是首先被用于更新下一个隐藏状态的。

$$h'^t_j = \tanh\left(W_{h \to h'}[\ c_j^{t+1} \parallel h_j^t\ ]\right).$$

Experiments：

	30 × 30, 4 agents, find[red]	50 × 50, 4 agents, find[red]	50 × 50, 4 agents, find[red, red, green, blue]
No communication	95.6%	83.3%	68.1%
No attention	**100.0%**	**89.5%**	82.7%
TarMAC	**100.0%**	**89.6%**	**84.9%**

Table 2: Success rates on 3 different settings of cooperative navigation in the SHAPES environment.

	Easy	Hard
No communication	84.9%	74.1%
CommNets (Sukhbaatar et al., 2016)	99.7%	78.9%
TarMAC 1-stage	**100.0%**	84.6%
TarMAC 2-stage	**100.0%**	**97.1%**

	Success rate
No communication	62.1%
No attention	64.3%
TarMAC	**68.9%**

Conclusion：

作者提出在每一个时间步一种有针对性通信和多阶段协作推理。三种不同环境的评估告诉我们，该方法能从下游具体的任务队伍 reward 中，直观学习注意力和改善行为。

未来挑战就是更大规模智能体数量或者更大规模状态空间。