arXiv:2302.09531v1 [cs.LG] 19 Feb 2023

AIIR-MIX：多智能体强化学习与注意力个体内在奖励混合网络相结合

WeiLi(✉) LI-WEI@SEU.EDU.CN**WeiyanLiu**
LIUWEIYAN@SEU.EDU.CN**ShitongShao**
SHAOSHITONG@SEU.EDU.CN**ShiyiHuang**
HUANGSHIYI@SEU.EDU.CN 东南大学仪器科学与工程学院，中国江苏省南京市，邮编 210096

摘要

推断每个智能体的贡献并为其分配相应奖励，是合作式多智能体强化学习（MARL）中的一个关键问题。以往研究试图通过设计内在奖励函数来解决这一问题，但这些研究只是将内在奖励与环境奖励相加，这使得其多智能体强化学习框架的性能不尽如人意。我们提出了一种名为注意力个体内在奖励混合网络（AIIR-MIX）的多智能体强化学习新方法，其贡献如下：（a）我们基于注意力机制构建了一种新型内在奖励网络，使团队协作更加有效。（b）我们提出了一种混合网络，能够根据环境条件的变化，以非线性、动态的方式结合内在奖励和外在奖励。我们在《星际争霸 II》对战游戏中，将 AIIR-MIX 与多种最先进方法（SOTA）进行了比较。结果表明，AIIR-MIX 表现出色，在平均测试胜率方面能够击败当前先进方法。为验证 AIIR-MIX 的有效性，我们还进行了消融研究。结果表明，AIIR-MIX 能够根据每个智能体的实际贡献，为其动态分配实时内在奖励。

关键词：多智能体强化学习、注意力机制、混合网络、内在奖励。

1. 引言

深度强化学习（DRL）是机器学习的一个重要分支。它利用神经网络近似智能体的最优动作决策或价值函数，实现表示能力的泛化。凭借深度学习强大的拟合能力，DRL 可作为解决复杂环境中智能体决策问题的有效方法，例如群体决策 Nguyen 等人 (2020年）、语音识别 Mousavi 等人 (2016年）；Shen 等人 (2019年）、自动驾驶 Shen 等人 (2019年）、自然语言处理 Young 等人 (2018），以及智能控制 Carlucho 等人 (2020年）。

近年来，多智能体环境成为一个新兴的热门研究主题，其中存在多个智能体进行交互与学习。然而，单智能体强化学习应用于多智能体环境时效果并不理想，因为完全集中式学习所产生的智能体联合动作空间过于庞大，难以学习到最优策略。因此，多智能体强化学习（MARL）是深度

© W. Li(✉)，W. Liu、S. Shao 和 S. Huang。

强化学习从单智能体扩展到多智能体，并形成了一系列解决这一问题的方法。集中式训练与分散式执行（CTDE）Foerster 等人（2016年）；Wang* 等人（2020年），是其中一种有效且应用广泛的方法。在这一范式中，中央控制器在训练期间管理所有智能体的观测、动作和奖励，而在执行期间不使用中央控制器及其价值网络。由于具备上述特征和优势，CTDE 已成为多智能体强化学习中广泛应用的范式，例如 COMA Foerster 等人（2018）、VDN Sunehag 等人（2017年）、QMIX Rashid 等人（2018）和 QTRAN Son 等人（2019年）。在我们的研究中，所提出的方法也将遵循 CTDE 范式。

在多智能体强化学习中，奖励函数极其重要。然而，在一些真实世界任务中，很难将所有情况形式化为奖励函数。在强化学习中探索新环境时，通常只能设置稀疏奖励。然而，在实际应用场景中，稀疏奖励仍面临样本效率低、探索困难等问题。这一问题在多智能体强化学习中尤为明显。由于多智能体强化学习本身存在不稳定环境和维度灾难等挑战，将其扩展到稀疏奖励设置会进一步增加策略学习的难度。在多智能体环境中，解决激励问题的一种有效方法是分配外在奖励，并为每个智能体构建内在奖励。近年来，许多研究者开始关注这些方向。LIIR Du 等人 (2019年)为每个智能体学习一个内在奖励函数，并持续更新该函数，以最大化从环境中获得的预期累积团队奖励。GIIR Wu 等人 (2021)通过使用内在奖励编码器为每个智能体生成独立的内在奖励，解决了懒惰智能体问题。OpenAIBerner 等人 (2019年)使用人工设定的中间奖励来加速学习。FTW Jaderberg 等人 (2019年)通过两层优化来学习智能体的内在奖励。然而，这两种方法都忽略了以下几点：**(a)** 在智能体之间建立依赖关系（即注意力机制）可以产生更精确的奖励；**(b)** 整合内在奖励和外在奖励，以更好地促进策略学习。

在本文中，我们提出了注意力个体内在奖励混合网络（AIIR-MIX）方法，以填补这一空白。AIIR-MIX 包括精确的内在奖励网络（AIIR）的生成模块，以及用于结合内在奖励和外在奖励的非线性混合网络（MIX）。在内在奖励生成方面，我们提出了一种基于注意力机制的内在奖励网络。我们假设每个智能体都有独立的内在奖励。当智能体进行团队协作时，它们的观测和动作极其相似。基于上述假设，利用注意力机制，根据每个智能体的观测和动作计算其在团队协作中的贡献，并为每个智能体生成更准确的内在奖励。通过最大化环境提供的标准累积折扣外在奖励，更新内在奖励网络。在结合内在奖励和外在奖励方面，我们提出了一种能够以非线性方式结合内在奖励和外在奖励的混合网络。将外在奖励输入超网络，以生成混合网络的权重。混合网络结合权重和内在奖励，为每个智能体输出全局奖励。此外，我们将内在奖励函数应用于演员-评论家算法，其中每个智能体的个体策略在相应代理评论家的指导下进行更新，

得益于这些改进，AIIR-MIX 能够生成更恰当的全局奖励，并减少奖励函数设计中的人为干预。

我们在星际争霸 II 微操基准上评估了 AIIR-MIX 方法Samvelyan 等人（2019）。实验结果表明，在同质和异质地图上，AIIR-MIX 的表现均优于 LIIR 和 QMIX 等主流算法。我们进行了消融实验，结果表明，单独使用 AIIR 和 MIX 时，其表现均优于基线算法。此外，我们对训练过程进行了可视化，展示了完整轨迹中注意力权重和内在奖励随时间推移而动态变化的过程。结果证明了基于注意力机制的内在奖励的有效性和重要性。

2. 相关工作

2.1. 稀疏奖励

在许多强化学习（RL）任务中，环境提供的奖励是稀疏的。因此，智能体可能需要经过许多步才能到达获得正奖励的状态。这种情况会导致效率低下、探索困难等诸多问题。

研究人员提出了一种内在动机方法来解决稀疏奖励问题，该方法通过设计内在奖励函数为智能体生成内在奖励，以提高学习效率。Pathak 等人（2017年）提出在自监督方式下，将像素空间中的预测误差作为好奇心奖励，以驱动探索。Strehl 和 Littman（2008）以表格形式记录已访问状态-动作对的次数，并将次数转换为内在奖励，再将其额外加到环境反馈的奖励上。Song 等人（2018）将生成对抗模仿学习扩展到多智能体领域，但对专家演示的需求限制了其通用性。Hao 等人（2019年）将生成对抗模仿学习与自我限制学习相结合，并将其应用于多智能体系统以促进多智能体协作，但当智能体数量较多时，仍未从根本上降低训练难度。

为了设计出能够更高效地解决稀疏奖励问题的内在奖励函数，我们可以利用注意力机制——一种用于建立智能体之间依赖关系的方法——来解决多智能体强化学习中的稀疏奖励问题。

2.2. 注意力机制

注意力机制是一种能够自动选择重要信息的方法，广泛应用于计算机视觉、自然语言处理和强化学习领域。近年来，注意力机制被引入多智能体强化学习（MARL），在一定程度上提升了智能体的学习效果。例如，多智能体注意力演员-评论家（MAAC）Iqbal 和 Sha (2019年)采用注意力机制对集中式评论家网络进行建模。注意力通信（ATOC）Jiang 和 Lu (2018)模型提出了一种带有注意力层的双向 LSTM 通信通道，其注意力机制允许每个智能体根据与其状态相关的重要性，关注来自其他智能体的信息。

受这些工作的启发，我们提出了一种基于注意力机制的内在奖励生成网络。不同于现有方法，本文利用注意力机制为每个智能体生成内在奖励。该网络能够自适应地处理来自其他智能体的历史信息，关注每个智能体在团队协作中的贡献，并生成更精确的内在奖励。同时，将内在奖励网络与演员-评论家架构相结合，以提升其性能。

3. 背景

一般而言，一个完全协作的多智能体问题可以描述为一个去中心化部分可观测马尔可夫决策过程（Dec-POMDP）Oliehoek 和 Amato（2016），由元组 $G =< \mathcal{N}, S, U, P, Z, O, r, \gamma, \rho_0 >$构成。 $\mathcal{N} = \{1, 2, \cdots, n\}$ 表示 n 智能体及 $i \in \mathcal{N}$的集合。$s \in S$ 是环境状态，其中包含所有智能体的全局信息。 U 是动作集。 Z 是智能体的观测集，每个智能体根据观测函数 $O(s, i)$： $S \times \mathcal{N} \rightarrow Z$获得自己的观测 $o_i \in Z$。在每个时间步，每个智能体 $i \in \mathcal{N} \equiv \{1, \cdots, n\}$ 根据当前观测 o_i，通过参数化策略网络 $\pi_i(o_i)$选择一个动作 u_i，形成联合动作 $u_i \in U \equiv U^n$，并根据转移函数$P(s'|s, u)$： $S \times U \times S \rightarrow [0, 1]$转移到下一个状态 s'。 $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_n\}$ 表示由每个智能体的策略组成的联合策略。为了区分不同的奖励，我们将环境提供的团队奖励记为外在奖励 $\mathbf{r}^{\text{ex}}$。将要学习的内在奖励集记为 $\mathbf{r}_t^{\text{in}} = \left\{r_i^{\text{in}}\right\}_{i=1}^n$，其中 t 是时间步的索引。$\mathbf{r}^{\text{ex}}(s, u_i)$： $S \times U \rightarrow \mathbb{R}$ 表示环境为每个智能体 i 提供的团队奖励。 ρ_0： $S \rightarrow \mathbb{R}$ 是初始状态s_0的分布。在完全协作的多智能体问题中，每个智能体都获得相同的 $\mathbf{r}^{ex}$，以促进协作行为。

协作多智能体问题的学习目标是， n 智能体学习一个由 π_i 参数化的策略网络 θ_i，以最大化全局累积折扣奖励集合 $\mathbf{r}_t^{\text{total}}$。也就是说，当 $\mathbf{r}_t^{\text{total}}$最大时，即可获得所有智能体的最优联合策略$\boldsymbol{\pi}^* = \operatorname{argmax}_{\pi} \mathbb{E}_{s_0, u_0, \ldots, s_n, u_n}\left[\sum_{t=0}^{T} \gamma^t \mathbf{r}_t^{\text{ex}}\right]$，其中 $\gamma \in [0, 1)$为折扣因子， T 为最大步数。

3.1. 策略梯度

强化学习的目标是为智能体找到最优行为策略，从而获得最大奖励。策略梯度方法的主要特点是直接对策略进行建模和优化。策略通常被建模为由 θ参数化的函数 $\pi_\theta(a|s)$。在每次训练迭代中，参数 θ 沿梯度 $\nabla_\theta J(\theta)$给出的方向进行调整，以寻找最优 θ^*，与该参数相关的梯度表示为：

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta}, u \sim \pi_\theta}\left[\boldsymbol{\Psi}(s, u) \nabla_\theta \log \pi_\theta(u|s)\right], \tag{1}$$

其中， d^{π_θ} 表示遵循策略的状态转移， $\boldsymbol{\Psi}(s, u)$是与智能体状态和动作相关的轨迹奖励。策略梯度算法被扩展到多智能体领域，每个智能体 $i \in \{1, \cdots, n\}$ 都有一个策略函数 $\pi_{\theta_i}(u_i|o_i)$。多智能体

策略梯度可表示为:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_{\theta_i} \sim \boldsymbol{\pi_\theta}} \left[\sum_{i=1}^{n} \boldsymbol{\Psi}(s, \mathbf{u}) \nabla_{\theta_i} \log \pi_{\theta_i}(u_i | o_i) \right], \tag{2}$$

其中 u_i 表示智能体 i的动作， o_i 表示智能体 i的观测，$\mathbf{u}$ 和 $\boldsymbol{\pi}$ 分别表示所有智能体的联合动作和策略。计算 $\boldsymbol{\Psi}(s, \mathbf{u})$有两种训练方法。一种是通过 REINFORCE 方法训练策略网络（Williams，1992）。另一种是将策略梯度引入 Actor-Critic框架，并通过评论家的梯度训练 Actor。本文选择第二种方法进行训练。优势函数是衡量智能体动作价值的一种常见方法，通常定义为 $A_{\boldsymbol{\pi}}(s, \mathbf{u}) = Q_{\boldsymbol{\pi}}(s, \mathbf{u}) - V_{\boldsymbol{\pi}}(s)$，其中 $Q_{\boldsymbol{\pi}}(s, \mathbf{u})$是$r(s, \mathbf{u}) + \gamma V_{\boldsymbol{\pi}}(s')$，用于帮助计算优势函数。

3.2. QMIX

QMIX Rashid 等人 (2018) 使用混合网络，将所有智能体网络的输出作为输入，并以单调方式进行混合，同时在训练过程中加入全局状态信息，以提升算法性能。在该方法中，混合网络的权重和偏置由各个超网络根据输入状态 s生成。生成的权重通过绝对值激活函数确保非负。每次训练迭代中，QMIX 的目标是最小化损失函数:

$$\mathcal{L}(\theta) = \sum_{i=1}^{b} \left[(y_i^{tot} - Q_{tot}(\boldsymbol{\tau}, \mathbf{u}, s; \theta))^2 \right], \tag{3}$$

其中，$y_i^{tot} = r + \gamma \max_{u'} Q_{tot}(\boldsymbol{\tau}, \mathbf{u}', s'; \bar{\theta}_i)$、 $\bar{\theta}_i$表示目标网络的参数，b表示从经验回放缓冲区采样的转移批量大小， $\boldsymbol{\tau}$ 表示联合动作-观测历史。受这一范式启发，我们采用前馈神经网络作为混合网络，将 $\mathbf{r}_i^{\text{in}}$和 $\mathbf{r}^{\text{ex}}$ 组合成全局奖励 $\mathbf{r}_i^{\text{total}}$，从而提升算法性能。

3.3. 注意力机制

对于注意力机制，输入为一些向量 $\mathcal{V} \equiv \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$。对于每个向量$\mathbf{v}_i$ 属于 $\mathcal{V}$，根据公式 4，注意力机制的输出为 $\widetilde{\mathbf{v}}_i$。

$$\widetilde{\mathbf{v}}_i = \sum_{j=1}^{n} \frac{\mathbf{exp}\left(\sigma\left(v_i, v_j\right)\right) v_j}{\sum_{k=1}^{n} \mathbf{exp}\left(\sigma\left(v_i, v_j\right)\right)}, \tag{4}$$

其中， $\sigma(\cdot, \cdot)$表示相似度度量函数。AIIR-MIX设计了一种注意力信息处理机制，用于计算智能体历史信息的相关性，使其能够自适应地识别和处理有意义的信息，并增强智能体之间的团队协作。

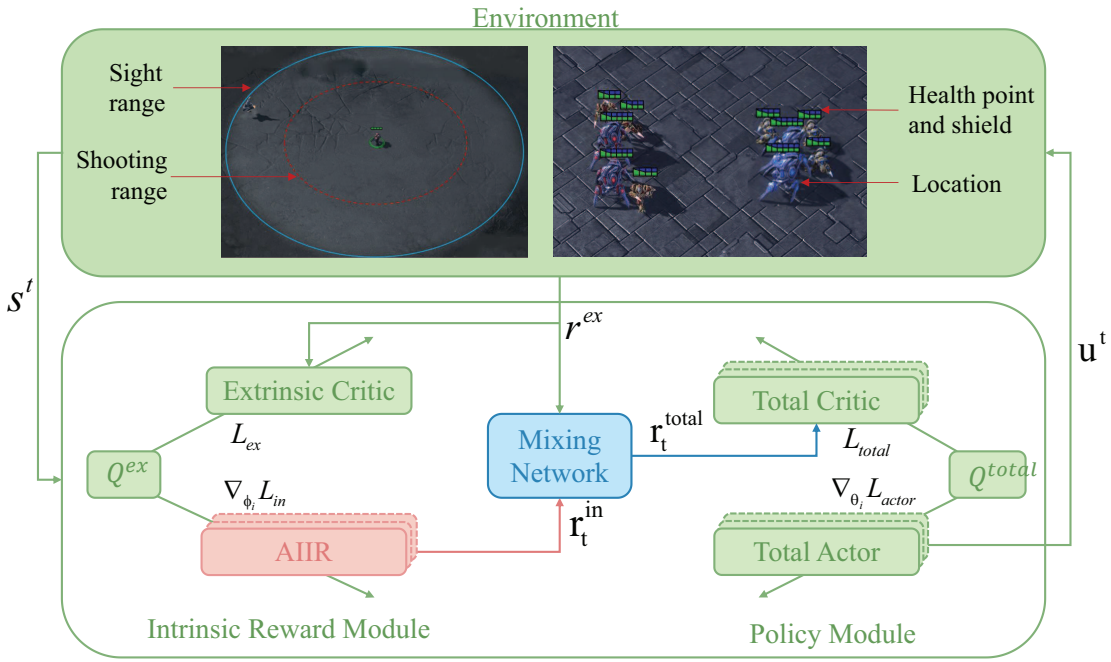


图 1：AIIR-MIX 整体框架的架构。在时间步 t，总 Actor 从环境中获取状态信息（例如健康状况、位置、护盾、智能体视野范围等）并生成动作。随后，从环境中获取 $\mathbf{r}^{\text{ex}}$，AIIR 根据状态信息和动作生成内在奖励集 $\mathbf{r}_t^{\text{in}}$。混合网络通过对 $\mathbf{r}_t^{\text{in}}$ 和 $\mathbf{r}^{\text{ex}}$ 进行非线性运算，得到全局奖励集 $\mathbf{r}_t^{\text{total}}$。

4. 方法

本节介绍一种名为 AIIR-MIX 的新方法，该方法基于 Actor-评论家框架。AIIR-MIX 的主要贡献是生成集合 $\mathbf{r}_t^{\text{in}}$（关于时间步，时间步为 t），并将其与 $\mathbf{r}^{\text{ex}}$ 结合，以非线性方式生成集合 $\mathbf{r}_t^{\text{total}}$。我们首先在图 1中展示 AIIR-MIX 的总体框架，然后在图 2中具体介绍 AIIR-MIX 的相关细节。

如图 1 所示，完整的 AIIR-MIX 框架由五部分组成：外在Critic网络、总体Critic网络、总体Actor网络、AIIR 和混合网络。AIIR 和混合网络构成了一个用于生成 $\mathbf{r}_t^{\text{total}}$的模块，这是 AIIR-MIX 中最重要的部分。具体而言，总体Actor网络由 n个Actor网络组成，每个参数为 θ_i 的Actor网络都会生成其对应智能体的策略。我们将当前状态 s_i 记为每个Actor网络的输入，并将智能体的策略 $\pi_{\theta_i}(u_i|s_i)$记为输出。同样，总体Critic网络中的每个参数为 ω_i的Critic网络都会评估一个智能体的策略，从而能够端到端地更新Actor网络。具体来说，每个Critic网络根据 $Q(s_i, u_i; \omega_i)$更新价值函数参数 ω_i，而每个Actor网络则沿着其对应Critic网络建议的方向，更新 $\pi_{\theta_i}(u_i|s_i)$的策略参数 θ_i。遵循 CTDE范式，总体Critic网络仅在训练期间使用。与 Lowe等（2017）；Foerster 等（2018）类似，总体

Critic网络也通过时序差分误差（TD-error）进行更新。但我们使用由混合网络结合 $\mathbf{r}^{\mathrm{ex}}$ 和 $\mathbf{r}_t^{\mathrm{in}}$生成的 $\mathbf{r}_t^{\mathrm{total}}$，而非 $\mathbf{r}^{\mathrm{ex}}$ 来计算 TD-error。为了更好地促进智能体之间的团队协作，我们基于注意力机制设计了 AIIR，以生成 $\mathbf{r}_t^{\mathrm{in}}$。凭借强大的表征能力，AIIR-MIX 能够更好地更新总体Critic网络和总体Actor网络。

根据贝尔曼方程，评论家网络的损失函数可定义为：

$$\mathcal{L}_{\mathrm{total}}\left(\omega_i\right)=\frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i^{\mathrm{total}}-Q(s_i,u_i;\omega_i)\right)^2\right], \tag{5}$$

其中 $i\in\mathcal{N}\equiv\{1,\cdots,n\}$和 $y_i^{\mathrm{total}}=r_i^{\mathrm{total}}+\gamma Q(s_i^{'},u_i^{'};\omega_i^{'})$。特别值得注意的是，我们将 r_i^{total}定义为智能体 i的全局奖励，它是 $\mathbf{r}_t^{total}$中的一个元素。每个演员网络均通过策略梯度进行更新：

$$\nabla_{\theta_i}J\left(\theta_i\right)=E_{\pi_{\theta_i}}\left[A_{\pi_{\theta_i}}\left(u_i|s_i\right)\nabla_{\theta_i}\log\pi_{\theta_i}\left(u_i|s_i\right)\right], \tag{6}$$

其中 $A_{\pi_{\theta_i}}(u_i|s_i)$ 是基于 $A_{\boldsymbol{\pi}}(s、\mathbf{u})$ 的优势函数（已在第 3.1 节介绍），而 $A_{\pi_{\theta_i}}(u_i|s_i)=r^{ex}(s_i、u_i)+\gamma V\pi_{\theta_i}(s_i)-V\pi_{\theta_i}(s_i^{'})$。演员网络的学习率为$\alpha$ 时，参数 θ_i 的更新可定义为：

$$\theta_{i,t+1}\leftarrow\theta_{i,t}+\alpha A_{\pi_{\theta_i}}\left(u_i|s_i\right)\nabla_{\theta_i}\log\pi_{\theta_i}\left(u_i|s_i\right). \tag{7}$$

参见图 1，所有智能体共享外在Critic网络。该外在Critic网络的参数为 η，其损失函数可定义为：

$$\begin{aligned}&\mathcal{L}_{ex}\left(\eta\right)=\left[y^{ex}-Q(\mathbf{s},\mathbf{u};\eta)\right]^2,\\&\textbf{where }\mathbf{s}\equiv\{s_i\}_{i=1}^n,\mathbf{u}\equiv\{u_i\}_{i=1}^n,y^{ex}=\mathbf{r}^{ex}+\gamma Q(\mathbf{s}',\mathbf{u}';\eta').\end{aligned} \tag{8}$$

接下来，我们将详细介绍 $\mathbf{r}_t^{\mathrm{in}}$ 和 $\mathbf{r}_t^{\mathrm{total}}$ 的生成过程。为了促进智能体之间的合作，并实现最大化 $\mathbf{r}_t^{\mathrm{in}}$ 全局奖励的目标，我们设计了一种基于注意力机制的内在奖励生成框架。

如图2c所示，我们首先定义一个特征提取器，用于处理智能体 s_i^t 的状态 u_i^t 和最后一次动作 i（时间步 t），其中，特征提取器由一系列连续的 ReLU-FC-ReLU-FC 组成。然后，$\mathbf{v}_i^t$ 既是特征提取器的输出，也是注意力机制的输入，表示智能体的局部注意力嵌入。借助注意力机制，我们可以得到所有智能体对之间的相关性。通常，智能体对之间的相关性由距离度量函数计算，并可表示为以下三种形式：

$$\mathrm{A}_{i,j}^t=\sigma\left(\mathbf{v}_i^t,\mathbf{v}_j^t\right),\textbf{where }\sigma\left(\mathbf{a},\mathbf{b}\right)=\begin{cases}<\mathbf{a},\mathbf{b}> & ,\quad\text{dot},\\ \dfrac{<\mathbf{a},\mathbf{b}>}{\parallel\mathbf{a}\parallel_2\cdot\parallel\mathbf{b}\parallel_2} & ,\quad\text{cosine},\\ \mathrm{MLP}\left(\mathbf{a},\mathbf{b}\right) & ,\text{MLP network},\end{cases} \tag{9}$$

由于余弦相似度具有归一化能力，我们使用它来计算智能体注意力嵌入之间的相关性。此外，我们对计算得到的相关性应用Softmax函数

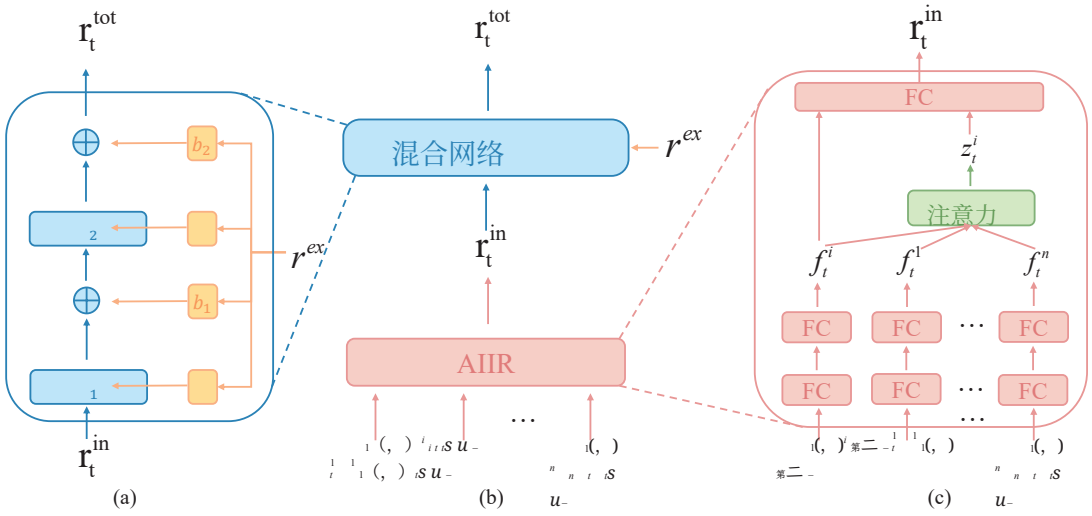


图 2：（a）混合网络结构。（b）生成全局奖励 $\mathbf{r}_t^{\text{total}}$的整体架构。（c）AIIR 结构。

基于计算出的相关性：

$$\widehat{\mathbf{A}}_{i,j}^t = \text{softmax}\left(\mathrm{A}_{i,:}^t\right) = \frac{\exp\left(\mathrm{A}_{i,j}^t\right)}{\sum_k \exp\left(\mathrm{A}_{i,k}^t\right)}. \tag{10}$$

智能体 i 的全局注意力嵌入 $\mathbf{z}_t^i$ 是根据加权系数对所有智能体的注意力嵌入进行加权求和得到的：

$$\mathbf{z}_i^t = \text{attn}\left(\widehat{\mathbf{A}}_{i,:}^t, \mathbf{v}_i^t\right) = \sum_j \widehat{\mathbf{A}}_{i,j}^t \cdot \mathbf{v}_{i,j}^t, \tag{11}$$

其中，$\mathbf{z}_i^t$ 包含智能体状态和动作的相似程度，能够学习智能体之间的相关性并促进团队协作。之后，将智能体 i 的局部注意力嵌入 $\mathbf{v}_i^t$ 和通过注意力机制获得的全局注意力嵌入 $\mathbf{z}_i^t$ 同时输入全连接层。最后，输出智能体 i 的 r_i^{in}以及所有内在奖励 $\mathbf{r}_t^{\text{in}}$的集合：

$$\mathbf{r}_t^{\text{in}} = \left\{r_i^{\text{in}}\right\}_{i=1}^n, \ \textbf{where } r_i^{\text{in}} = \mathbf{FC}(\mathbf{z}_i^t). \tag{12}$$

如图2a所示，$\mathbf{r}_t^{\text{in}}$由 AIIR 输出，并通过混合网络与 r^{ex} 进行非线性组合，输出$\mathbf{r}_t^{\text{total}}$。在以往研究中，LIIRDu 等人(2019年) 和 GIIRWu 等人(2021) 均通过加权求和将$\mathbf{r}_t^{in}$和$\mathbf{r}^{\text{ex}}$ 结合起来：

$$\mathbf{r}_t^{\text{total}} = \mathbf{r}^{\text{ex}} + \lambda \mathbf{r}_t^{\text{in}}. \tag{13}$$

在 AIIR-MIX 中，我们采用非线性方法将 $\mathbf{r}_t^{\text{in}}$和 r^{ex} 结合起来，动态生成 $\mathbf{r}_t^{\text{total}}$，使智能体能够在每个时间步获得更准确的奖励，从而

让智能体更容易选择最优策略。混合网络的权重 $\{\mathcal{W}_1, \mathcal{W}_2\}$和偏置 $\{\mathbf{b}_1,\mathbf{b}_2\}$ 由一个独立的超网络生成，如图2a左侧所示。将 r^{ex} 用作超网络的输入，超网络通过不同的线性层生成相同的权重 $\{\mathcal{W}_1, \mathcal{W}_2\}$ 和偏置 $\{\mathbf{b}_1, \mathbf{b}_2\}$，并分别将其作为混合网络的权重和偏置输出。结合梯度下降法与超网络来更新混合网络，可以动态调整混合网络的权重，从而使其能够结合$\mathbf{r}_t^{\text{in}}$ 和 r^{ex}:

$$\mathbf{r}_t^{total} = \left\{ r_i^{\text{in}} \times \mathcal{W}_1 \times \mathcal{W}_2 + \mathbf{b}_1 \times \mathcal{W}_2 + \mathbf{b}_2 \right\}_{i=1}^{n}. \tag{14}$$

然后，$\mathbf{r}_t^{\text{total}}$由 AIIR-MIX 得出，其工作原理在图 2b 中有更详细的介绍。

5. 实验

本节中，我们在星际争霸多智能体挑战赛（SMAC）环境中评估 AIIR-MIX 方法。该环境由 Samvelyan 等人 (2019年) 提出，现已成为评估多智能体强化学习方法的基准。我们将 AIIR-MIX 与最先进的多智能体强化学习方法进行比较，例如 LIIR Du 等人 (2019年)、QMIX Rashid 等人 (2018)、COMA Foerster 等人 (2018)、QTRANSon 等人 (2019年)。我们通过消融实验展示 AIIR 和混合网络的有效性与合理性。最后，为了更清晰地分析智能体的学习过程，我们对每个时间步的注意力权重和内在奖励进行可视化。

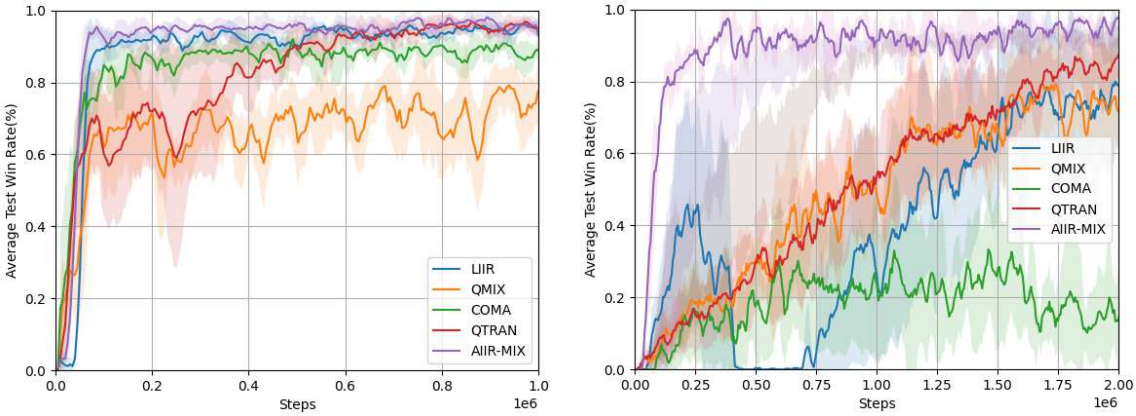
5.1. 星际争霸 II 微操

星际争霸多智能体挑战赛（SMAC）环境是一个基于即时战略游戏《星际争霸 II》的实验环境。与完整的《星际争霸 II》相比，它更侧重于各智能体的微观策略，而非宏观操作；也就是说，SMAC 关注如何控制各智能体击败敌人，而不考虑发展经济、进行资源调度等高层宏观操作。

表 1：不同场景中的地图。

Name	友军单位	敌军单位	Type
2s3z	2 个 Stalker，3 个狂热者	2 个 Stalker，3 个狂热者	异质
3s5z	3名追猎者 5名狂热者	3名追猎者 5名狂热者	异质
8m	8 名陆战队员	8名陆战队员	同质
MMM	1 架医疗运输机 2 名掠夺者 7 名陆战队员	1 架医疗运输机 2 名掠夺者 7 名陆战队员	异质

在实验中，本文采用 SMAC 的所有默认设置，包括游戏难度设置、射程和观测范围。射程和观测范围均为具有一定半径的圆形。只有处于观测范围内的智能体才能进入视野，只有处于射程内的智能体才能受到攻击。如图1所示，在环境中，智能体的属性包括武器冷却时间（CD）、生命值（HP）、护盾（2S3Z 和 3S5Z）、单位类型、相对距离



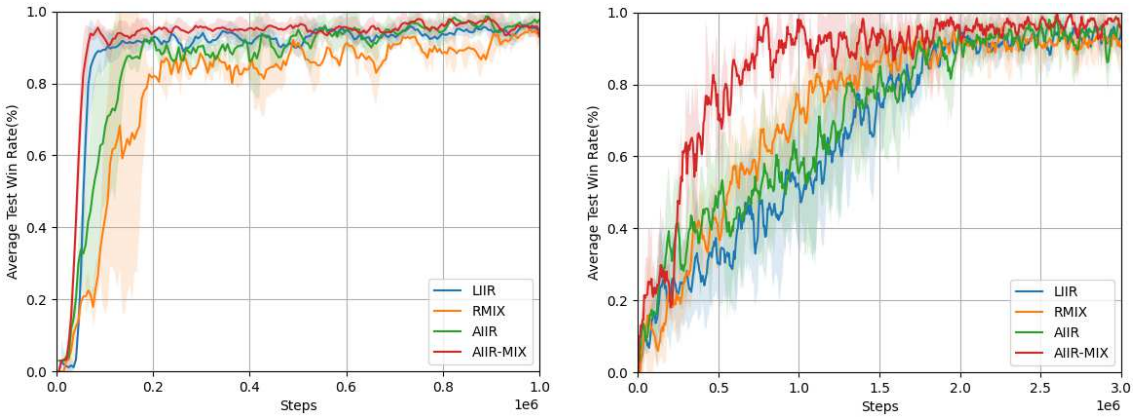
(a) 8m (b) MMM

(c) 2s3z (d) 3s5z

图3：LIIR、QMIX、COMA、QTRAN 和 AIIT-MIX 在同质和异质场景下的平均测试胜率

被观测单位的属性，以及上一个动作。智能体的动作空间由四种离散动作组成：移动[方向], 攻击[敌方单位编号], 停止和无操作。智能体移动空间包含四个方向：东、南、西和北。攻击动作需要指定射程内的敌方单位编号。我们选取四个具有挑战性的对称场景——2s3z、3s5z、MMM 和 8m，来评估算法性能。2s3z、3s5z 和 MMM 是异质地图，而 8m 是同质地图。SMAC 不同地图的场景详情见表1。不同的游戏角色具有不同的生命值、攻击力和射程。

我们使用不同的随机种子进行 5 次独立运行来训练这些方法。每次运行中，每经过 5000 个训练时间步，就在 20 个独立评估回合中评估这些方法。如果在时间限制内击败所有敌军单位，该评估回合便视为获胜回合，随后计算获胜回合所占的百分比，即胜率。



(a) 8m (b) 2s3z

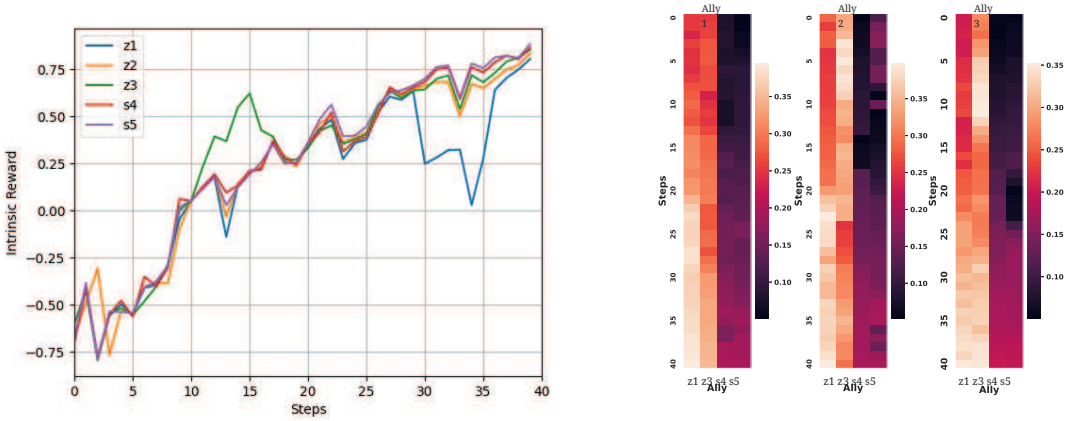
图4：AIIR-MIX 在 8m 和 2s3z 上的平均测试胜率及消融结果。

5.2. 对比

图 3 展示了这些方法在 SMAC 的 4 张不同地图上的表现。在所有基线算法中，QMIX 和 QTRAN 是当前最先进的主流价值分解算法。在 MMM 和 2s3z 中，QMIX 和 QTRAN 均取得了良好的表现。QMIX 在 3s5z 中表现良好。然而，QMIX 和 QTRAN 的收敛速度比其他最先进算法慢得多，而且 QMIX 在 8m 中的表现相对较差。这些结果表明，基于价值分解的 QMIX 和 QTRAN 在一定程度上能够掌握异质场景，但在同质场景中的表现相对较差。COMA 在 MMM、2s3z 和 3s5z 场景中表现不佳，这一结果证实，反事实策略梯度与 TD优势策略梯度在引导演员网络方面存在性能差距。LIIR 能够掌握同质场景，例如 8m；但在 MMM、2s3z 和 3s5z 等异质场景中，其收敛速度较慢且表现不佳。在所有场景中，AIIR-MIX 算法的表现始终优于其他算法。这一结果表明，将注意力机制生成的内在奖励与外在奖励相结合的混合网络，能够显著提升训练策略的表现。

5.3. 消融实验

我们开展了消融实验，以研究注意力机制对 AIIR 的影响，以及混合网络中非线性变换的必要性。一方面，我们通过将 AIIR 与 RMIX（即混合网络）进行比较，分析注意力机制在 AIIR 中的重要性。RMIX 采用与 LIIR 相同的内在奖励网络结构，而不是注意力机制来生成内在奖励。另一方面，我们研究了非线性混合网络的必要性。我们将混合网络替换为线性混合网络，在该网络中，每个智能体的内在奖励和外在奖励仅经过加权求和，以生成全局奖励。



(a) 内在奖励　　（b）注意力权重

图5：AIIR-MIX中的内在奖励和注意力权重。

图 4 展示了 AIIR-MIX 及其在 SMAC基准测试 8m 和 2s3z 上的消融实验结果。可以看出，在实验中，AIIR-MIX 的表现优于其所有消融版本。图 4a 显示，在同质场景中，与基线算法 LIIR 相比，RMIX 和 AIIR 由于架构的复杂性而导致学习速度变慢。然而，当基于注意力的内在奖励函数与非线性混合网络共同应用于 AIIR-MIX 时，这种改进能够弥补复杂架构所导致的训练速度下降。图 4b 显示，在异质场景中，需要使用非线性混合网络才能获得更好的性能。此外，AIIR-MIX 与 RMIX 的性能对比表明，精确的内在奖励以及通过非线性方式与外在奖励相结合的重要性。

5.4. 可视化

除了评估第5.2节中训练得到的策略的性能外，我们更想了解注意力机制对学习内在奖励的贡献，以及学到的内在奖励对策略学习的贡献。为了弄清上述两个过程中学到了什么，我们提出显式地可视化轨迹中每个时间步的注意力权重和内在奖励。为清晰起见，我们选择智能体数量相对较少的2s3z场景进行分析。图5 展示了智能体的注意力权重和内在奖励。图6 展示了智能体的一些辅助快照。上述图中包含智能体类型（s表示Stalker，z表示Zealot）和智能体编号（从1到5）。

从图中可以看出：5a 在时间步 10 之后，友方Zealot 3 的内在奖励大幅增加。图6a 和图6b清楚地说明了这一现象的原因。时间步 10 之前，生命值较低的友方Zealot 3 正在攻击敌方Zealot 3。时间步 10 之后，友方Zealot 2 开始攻击敌人，而友方Zealot 3 停止开火并逃离。当智能体生命值不足时，避免正面攻击敌人显然是良好的行为。时间步 30 之后，友方Zealot 1 的内在奖励大幅下降



(a) t=5 (b) t=15

(c) t=23 (d) t=34

图6：2s3z中的一些辅助快照。

。6d所示，友方Zealot 1 正与友方Zealot 2、4 和 5 一同攻击敌方Zealot 2。然而，此时友方Zealot 1 的生命值非常低，继续攻击敌人并非良好的策略。总体而言，智能体选择良好策略时，内在奖励会增加；选择糟糕策略时，内在奖励则会减少。这种更精确的奖励有助于智能体快速学习良好策略。

如图 5a所示，狂热者类型智能体的内在奖励波动更大。为了更详细地分析，我们将狂热者类型智能体的注意力权重可视化。图 5b展示了三个狂热者对其他友方智能体的注意力权重。显然，狂热者总是更加关注另外两个友方狂热者。注意力权重的升高表明智能体之间出现了协作行为，包括图 6中时间步23时友方Zealot 1对友方Zealot 2的关注，以及时间步5时友方Zealot 2对友方Zealot 3的关注。也就是说，注意力机制在一定程度上促进了智能体之间的协作行为。

6. 结论

本文提出了 AIIR-MIX，这是一种新颖的多智能体强化学习算法，它通过注意力机制为每个智能体学习个体内在奖励，使智能体能够获得不同的奖励，从而促进其学习。此外，我们首次设计了非线性混合网络来结合内在奖励和外在奖励，而非简单地进行线性求和，从而根据环境的变化动态生成每个智能体的全局奖励。在星际争霸 II 对战游戏中进行的实验结果表明，我们的方法能够训练出

与若干最先进的多智能体强化学习方法相比表现更好的策略。我们还进行了消融研究，以验证基于注意力机制的内在奖励网络和非线性混合网络的有效性。此外，我们对内在奖励和注意力权重进行了可视化，以说明内在奖励网络如何为每个智能体分配适当的奖励并促进智能体之间的协作。

致谢

本研究部分得到中国航空科学基金（项目编号：20200058069001）资助，部分得到中央高校基本科研业务费（项目编号：2242021R41094）资助。

参考文献

Christopher Berner、Greg Brockman、Brooke Chan、Vicki Cheung、Przemys law D˛ebiak、Christy Dennison、David Farhi、Quirin Fischer、Shariq Hashme、Chris Hesse 等。《Dota 2》中的大规模深度强化学习。arXiv 预印本 arXiv:1912.06680，2019年。网址：https://doi.org/10.48550/arXiv.1912.06680。

Ignacio Carlucho、Mariano De Paula 和 Gerardo G Acosta。面向移动机器人的多输入多输出 PID 控制的一种自适应深度强化学习方法。《ISA Transactions》，102：280–294，2020年。doi：10.1016/j.isatra.2020.02.017。网址：https://doi.org/10.1016/j.isatra.2020.02.017。

Yali Du、Lei Han、Meng Fang、Ji Liu、Tianhong Dai 和 Dacheng Tao。LIIR：在多智能体强化学习中学习个体内在奖励。神经信息处理系统进展，32，2019年。URLhttps://proceedings.neurips.cc/paper/2019/file/07a9d3fed4c5ea6b17e80258dee231fa-Paper.pdf。

Jakob Foerster、IoannisAlexandros Assael、Nando De Freitas 和 Shimon Whiteson。学习通过深度多智能体强化学习进行通信。神经信息处理系统进展，29，2016年。URLhttps://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf。

Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。载于AAAI 人工智能大会论文集，第32卷，新奥尔良，路易斯安那州，2018年4月。AAAI。URL https://ojs.aaai.org/index.php/AAAI/article/view/11794。

Xiaotian Hao、Weixun Wang、Jianye Hao 和 Yaodong Yang。合作式多智能体系统中的独立生成对抗式自我模仿学习。arXiv 预印本 arXiv:1909.11468，2019年。网址：https://doi.org/10.48550/arXiv.1909.11468。

Shariq Iqbal 和 Fei Sha。面向多智能体强化学习的演员-注意力-评论家。在国际机器学习会议，第 2961–2970 页，美国长滩，2019年6月。PMLR。网址：https://proceedings.mlr.press/v97/iqbal19a.html。

Max Jaderberg、Wojciech M Czarnecki、Iain Dunning、Luke Marris、Guy Lever、Anto- nio Garcia Castaneda、Charles Beattie、Neil C Rabinowitz、Ari S Morcos、Avraham Ru- derman 等。通过基于群体的强化学习在三维多人游戏中实现人类水平表现。*Science*，364(6443):859–865，2019年。doi：10.1126/science.aau6249。网址：https://doi.org/10.1126/science.aau6249。

Jiechuan Jiang 和 Zongqing Lu。学习用于多智能体协作的注意力通信。神经信息处理系统进展，第 31 卷，2018。链接：
https://proceedings.neurips.cc/paper/2018/file/6a8018b3a00b69c008601b8becae392b-Paper.pdf。

Ryan Lowe、Yi I Wu、Aviv Tamar、Jean Harb、OpenAIPieter Abbeel 和 Igor Mordatch。适用于混合合作竞争环境的多智能体演员-评论家。神经信息处理系统进展，第 30 卷，2017年。链接：https://doi.org/10.48550/arXiv.1706.02275。

Seyed Sajad Mousavi、Michael Schukat 和 Enda Howley。深度强化学习：概述。载于 SAI智能系统会议论文集，第 426–440 页，英国伦敦，2016年 9 月。Springer。doi：10.1007/978-3-319-56991-8 32。链接：https://doi.org/10.1007/978-3-319-56991-8_32。

Thanh Thi Nguyen、Ngoc Duy Nguyen 和 Saeid Nahavandi。《多智能体系统的深度强化学习：挑战、解决方案与应用综述》。IEEE 控制论汇刊，50(9):3826–3839，2020年。doi：10.1109/TCYB.2020.2977374。URLhttps://doi.org/10.1109/TCYB.2020.2977374。

Frans A Oliehoek 和 Christopher Amato。《去中心化部分可观测马尔可夫决策过程简明导论》。Springer，2016年。URL https://doi.org/10.1007/978-3-319-28929-8。

Deepak Pathak、Pulkit Agrawal、Alexei A Efros 和 Trevor Darrell。通过自监督预测进行好奇心驱动的探索。载于 国际机器学习会议，第2778–2787页，澳大利亚新南威尔士州悉尼，2017年8月。PMLR。URLhttps://proceedings.mlr.press/v70/pathak17a.html。

Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。Qmix：面向深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议，第 80 卷，第 4295–4304 页，瑞典斯德哥尔摩，2018 年 9 月。PMLR。网址：https://doi.org/10.48550/arXiv.1803.11485。

Mikayel Samvelyan、Tabish Rashid、Christian Schroeder De Witt、Gregory Farquhar、Nantas Nardelli、Tim GJ Rudner、Chia-Man Hung、Philip HS Torr、Jakob Foerster 和 Shimon Whiteson。星际争霸多智能体挑战赛。arXiv 预印本 arXiv:1902.04043，2019年。网址：https://doi.org/10.48550/arXiv.1902.04043。

Yih-Liang Shen、Chao-Yuan Huang、Syu-Siang Wang、Yu Tsao、Hsin-Min Wang 和 Tai-Shih Chi。基于强化学习的语音增强，用于实现稳健的语音识别。载于 ICASSP 2019-2019 IEEE声学、语音

和信号处理国际会议 (ICASSP)，第 6750–6754 页，英国布莱顿，2019 年 5 月。IEEE。doi: 10.1109/ICASSP.2019.8683648。网址：https://doi.org/10.1109/ICASSP.2019.8683648。

Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero 和 Yung Yi。Qtran：通过变换学习因式分解，以实现协作式多智能体强化学习。载于国际机器学习会议，第 97 卷，第 5887–5896 页，美国长滩，2019年 9 月。PMLR。网址：https://doi.org/10.48550/arXiv.1905.05408。

Jiaming Song、Hongyu Ren、Dorsa Sadigh 和 Stefano Ermon。多智能体生成对抗模仿学习。神经信息处理系统进展，第 31 卷，2018年。网址：https://proceedings.neurips.cc/paper/2018/file/240c945bb72980130446fc2b40fbb8e0-Paper.pdf。

Alexander L Strehl 和 Michael L Littman。马尔可夫决策过程的基于模型的区间估计分析。计算机与系统科学杂志，74(8):1309–1331，2008年。doi: 10.1016/j.jcss.2007.08.009。网址：https://doi.org/10.1016/j.jcss.2007.08.009。

Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vinicius Zam- baldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z Leibo、Karl Tuyls 等人。用于合作式多智能体学习的价值分解网络。arXiv 预印本 arXiv:1706.05296，2017年。网址：https://doi.org/10.48550/arXiv.1706.05296。

Tonghan Wang*、Jianhao Wang*、Yi Wu 和 Chongjie Zhang。基于影响力的多智能体探索。载于国际学习表征会议，埃塞俄比亚亚的斯亚贝巴，2020年4月。ICLR。网址：https://openreview.net/forum?id=BJgy96EYvr。

Ronald J Williams。用于连接主义强化学习的简单统计梯度跟踪算法。机器学习，8(3):229–256，1992年。网址：https://doi.org/10.1007/BF00992696。

Haolin Wu、Hui Li、Jianwei Zhang、Zhuang Wang 和 Jianeng Zhang。为合作式多智能体强化学习生成个体内在奖励。《国际先进机器人系统杂志》，18(5):17298814211044946，2021。doi: 10.1177/17298814211044946。网址https://doi.org/10.1177/17298814211044946。

Tom Young、Devamanyu Hazarika、Soujanya Poria 和 Erik Cambria。基于深度学习的自然语言处理近期趋势。*IEEE Computational IntelligenCeMagazine*，13(3):55–75，2018。doi: 10.1109/MCI.2018.2840738。网址https://doi.org/10.1109/MCI.2018.2840738。