

基于动态自选择参数共享的合作多智能体强化学习算法

王涵, 俞扬, 姜远

(计算机软件新技术国家重点实验室(南京大学), 江苏 南京 210023)

摘要: 在多智能体强化学习的研究中, 参数共享作为学习过程中一种信息集中的方式, 可以有效地缓解不稳定性导致的学习低效率。但是, 在实际应用中智能体使用同样的策略往往会带来不利影响。为了解决此类过度共享的问题, 提出了一种新的方法来赋予智能体自动识别可能受益于共享参数的智能体的能力, 并且可以在学习过程中动态地选择共享参数的对象。具体来说, 智能体需要将历史轨迹编码为可表示其潜在意图的隐信息, 并通过与其余智能体隐信息的对比选择共享参数的对象。实验表明, 提出的方法在多智能体系统中不仅可以提高参数共享的效率, 同时保证了策略学习的质量。

关键词: 多智能体系统; 强化学习; 参数共享

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202214

A cooperative multi-agent reinforcement learning algorithm based on dynamic self-selection parameters sharing

WANG Han, YU Yang, JIANG Yuan

State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing 210023, China

Abstract: In multi-agent reinforcement learning, parameter sharing can effectively alleviate the inefficiency of learning caused by non-stationarity. However, maintaining the same policy for all agents during learning may have detrimental effects. To solve this problem, a new approach was introduced to give agents the ability to automatically identify agents that may benefit from parameter sharing and dynamically share parameters with them during learning. Specifically, agents needed to encode empirical trajectories as implicit information that can represent their potential intentions, and selected peers to share parameters by comparing their intentions. Experiments show that the proposed method not only can improve the efficiency of parameter sharing, but also ensure the quality of policy learning in multi-agent system.

Key words: multi-agent system, reinforcement learning, parameter sharing

0 引言

多智能体强化学习 (multi-agent reinforcement learning, MARL) 旨在在共享环境中联合训练多个智能体从而完成给定的任务^[1-4]。近期, 多智能体强化学习领域中应用非常广泛的是中心化学习去中心化执行^[5] (centralized learning decentralized execution, CTDE) 的训练框架。CTDE 框架使智能体

在学习的过程中可以共享经验池中的历史轨迹, 进行必要的信息互补, 因此被广泛应用于多智能体强化学习的算法设计和实现中, 例如: 联合动作学习^[6] (joint action learning)、智能体建模^[7] (agent modeling)、多智能体通信^[8-10] (multi-agent communication) 以及信任分配问题^[11-13] (credit assignment)。

然而, 当多智能体系统 (multi-agent system, MAS) 规模较大时, 将 MARL 算法扩展到这样的

收稿日期: 2021-12-15; 修回日期: 2022-01-14

通信作者: 俞扬, yuy@lamda.nju.edu.cn

基金项目: 国家自然科学基金资助项目 (No.61876077)

Foundation Item: The National Natural Science Foundation of China (No.61876077)

系统中便遇到了困难。为了协同完成任务，智能体需要与其他同伴在部分可观测的前提下无缝地协同适应，共同学习。因此学者们希望智能体能够进行有效的信息集中，以进行必要的信息互补。而最直接的做法就是进行全智能体范围内的策略参数共享^[14]，即完全参数共享，如图1(a)所示。该方法之所以有效，主要是在许多MAS中智能体的观测(observation)和奖赏(reward)的结构是相似的。学者们将这种相似性称为智能体的意图(intent)^[15-17]，它通过智能体与环境交互时的轨迹体现出来。尽管完全参数共享在若干工作中发挥了一定的作用，并且得到了理论工作的支持，但大多数情况下它除了作为一个实现细节被提及，并没有得到广泛的应用。这是因为完全的参数共享也可能对最终智能体策略的收敛有不利影响。

人类在日常生活中可以依据所处环境自然地选择自己的合作伙伴。并且做选择的过程基于对他人意图的了解，而不需要明确地掌握他们的具体行为(例如开车、走路)。受上述人类交互情景的启发，本文提出：智能体在学习的过程中可以同人类一般，即通过互相分享意图来决定共享策略参数的对象，并且智能体的意图可以随着时间的推移而改变。

因此，本文介绍了一种算法——动态参数共享(dynamic parameter sharing, D-PS)算法。其中的动态指的是在学习过程中的不同阶段，智能体都可以依据当下的分析，自由地选择共享参数的对象。选择的过程是观察每个智能体的意图，然后对全体智能体的意图应用无监督聚类算法进行分组，同组的智能体便可以进行参数共享。实验部分提供了对本文提出的D-PS算法的测试，测试结果说明：D-PS算法可以收敛到更高的累计奖赏，同时执行效率也高于基线(baseline)算法。此外，智能体在了解其他智能体意图的基础上进行学习有助于其有意识地改变政策，以影响其他智能体的策略，因此更有助于协同完成任务。

1 合作多智能体强化学习下的参数共享概述

1.1 研究背景

通常，多智能体强化学习过程可以被建模成马尔可夫博弈^[18]，记作 $(\mathcal{N}, \mathcal{S}, \{\mathcal{O}^i\}_{i \in \mathcal{N}}, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, \dots, \{R^i\}_{i \in \mathcal{N}})$ ，其中包括智能体集合 $i \in \mathcal{N} = \{1, \dots, N\}$ 、状态空间 $\mathcal{S}$ 、联合观测空间 $\mathcal{O} = \mathcal{O}^1 \times \dots \times \mathcal{O}^N$ ，以及联合动作空间 $\mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^N$ 。每个智能体 i 仅可得

到关于当前状态 s 的局部观测 $\mathbf{o}^i \in \mathcal{O}^i$ 。转移函数 P 定义了基于当前状态和动作所能到达的下一状态的分布 $\mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ ；而 $R^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ 定义了智能体 i 在 t 时刻接收到的奖赏函数 r_t^i 。算法的学习目标是找到一组使每个智能体均可获得最大累计奖赏 $G^i = \sum_{t=0}^T \gamma^t r_t^i$ 的策略 $\pi = (\pi_1, \dots, \pi_N)$ ，即求出 $\forall_i : \pi_i \in \arg \max_{\pi_i} / E[G^i | \pi_i, \pi_{-i}]$ ，其中 π_{-i} 定义为 $\pi_{-i} = \pi \setminus \{\pi_i\}$ ， γ 是折扣因子， T 是一个执行周期(episode)的步长。

在多智能体强化学习算法的范畴内，学者们通常使用演员评判家(actor-critic, AC)算法求得每个智能体的最优策略。具体来说，智能体的策略损失可以被定义为：

$$y_i = r_t^i + \gamma V(\mathbf{o}_{t+1}^i; \mathbf{v}_i) \quad (1)$$

其对应的值损失函数记作 $\mathcal{L}(\mathbf{v}_i) = \|V(\mathbf{o}_t^i; \mathbf{v}_i) - y_i\|^2$ ，其中 $y_i = r_t^i + \gamma V(\mathbf{o}_{t+1}^i; \mathbf{v}_i)$ 。

1.2 相关工作

近年来，随着强化学习研究领域涌现出许多高效的算法，MARL领域也逐渐受到广大学者的关注。为了解决MAS中的不稳定性带来的学习困难，很多工作借助智能体策略参数共享的技术来对算法进行辅助。尽管当时学者们发现参数共享确实在某些情况中会产生正面效果，但它并没有得到理论工作的支持，除了作为一个实现细节被提及，也没有得到太多关注。

直到最近，在Terry J K等人^[18]的工作中，参数共享首次得到了理论上的支持：作者通过将多智能体的学习过程建模为多智能体信息化学习过程(multi-agent informational learning process, MAILP)，从而研究学习过程中的信息传递。该参考文献给出了在学习过程中缓解MARL中的不稳定性的收敛性分析，并提供了一系列实验支撑。同时，Christianos F等人^[19]认为，在学习之前人们可以分析智能体的能力和目标，从而对其进行分组，同组的智能体可以从共享参数中受益，如图1(b)所示。但分组一旦确定，智能体在后续的学习过程中需要维持这个分组不变。这就导致了智能体在学习过程中不够灵活，也就导致了学习的低效性。本文认识到这种分组可以而且应该动态地改变，因此提出了允许智能体在学习的过程中实时地选择共享参数的对象的动态自选择的参数共享方式。

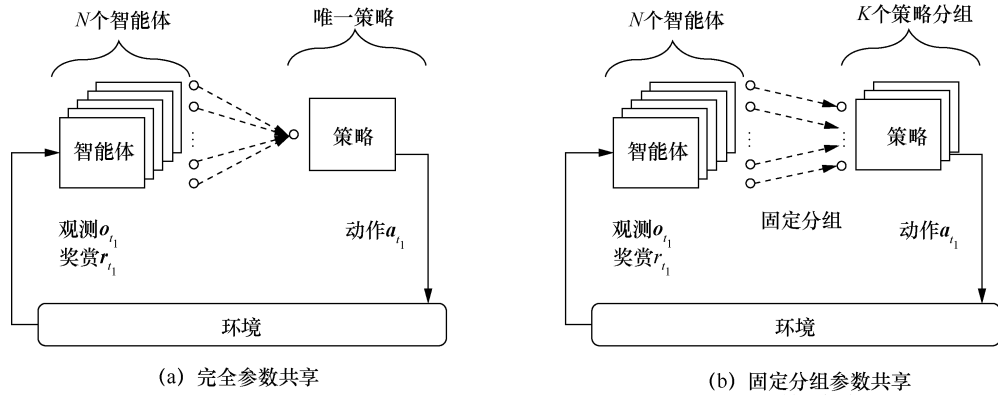


图 1 现有的参数共享工作框架

MARL 的发展在一定程度上得益于深度学习领域的突破。在 MARL 领域的研究工作中不难发现,对智能体的行为和目标进行分析的主要手段就是通过神经网络的表示能力对智能体的轨迹等可观测信息进行抽象,获取高层次的隐特征。在 MAS 中,智能体可以通过对其轨迹的抽象总结出自己的意图,或者推测其他智能体的意图。一些学者提出了基于意图分析的多智能体通信框架^[20-22],该框架允许智能体将意图传送给其他智能体,从而在意图传播的过程中完成多智能体的行为协同。在本文中,智能体的意图(抽象特征)可以借助于神经网络的表示能力被捕获,并且本文允许智能体在学习过程中自主地分析自身以及其他智能体的意图,从而进行共享参数对象的选择。

2 动态自选择的参数共享

为了提高参数共享的有效性,本文提出在学习过程中赋予智能体选择参数共享对象的自由度,并且在训练期间允许智能体随时调整分组。在 MAS 中,本文假设 N 个智能体可以划分为 K 个策略分组,分组 k 中的每个智能体使用并更新共享策略 π_k 。一

旦确定了分组,便确定了每个智能体 i 到策略 π_k 的映射。共享参数的智能体可以从其策略网络的共享表示中获益。动态可选择共享对象的 D-PS 算法结构如图 2 所示。

2.1 从智能体轨迹中提取潜意图

在进行分组前,智能体需要学习如何从本地的经验轨迹中抽象出自身意图。具体来说,智能体 i 的经验记作 τ^i ,它是一系列局部状态、行为和奖励的轨迹。智能体 i 在与环境第 k 次交互后,利用本地经验 τ_k^i 来提取其意图 z_{k+1}^i 。意图 z_{k+1}^i 会指导智能体下一次与环境交互中表现出的行为,即影响策略 $\pi_{\theta^i}(a^i | o^i, z_{k+1}^i)$ 。为此,可以定义一个编码器 E_{ϕ} ,以 ϕ 为参数。编码器以历史经验 τ_k^i 为输入,输出潜在意图 z_{k+1}^i ,编码器结构如图 3 (a) 所示。

既然智能体的潜在意图能够影响其后续的行为,可以使用解码器 D_{ψ} 来解码潜在意图 z_{k+1}^i ,并重构下一次智能体与环境交互时的状态转移和奖励函数。因此,给定智能体 i 一系列轨迹 $\tau_{1:M}^i$ 后,可以通过优化最大似然目标来训练编码器 E_{ϕ} 和解码器 D_{ψ} :

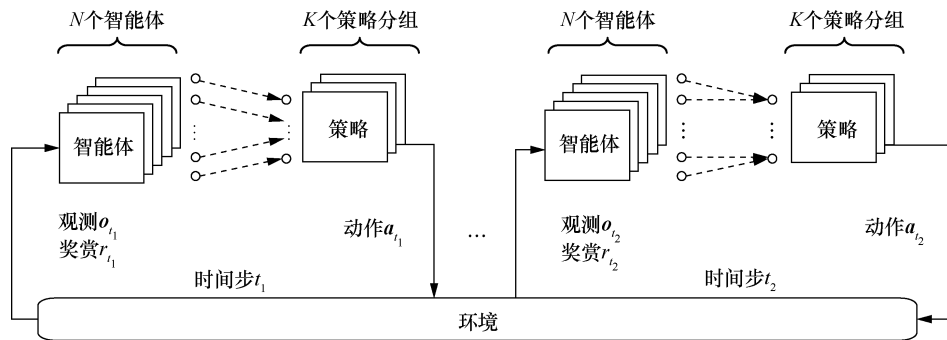


图 2 动态可选择共享对象的 D-PS 算法结构

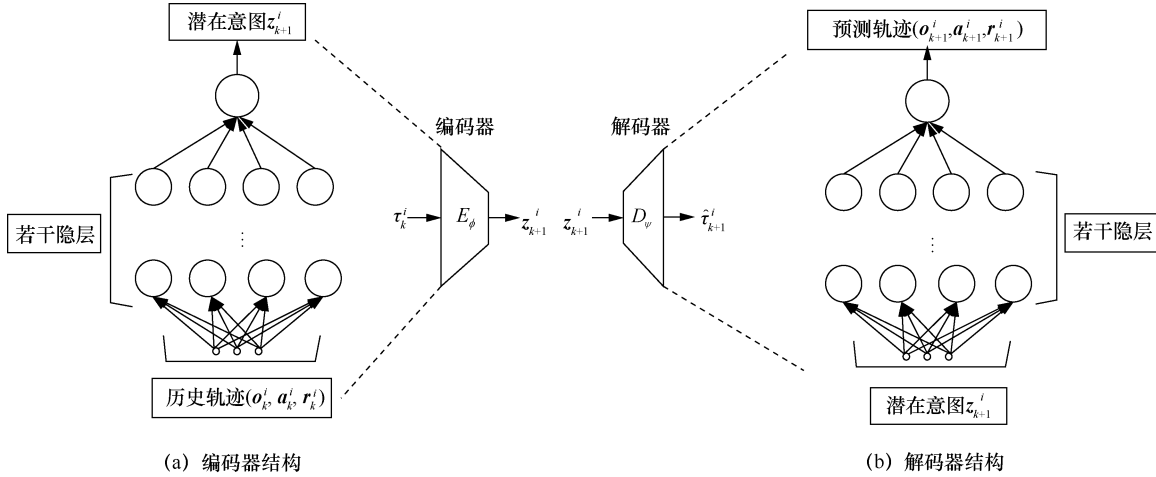


图3 智能体的编码器、解码器结构

$$\begin{cases} \mathcal{J}_{\text{vae}} = \sum_{k=2}^M \sum_{t=1}^H \log p_{\phi, \psi}(\mathbf{o}_{k,t+1}^i, \mathbf{r}_{k,t}^i | \mathbf{o}_{k,t}^i, \mathbf{a}_{k,t}^i, \tau_{k-1}^i) \\ \max_{\phi, \psi} \mathcal{J}_{\text{vae}} \end{cases} \quad (2)$$

由于编码器生成的意图会指导智能体的行为，因此在更新编码器时，不仅需要最大化似然目标，还要智能体的 critic 损失最小化，即：

$$\phi = \phi - \alpha_{\phi} \nabla_{\phi} (\mathcal{J}_Q - \mathcal{J}_{\text{vae}}) \quad (3)$$

而更新解码器时，只需要：

$$\psi = \psi + \alpha_{\psi} \nabla_{\psi} \mathcal{J}_{\text{vae}} \quad (4)$$

最后，利用解码器 D_{ψ} 为智能体生成潜意图，通过对这些潜意图进行相似性对比从而进行分组，相似性高的智能体可以被分为一组。简单起见，笔者在实验过程中通过 K -means 进行聚类以确定分组，分组完成后，同组智能体便可以进行参数共享。

2.2 利用潜意图进行多智能体强化学习

在预知了自身的潜意图后，智能体可以灵活地分析环境中其余智能体的意图并做出反应。例如，在进行巡航的多智能体系统中，距离较近的智能体可以共享策略，加强交互以增进局部区域的探索效率。事实上，智能体在对潜意图做出反应的同时，也在主动影响着这个意图，以在重复的交互中最大化其累计奖赏：

$$\mathcal{J}(\theta^i) = \max_{\theta^i} \sum_{k=1}^M \gamma^k E_{\rho_{\pi_{\theta^i}}} \left[\sum_{t=1}^H R(\mathbf{o}_{k,t}^i, \mathbf{z}_k^i) \right] \quad (5)$$

从而可以获得更高累计奖赏的轨迹。因此，分组的更新可以伴随着策略的更新同时进行，潜意图影响指导智能体的策略学习如图4所示。

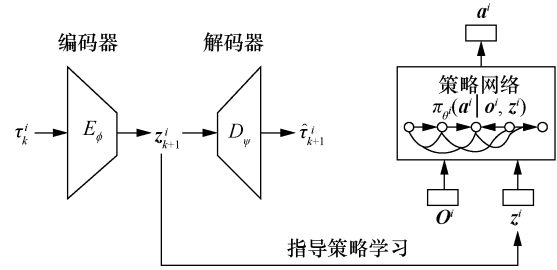


图4 潜意图影响指导智能体的策略学习

本文以多智能体深度确定性策略梯度 (multi-agent deep deterministic policy gradient, MADDPG) 算法为基础，在分组后，每个分组内的智能体 i 通过对式 (6) 求梯度来更新 critic 网络：

$$\begin{cases} \mathcal{J}_Q = E_{\mathbf{o}, \mathbf{a}, \mathbf{r}, \mathbf{o}'} [(Q^i(\mathbf{o}, \mathbf{a}_1, \dots, \mathbf{a}_N) - y)^2] \\ y = r_i + \gamma \bar{Q}^i(\mathbf{o}', \mathbf{a}_1', \dots, \mathbf{a}_N') |_{\mathbf{a}_j' = \pi_j(\mathbf{o}_j)} \end{cases} \quad (6)$$

对于 actor 网络，可以依据式 (7) 进行梯度更新：

$$\nabla_{\theta^i} \mathcal{J}(\theta^i) = E_{\mathbf{o}, \mathbf{a} \sim \mathcal{B}} [\nabla_{\theta^i} \pi_i(\mathbf{a}_i, \mathbf{o}_i) \nabla_{\mathbf{a}_i} Q^i(\mathbf{o}, \mathbf{a}_1, \dots, \mathbf{a}_N) |_{\mathbf{a}_i = \pi_i(\mathbf{o}_i)}] \quad (7)$$

动态参数共享的多智能体强化学习算法伪代码见算法1。

算法1 动态参数共享的多智能体强化学习算法

- 1: 初始化学习速率 α_Q 、 α_{π} 、 α_{ϕ} 以及 α_{ψ}
- 2: 随机初始化参数 θ_Q 、 θ_{π} 、 θ_{ϕ} 以及 θ_{ψ}
- 3: 初始化空经验池 $\mathcal{B}$
- 4: 令全体智能体的意图特征 $\mathbf{Z}^0 = \{\mathbf{z}_1^0, \dots, \mathbf{z}_N^0\}$ 为零向量，即 $\mathbf{z}_i^0 \leftarrow \vec{0}$
- 5: for 训练次数 $i = 1, 2, \dots$ do
- 6: 利用策略 $\pi_{\theta} = \{\pi_{\theta}^1, \dots, \pi_{\theta}^N\}$ 收集交互轨迹 $\tau^i = \{\tau_1^i, \dots, \tau_N^i\}$

```

7:   更新经验池  $\mathcal{B}[i] \leftarrow \tau^i$ 
8:   for  $N$  个智能体  $j = 1, 2, \dots, N$  do
9:     从经验池中采集若干连续序列对
        $(\tau, \tau') \sim \mathcal{B}$ 
10:    更新编码器  $\phi = \phi - \alpha_\phi \nabla_\phi (\mathcal{J}_Q - \mathcal{J}_{\text{vae}})$ 
11:    更新解码器  $\psi = \psi + \alpha_\psi \nabla_\psi \mathcal{J}_{\text{vae}}$ 
12:    更新意图特征  $\mathbf{z}_j \leftarrow \varepsilon_\phi(\tau_j^i)$ 
13:  end for
14:  用  $K$ -means 对意图特征将智能体分为  $K$  组
15:  if 初次进行分组 then
16:    同组智能体随机选择一个智能体的参数
    共享
17:  else
18:    新入组的智能体共享组内智能体的策略
    参数
19:  end if
20:  for  $K$  个分组  $g = 1, 2, \dots, K$  do
21:    for 分组中的每个智能体 do
22:      依据已采得的轨迹组  $\tau$  更新其 critic
      网络
23:      依据已采得的轨迹组  $\tau$  更新其 actor
      网络
24:    end for
25:  end for
26: end for

```

3 实验结果

3.1 实验环境简介

本文在 3 个合作多智能体环境中进行了实验，以说明 D-PS 算法在不同场景中的适用性。实验环境包括《星际争霸》游戏、多机器人仓库、追及 (pursuit) 问题以及粒子群体扩散环境，如图 5 所示。

为了保证算法的稳定性，本文对分组的更新进行了限定，智能体每次更新分组时，新增和删除的

同伴数之和不得超过 2 (最多进行 1 个智能体的替换)：例如某次更新后智能体 1、4 和 5 被分为一组，下次更新后聚类产生的分组是智能体 1、2 和 6；但如果直接将智能体 4 和 5 替换，则替换的智能体数量超过了限制，因此需要进行相似度 (欧氏距离) 的比较，替换当前潜意图最不一致的同伴。

《星际争霸》游戏^[22]是一个被广泛用于 MARL 算法测试的环境。在《星际争霸》游戏场景中，不同类型的多个智能体需要合作完成任务。例如，MMM2 (medivac marauders marines) 环境包含 3 种具有不同属性的智能体 (7 个“陆战队员”、2 个“掠夺者”和 1 个“医疗兵”)，它们根据属性差异被自然地分成了 3 组。因此，MMM2 环境可以作为算法分组有效性的判断依据。

在追及问题中，30 个蓝色躲避者 (evader) 和 8 个红色追赶者 (pursuer)，在边长为 16 的正方形区域活动。躲避者随机移动，追赶者被算法控制。每当追赶者完全包围一个躲避者时，周围的每个追赶者都会得到 5 分的奖励，而躲避者则会从环境中消失。追赶者每次接触躲避者都会得到 0.01 分的奖励。追赶者的动作空间是离散的，可以选择向上、向下、向左、向右以及静止不动。每个追赶者的观测空间是一个以自己为中心的边长为 7 的网格。一轮追及的步长是 500 步。

多机器人仓库中的多个机器人具有不同的功能 (被标注为不同的颜色)。机器人可以旋转、向前移动、拿起或放下架子。机器人的观测空间是一个以自己为中心的边长为 3 的正方形区域。机器人获得奖赏的前提是到达正确颜色的货架。因为该场景中智能体功能不同并且奖赏稀疏，因此学者们普遍认为多机器人仓库是一个比较困难的任务。

粒子群体扩散环境是基于多智能体粒子环境 (multi-agent particle environment, MPE) 创建的盲粒子扩散。由不同颜色的地标和许多被分配了颜色的粒子群体组成。每个粒子看不到自己或者其他粒子的颜

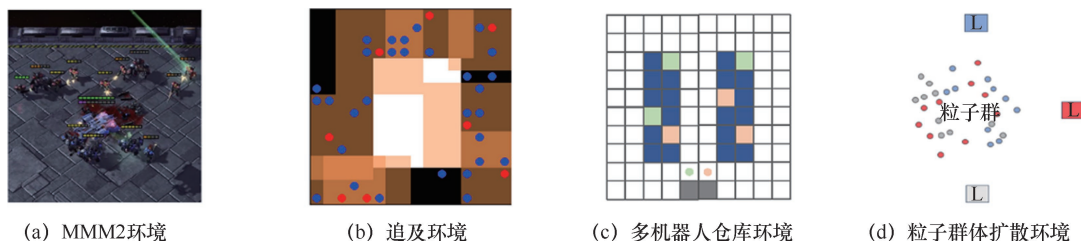


图5 实验环境展示

色,但需要移动到正确颜色的地标上。这种环境中有两个重要变量:智能体的数量和地标颜色的数量。

3.2 对比算法简介

本文将 D-PS 算法与下面描述的其他几种参数共享方法进行比较。

① 无参数共享 (no parameter sharing, N-PS) 算法:所有智能体都有自己的一组参数,并且梯度之间不存在重叠,例如 MADDPG 算法。

② 完全参数共享 (full parameter sharing, F-PS) 算法:所有智能体共享同一组策略参数,并且在学习过程中保持这种设定不变。

③ 预选择参数共享 (selective parameter sharing, S-PS) 算法:在学习之前,通过在环境中预采样得到的轨迹对智能体进行能力和目标分析,对其进行分组,然后在学习的过程中保持这样的分组不变。

3.3 实验结果

本节主要对算法的运行性能进行分析,主要包括算法效果和分组变化展示。算法效果主要讨论算法在学习过程中的表现;分组变化展示则提供了算法在学习过程中指导智能体进行动态分组的依据。

3.3.1 算法学习曲线展示

从图 5 可以看出,本文提出的 D-PS 算法在 3 种

场景中均呈现出预期的表现。具体来说:首先,在图 5 (a) 中的 MMM2 环境中进行聚类时, K -means 算法的超参数设为 3。MMM2 环境中 3 种智能体的观测、动作空间有显著差异。图 6 (a) 中的 S-PS 算法在学习前通过充分的采样已经得到了不同种类智能体的特有信息,在训练开始前就比较准确地确定了共享参数的智能体分组,因此 S-PS 算法有较好的初始表现。而本文提出的 D-PS 算法以及 N-PS 算法、F-PS 算法均没有经过预先采样的步骤,因此在学习的初始阶段没有表现出较好的性能。但是随着学习过程的推进,动态的参数共享体现出了优势:如图 6 (a) 所示,在训练不到 100 轮的时候, D-PS 算法就已经取得了相比 S-PS 算法更高的累计奖赏。而图 6 (a) 中 D-PS 算法在最优解附近呈现出的抖动主要来自更新自编码器带来的误差。

追及问题中的 8 个追赶者是同质智能体,它们的观测空间和动作空间是一致的。因此图 6 (b) 中的 F-PS 算法相较于 N-PS 算法就体现出了明显的优势。智能体的分组更多依赖于它们当前所处的位置以及躲避者的位置。而环境中的躲避者也是运动的,因此,本文提出的 D-PS 算法相较于预先确定分组的 S-PS 算法能得到更高的累计奖赏,并且训

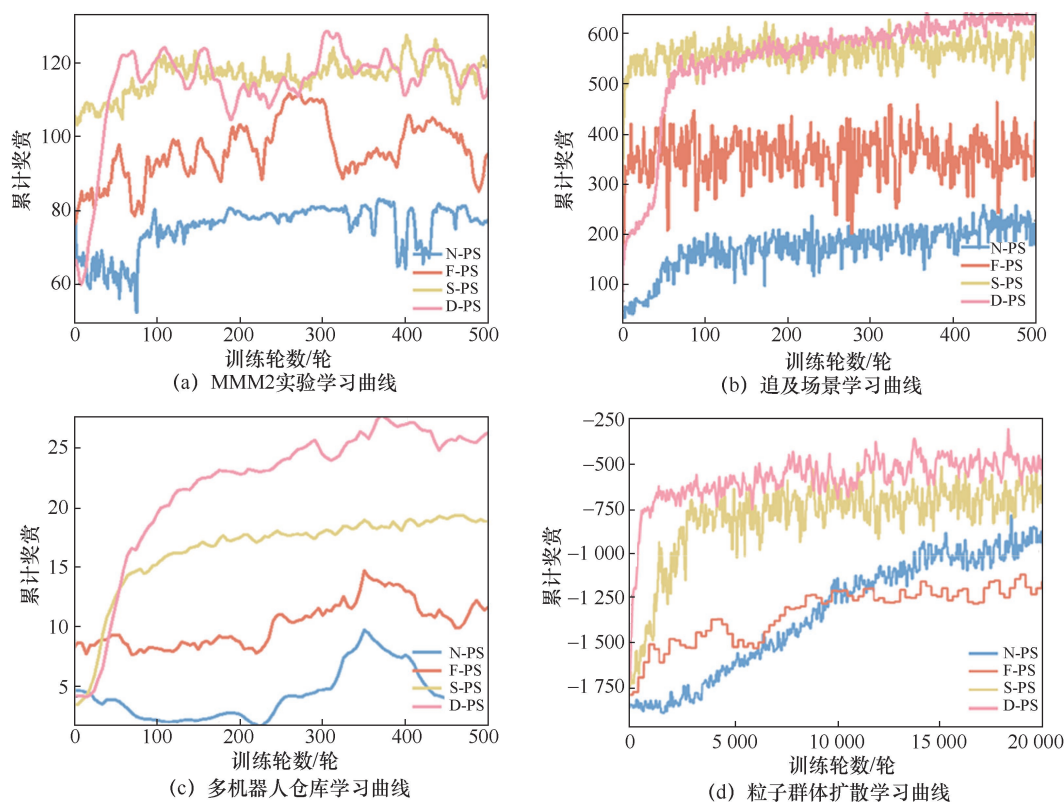


图 6 4 种实验场景中的算法学习曲线

练前期(约200轮)就呈现出更稳定的表现。

在多机器人仓库环境中,本文假设环境中有9个智能体,每3个智能体共享同一个颜色,对应同一个任务。与MMM2不同的是,由于智能体的观测和动作空间完全一致,智能体更需要准确提取自身的意图,以保证分组的有效性。通过图6(c)不难发现,N-PS算法仅依靠共享经验池进行中心化的学习难以奏效;F-PS算法要求功能不同的智能体共享同一策略,没有为学习过程带来实质性的收益;S-PS算法虽然能够表现出相比N-PS算法和F-PS算法更有效的性能,但是由于它在学习过程中不能及时更新分组,因此收敛到了局部最优解。而本文提出的D-PS算法则通过动态分组进行了分组纠正,使智能体突破了局部最优解的限制。

对于粒子群体扩散环境,本文首先将地标颜色的数量设为3,每种地标颜色对应10个智能体,因此环境中有30个智能体。如果所有智能体共享一个策略,它们将无法感知自己对应的颜色,只能学会移动到一个局部最优的地点。由于环境中智能体数量较多,如果不进行有效的参数共享,会使得训练效率大大降低,如图6(d)中的N-PS算法,在其他方法均稳定时,N-PS算法依然没有收敛。F-PS算法由于预分组的指导,在学习初期能够很快进行收敛,但是无法及时对错误分组的智能体进行调整,因此后期呈现出较大的波动。而D-PS算法在进行动态分组学习的过程中既可以保持较快的收敛速度又可以维持较稳定的表现。

3.3.2 分组变化展示

笔者将MMM2和多机器人仓库实验作为代表性场景进行D-PS算法在学习过程中的若干时刻分组情况的展示。

首先在MMM2实验中,如图7所示,笔者发现S-PS算法通过预采样分组后,已经准确得到了与智能体角色相吻合的分组,而D-PS算法则需要在学习过程中随着策略优化的过程进行分组更新。

在多机器人仓库场景中,智能体通过有效的共享参数,互补各自收到的奖赏信息,为优化策略提供有效信息。但因为智能体的任务不同(对应不同颜色),智能体更需要准确把握自身的潜在意图以产生有效分组。多机器人仓库任务中S-PS算法和D-PS算法的分组情况如图8所示。从图8可以得出,S-PS算法虽然在预采样阶段获取了大量样本,但是依旧没有做出充分的分组判断,并且在智能体学习的过程中该分组不能实时更新,从而导致了智能体的策略优化方向走向了局部最优。而D-PS算法则规避了固定分组产生的不良后果,通过对策略、编码器和解码器的及时更新保证了分组的有效性。

4 结束语

本文主要针对MARL中现有的参数共享进行讨论和分析,提出了一种动态的可选择的参数共享方法。“可选择性”为智能体提供了选择参数共享对象的自由度,而“动态性”赋予智能体在学习过程中调整参数共享对象的灵活性。本文的实验表

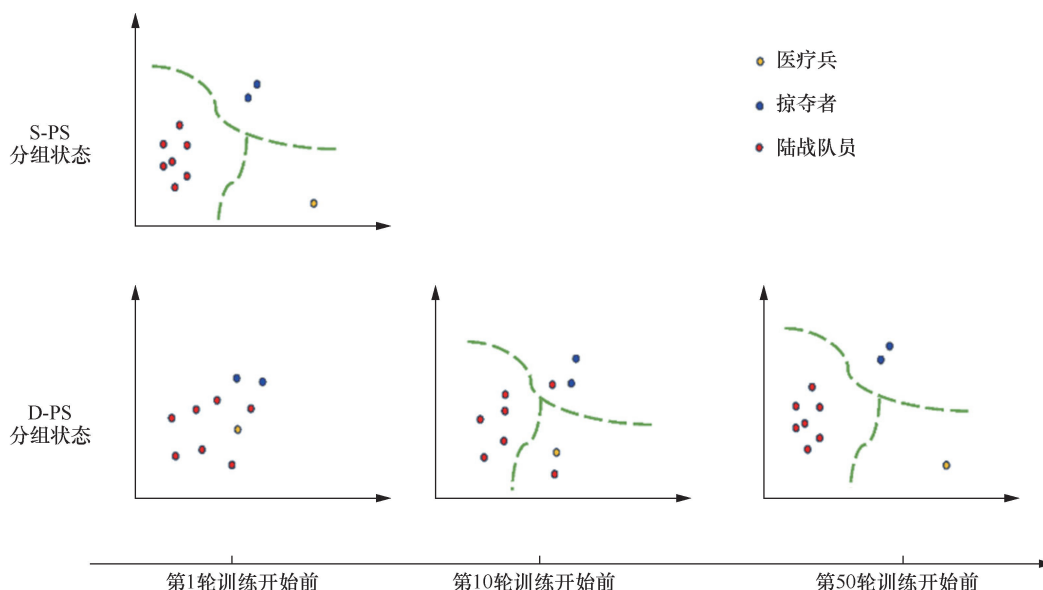


图7 MMM2任务中S-PS算法和D-PS算法的分组情况

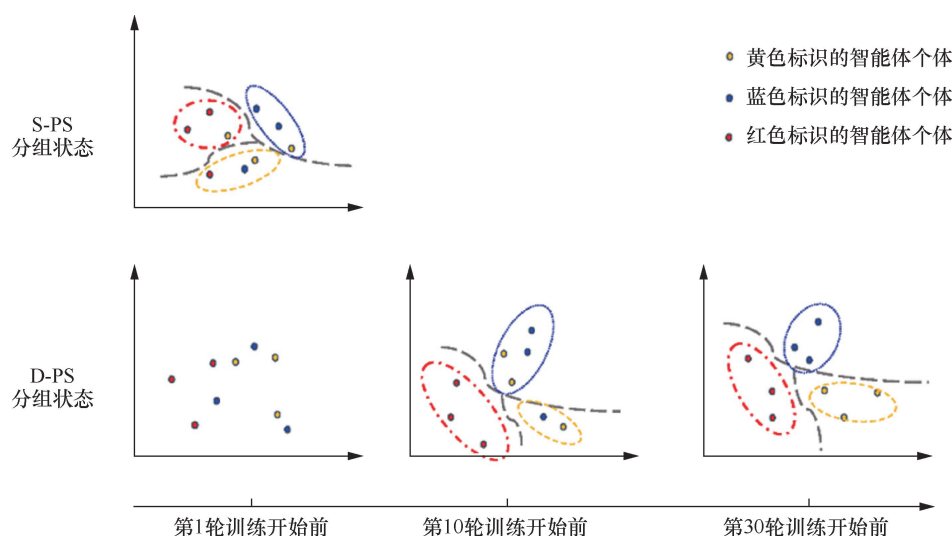


图8 多机器人仓库任务中 S-PS 算法和 D-PS 算法的分组情况

明,不加区别地在全体智能体范围内共享参数可能会使学习变得更加困难,加剧不稳定性的不良后果;而预先确定分组虽然在一定程度上提高了学习性能,但是无法保证智能体在学习过程能够保持高效的策略更新。而本文提出的方法可以帮助智能体在学习过程中实时地识别异质智能体,并将它们分配到不同的分组。结果表明,该方法在累计奖赏收敛速度方面优于其他参数共享方法,并获得了更高的累计奖赏值。

参考文献:

- [1] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.
- [2] 王飞跃, 曹东璞, 魏庆来. 强化学习: 迈向知行合一的智能机制与算法[J]. 智能科学与技术学报, 2020, 2(2): 101-106.
WANG F Y, CAO D P, WEI Q L. Reinforcement learning: toward action-knowledge merged intelligent mechanisms and algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(2): 101-106.
- [3] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 314-326.
LIU Z Y, MU C X, SUN C Y. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 314-326.
- [4] 王金予, 魏欣然, 石文磊, 等. 强化学习在资源优化领域的应用[J]. 大数据, 2021, 7(5): 131-149.
WANG J Y, WEI X R, SHI W L, et al. Applications of reinforcement learning in the field of resource optimization[J]. Big Data Research, 2021, 7(5): 131-149.
- [5] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017.
- [6] CONITZER V, SANDHOLM T. AWESOME: a general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents[J]. Machine Learning, 2007, 67(1/2): 23-43.
- [7] EVERETT R, ROBERTS S. Learning against non-stationary agents with opponent modelling and deep reinforcement learning[C]//Proceedings of AAAI Spring Symposium Series. [S.l.:s.n.], 2018.
- [8] FOERSTER J N, ASSAEL Y M, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning[C]//Proceedings of Advances in Neural Information Processing Systems. New York: ACM Press, 2016: 2137-2145.
- [9] SUKHBAATAR S, SZLAM A, FERGUS R. Learning multiagent-communication with back propagation[C]//Proceedings of Advances in Neural Information Processing Systems. New York: ACM Press, 2016: 2244-2252.
- [10] KIM D, MOON S, HOSTALLERO D, et al. Learning to schedule communication in multi-agent reinforcement learning[C]//Proceedings of the 7th International Conference on Learning Representations. [S.l.:s.n.], 2019.
- [11] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning[J]. arXiv preprint, 2017, arXiv:1706.05296.
- [12] RASHID T, SAMVELYAN M, WITT C S D, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[C]//Proceedings of the 35th International Conference on Machine Learning. New York: JMLR, 2018: 4292-4301.
- [13] GUPTA J K, EGOROV M, KOCHENDERFER M J. Cooperative multi-agent control using deep reinforcement learning[C]//Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. New York: ACM Press, 2017: 66-83.
- [14] BAI H Y, CAI S J, YE N, et al. Intention-aware online POMDP planning for autonomous driving in a crowd[C]//Proceedings of 2015 IEEE International Conference on Robotics and Automation. Piscataway:

IEEE Press, 2015: 454-460.

- [15] SADIGH D, SASTRY S S, SESHIA S A, et al. Information gathering actions over human internal state[C]//Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2016: 66-73.
- [16] XIE A N, LOSEY D P, TOLSMA R, et al. Learning latent representations to influence multi-agent interaction[C]//Proceedings of the 4th Conference on Robot Learning. New York: JMLR, 2020: 575-588.
- [17] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning[M]//Machine learning proceedings 1994. Amsterdam: Elsevier, 1994: 157-163.
- [18] TERRY J K, GRAMMEL N, HARI A, et al. Revisiting parameter sharing in multi-agent deep reinforcement learning[J]. arXiv preprint, 2020, arXiv:2005.13625.
- [19] CHRISTIANOS F, PAPOUDAKIS G, RAHMAN A, et al. Scaling multi-agent reinforcement learning with selective parameter sharing[C]//Proceedings of the 38th International Conference on Machine Learning. New York: JMLR, 2021: 1989-1998.
- [20] QI S Y, ZHU S C. Intent-aware multi-agent reinforcement learning[C]//Proceedings of 2018 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2018: 7533-7540.
- [21] KIM W J, PARK J G, SUNG T C. Communication in multi-agent reinforcement learning: intention sharing[C]//Proceedings of the 9th International Conference on Learning Representations. [S.l.:s.n.], 2021.
- [22] SAMVELYAN M, RASHID T, WITT C S, et al. The StarCraft multi-agent challenge[C]// Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM Press, 2019: 2186-2188.

[作者简介]



王涵（1992- ），女，计算机软件新技术国家重点实验室（南京大学）博士生，主要研究方向为机器学习、强化学习。



俞扬（1982- ），男，计算机软件新技术国家重点实验室（南京大学）教授，主要研究方向为人工智能、机器学习、强化学习、演化学习。



姜远（1976- ），女，博士，计算机软件新技术国家重点实验室（南京大学）教授，主要研究方向为人工智能、机器学习。