多机智能对抗——基于 MADDPG 的对抗仿真

1 概述

近年来，随着人工智能技术的快速发展，多智能体系统（Multi-Agent Systems, MAS）在军事、工业以及科研等多个领域中得到了广泛应用[1]。其中，多智能体之间的对抗与协作问题，逐渐成为该领域研究的一个重要方向。相比传统依赖预设规则的对抗方式，基于强化学习的方法具有更强的适应能力，能够在复杂多变的环境中，通过与环境的不断交互，自主摸索并优化策略，从而提升系统的稳定性和灵活性。

DDPG[2]（deep deterministic policy gradient）是一种专门针对连续动作空间设计的强化学习算法，它采用了 Actor-Critic 架构，并结合了深度神经网络来增强模型表达能力。不过，在多智能体场景中，由于每个智能体的策略更新会影响到其他智能体的学习过程，环境呈现出明显的非平稳性，进而影响训练效果。MADDPG[3] (Multi-agent deep deterministic policy gradient)的出现解决了这个问题。本文实现了 MPE-simple_tag 环境下利用 MADDPG 算法训练的多智能体的对抗仿真。

2 研究内容

2.1 实验环境

MPE（Multi-Agent Particle Environment）是由 OpenAI 开发的多智能体粒子环境。它为多智能体强化学习（MARL）提供了一组轻量级、可自定义的仿真环境，涵盖合作、竞争和混合场景，适用于测试和开发多智能体系统中的学习算法。本文使用 MPE 下的 simple_tag 任务进行测试：

simple_tag（捕食者-猎物环境）：好智能体速度更快，并且希望避免被对手击中。对手智能体速度较慢，并且希望击中好的代理。障碍物挡住了去路。

- 智能体类型：
  - ✓ 好智能体（Good Agents）：通常为绿色，速度较快，目标是避免被对手捕捉。
  - ✓ 对手智能体（Adversary Agents）：通常为红色，速度较慢，目标是捕捉好智能体。
- 障碍物：环境中存在固定的障碍物，阻碍智能体的移动路径。
- 奖励机制：
  - ✓ 对手成功捕捉好智能体时获得正奖励。
  - ✓ 好智能体被捕捉时受到惩罚。
  - ✓ 好智能体离开边界区域会受到额外惩罚。
- 观察空间：每个智能体的观察包括自身速度和位置、其他智能体的相对位置和速度，以及障碍物的相对位置。
- 动作空间：默认使用离散动作空间，包括：无动作、向左、向右、向下、向上。

2.2 DDPG 算法

DDPG(深度确定性策略梯度) 算法是用来处理无限动作空间的环境并且使用离线策略的算法，它学习一个最佳的确定性策略，用梯度上升的方法使长期奖励最大化。DDPG 有四个关键部分：

1. Actor（策略网络）$\mu_\theta(s)$。智能体的决策策略。输入当前状态s_t，输出确定性动作a_t。引入随机噪声$\mathcal{N}$进行探索$a_t = \mu(s_t) + \mathcal{N}$。使用神经网络逼近最优的策略函数$\mu_{\theta^*}(s)$。
2. Critic（Q 网络）$Q_\omega(s,a)$。评价动作好坏的函数。输入状态s和动作 a，输出 Q 值。通过最小化目标损失（时序差分）学习。

   $$L = \mathbb{E}[(Q(s,a) - \Big(r + \gamma Q'\big(s', \mu'(s)\big)\Big))^2]$$

   其中Q'和μ'是目标网络
3. 目标网络（Target Networks）$\mu_{\theta^-}, Q_{\omega^-}$ 目标网络为主网络提供稳定目标，解决非稳态问题和训练发散。它们的参数θ^-, ω^-是在每一时间步结束时由 Actor 和 Critic 以及上一时间步的自身继承而来的。

   $$\omega^- \leftarrow \tau\omega + (1-\tau)\omega^- , \theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$$

   τ是一个比较小的数
4. 经验回放（Replay Buffer）。将智能体的每个经验 $(s_t,\ a_t,\ r_t,\ s_{t+1})$存入回放池 R，训练时随机抽取一批历史数据进行回忆学习，打破数据间的相关性。

DDPG Algorithm

随机噪声可用 $\mathcal{N}$ 表示，用随机的网络参数 ω 和 θ 分别初始化 Critic 网络 $Q_\omega(s,a)$ 和 Actor 网络 $\mu_\theta(s)$
复制相同的参数 $\omega^- \leftarrow \omega,\quad \theta^- \leftarrow \theta,\quad$ 分别初始化目标网络 $Q_{\omega^-}, \mu_{\theta^-}$
初始化经验回放池 $\mathcal{R}$
for 序列 $e = 1 \to E$ **do**
  初始化随机过程 $\mathcal{N}$ 用于动作探索
  获取环境初始状态 s_1
  for $t = 1 \to T$ **do**
    根据当前策略和噪声选择动作 $a_t = \mu_\theta(s_t) + \mathcal{N}$
    执行动作 a_t，获得奖励 r_t，环境状态变为 s_{t+1}
    将 (s_t, a_t, r_t, s_{t+1}) 存储进回放池 $\mathcal{R}$
    从 $\mathcal{R}$ 中采样 N 个元组 $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^N$
    对每个元组，用目标网络计算：$y_i = r_i + \gamma Q_{\omega^-}(s_{i+1}, \mu_{\theta^-}(s_{i+1}))$
    最小化目标损失：$L = \dfrac{1}{N}\sum_{i=1}^{N}(y_i - Q_\omega(s_i, a_i))^2$
    以此更新当前 Critic 网络
    计算策略梯度以更新 Actor 网络：$\nabla_\theta J \approx \dfrac{1}{N}\sum_{i=1}^{N} \nabla_\theta \mu_\theta(s_i) \nabla_a Q_\omega(s_i, a)\Big|_{a=\mu_\theta(s_i)}$
    更新目标网络：
    $\omega^- \leftarrow \tau\omega + (1-\tau)\omega^-,\quad \theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$
  end for
end for

2.3 MADDPG 算法

MADDPG（多智能体深度确定性策略梯度）在 DDPG 的基础上做了改进，提出了集中式训练、分布式执行的方案：所有智能体共享一个中心化的 Critic 网络，该 Critic 网络在训练的过程中同时对每个智能体的 Actor 网络给出指导，而执行时每个智能体的 Actor 网络则是完全独立做出行动，即去中心化地执行。这种架构很好地缓解了环境非平稳带来的干扰，提高了整体系统的稳定性与多智能体之间的协作效果。

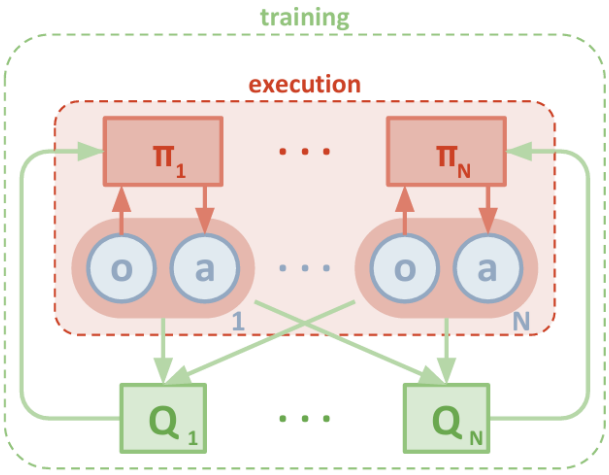


fig.1 Overview of multi-agent decentralized actor, centralized critic approach.

2.3 研究过程

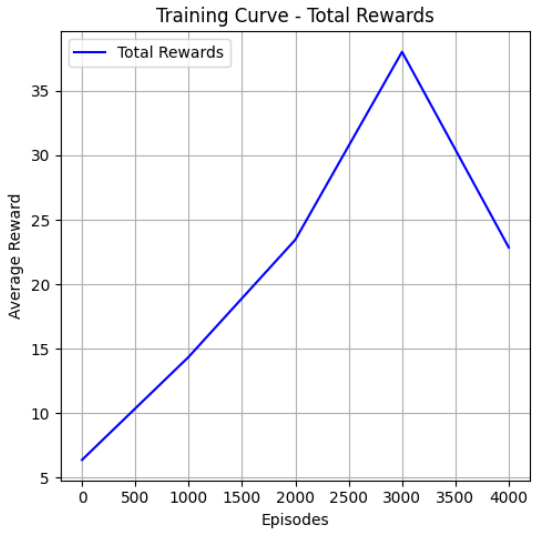
1. 安装并启用 openAI/Multiagent-particle-envs 和 openai/maddpg
2. 按以下表格进行参数配置：

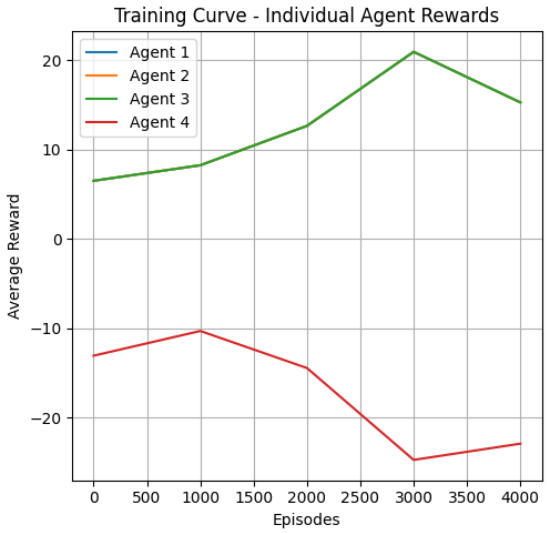
argument	setting
--scenario	simple_tag
--num-episodes	5000/10000/15000/20000
--good-policy	maddpg
--adv-policy	maddpg
--lr	1e-2
--batch-size	512
--exp-name	exp1/exp2/exp3/exp4

3. 一共经过 4 次训练，绘制了 4 幅 totalReward-individualReward 图。

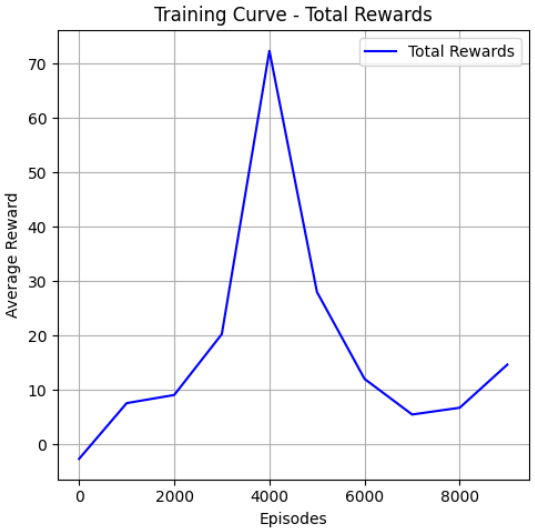
3 仿真结果

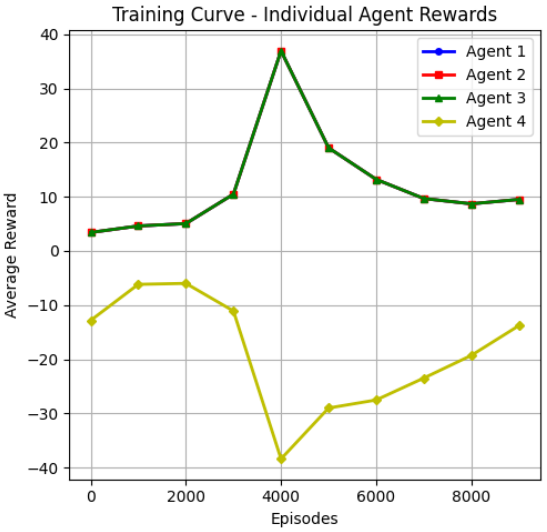
episodes =5000：



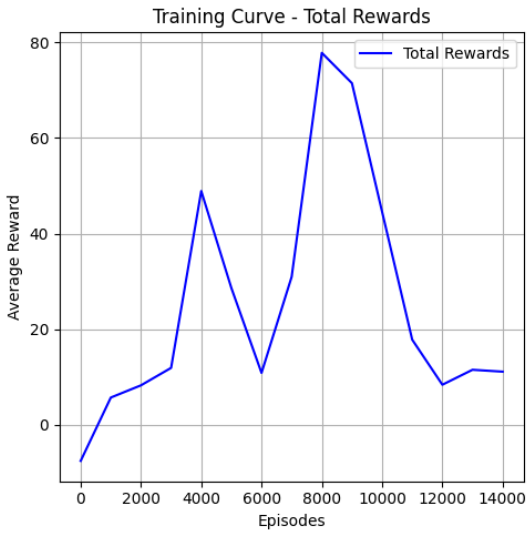


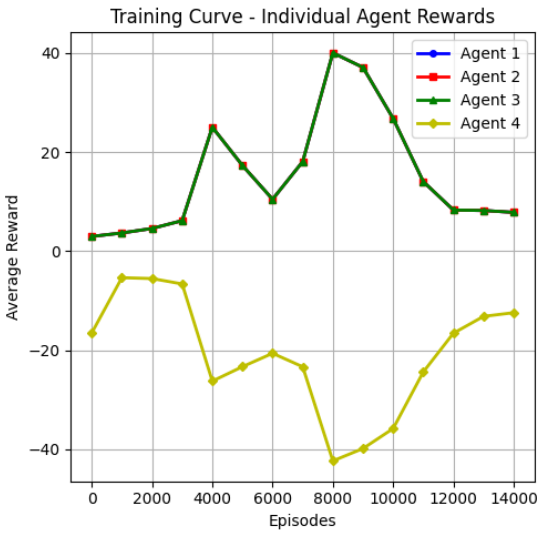
episodes =10000：





episodes =20000：





参考文献

[1]汤浩. 多智能体对抗博弈方法及仿真技术研究[D]. 四川:电子科技大学,2022.

[2] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra: Continuous control with deep reinforcement learning. ICLR (Poster)

[3] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, Igor Mordatch: Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. NIPS 2017: 6379-6390.

[4]动手学习强化学习. https://hrl.boyuai.com/chapter/3/.