**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）。智能体的决策策略。输入当前状态，输出确定性动作。引入随机噪声进行探索。使用神经网络逼近最优的策略函数。
2. Critic（Q网络）。评价动作好坏的函数。输入状态和动作a，输出Q值。

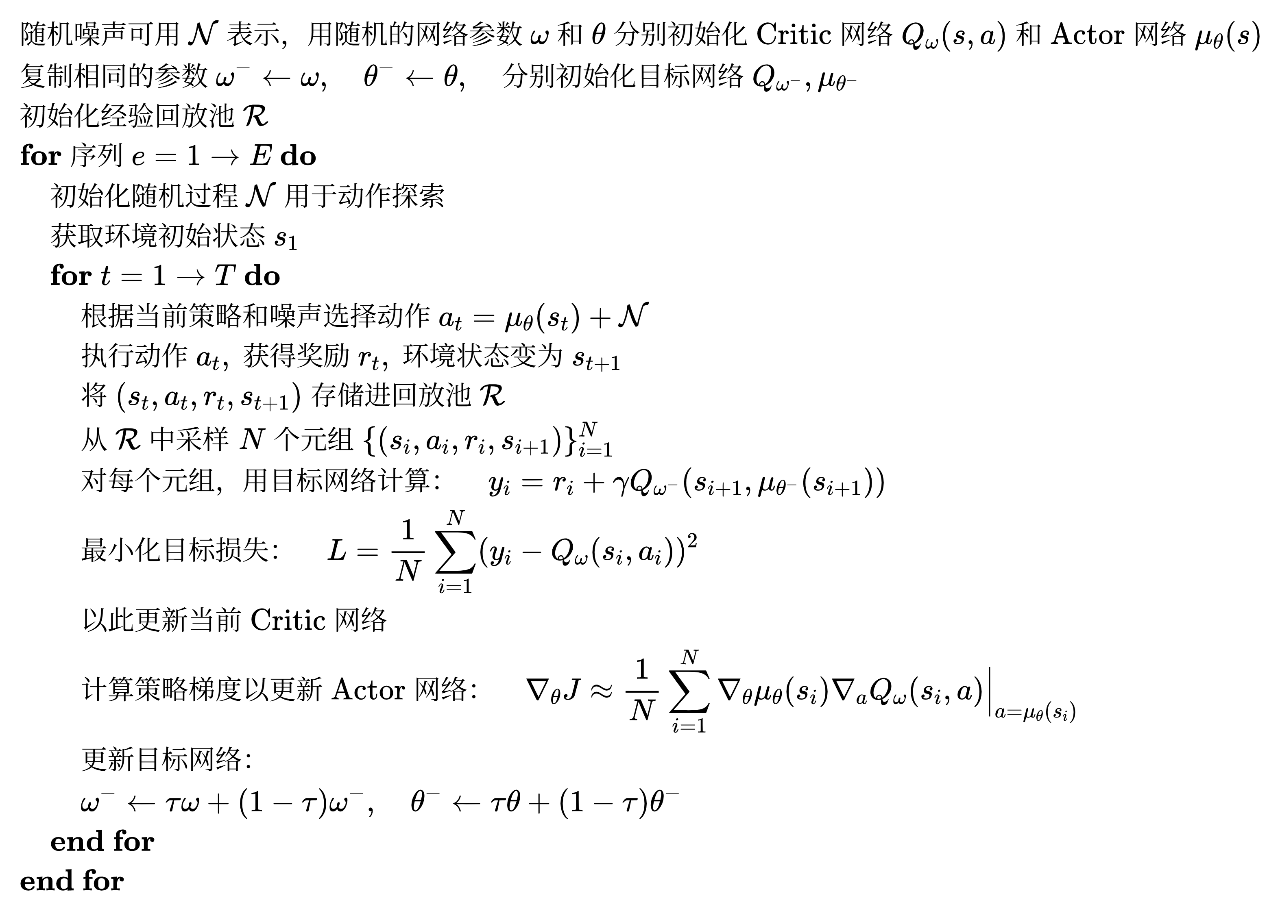
通过最小化目标损失（时序差分）学习。

其中和是目标网络

1. 目标网络（Target Networks） 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。

是一个比较小的数

1. 经验回放（Replay Buffer）。将智能体的每个经验 (,,,)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**

**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。

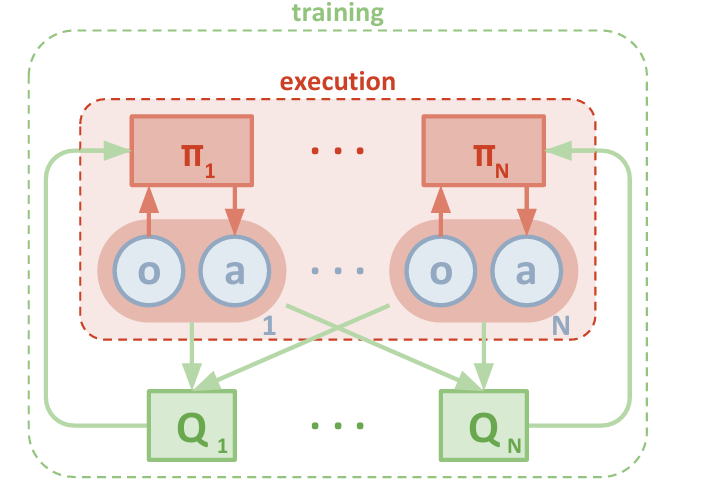


fig.1 Overview of multi-agent decentralized actor, centralized critic approach.

**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 |

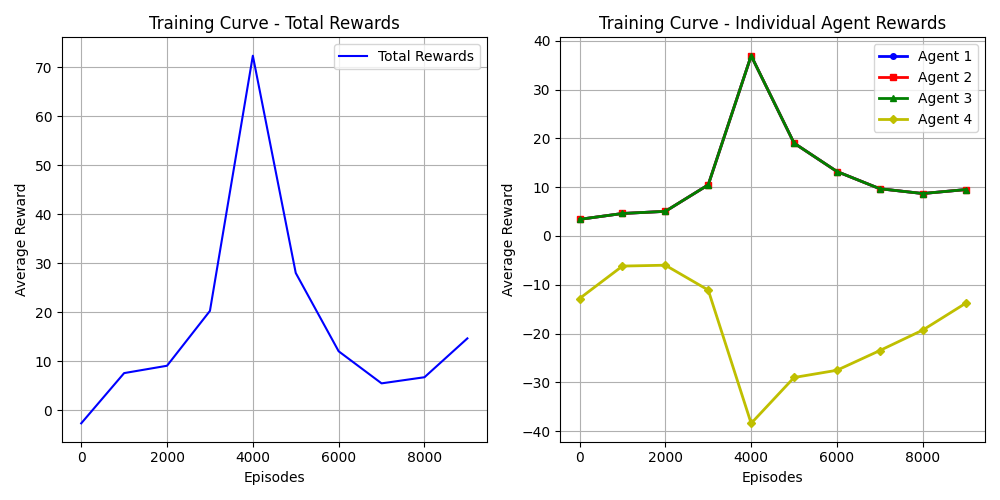
1. 一共经过4次训练，绘制了4幅totalReward-individualReward图。

**3仿真结果**

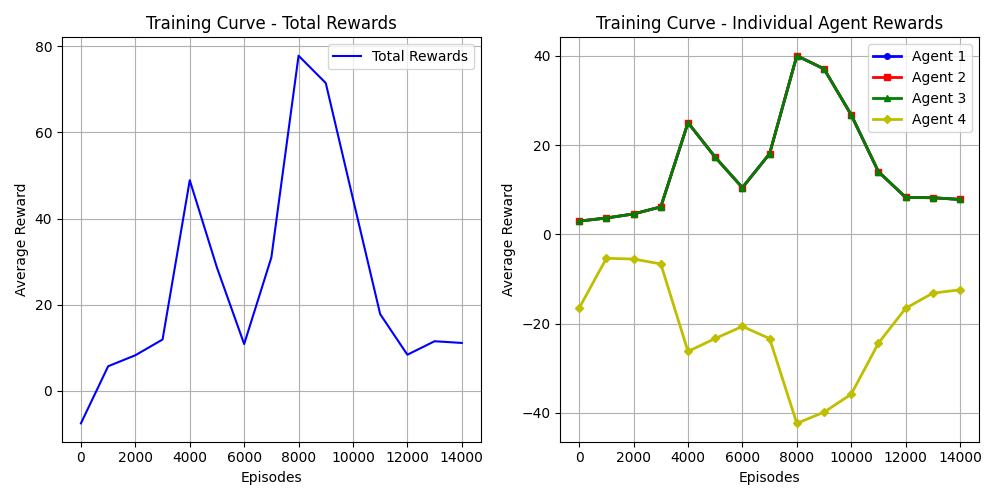
**episodes =5000：**



**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/.