266YYDS 参赛算法简介

学　　校：同济大学
团队成员：何子辰，宋春伟

0. 算法名称:

👉 Multi-Stage Transfer-TD3

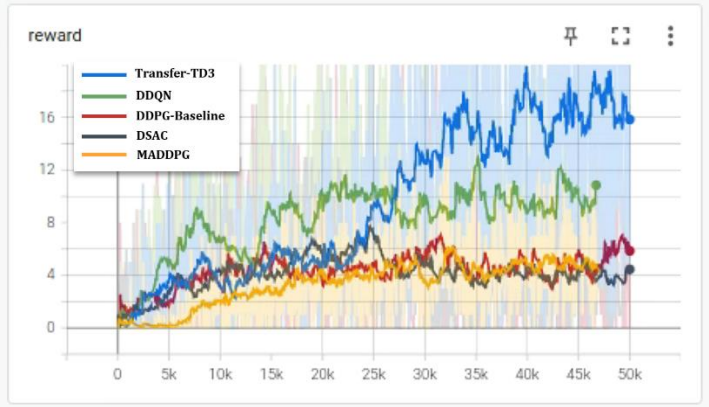

Fig.1 结果对比

1. 基础智能体选择

👉 TD3 算法

- ✓ Target Network
- ✓ Clipped Double Q
- ✓ Exploration Decay
- ✓ Delay Update

Pros: 目标网络使用是 Q 值学习种一种常见的方法，有效减少方差；延迟更新减小每次更新误差，解决估值函数与策略函数的耦合问题；双 Q 学习避免 Q 值的高估计问题。

👉 超参数设置

参数类型	数值
Max episodes	5e4
Episode length	200
Output activation	Gumbel Soft max
Tau	0.01
Buffer size	1e5
Gamma	0.95
Learning rate	1e-4
Batch size	256
Update nums per episode	2
Eps init	0.8
Eps decay speed	0.99998

2. Tricks

👉 Multi-Stage Training Paradigm 多阶训练范式;

- ✓ Basic TD3 参数作为 Actor 初始化参数
- ✓ 配合 eps decay，平衡探索利用
- ✓ 多阶训练
  - ■ 1st: vs. greedy policy (×2) & pretrained ddpg agent (×1)
  - ■ 2nd: vs. pretrained td3 policy (×3)
  - ■ 3rd: vs. pretrained ddqn (×3)

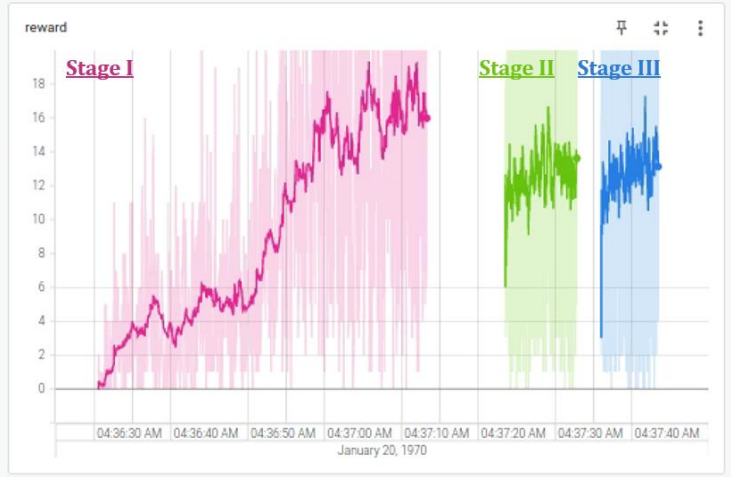

Fig.2 多阶算法训练过程

👉 延迟更新，每个 episode，更新两次参数;

👉 后续 Stage 训练，增加奖励函数中，done 状态下获胜奖励与失败惩罚度。