

一、nips - 2017 - « MADDPG »

1. MADDPG 提出原因：

传统RL方法 $\left\{ \begin{array}{l} \text{DQN 对 non-stationary 环境差} \\ \text{PG 高变异性 / Model 不收敛} \end{array} \right.$

提出创新 $\left\{ \begin{array}{l} \text{① 执行过程 Agent 只需局部信息} \\ \text{② 不需要 Model 与级且 Agent 间有插件重用} \\ \text{③ 应用于各种环境 Competitive Cooperative mix} \end{array} \right.$

总结 MADDPG 中心式训练，分布式执行。

策略函数只对某一个物体的场下 Agent 同步。

2. MADDPG 框架

decentralized actor
centralized critic A-C 架构

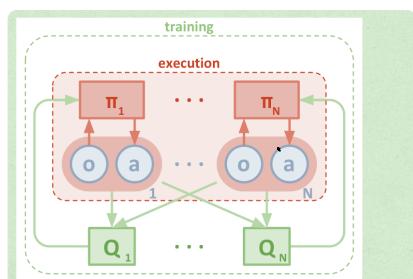


Figure 1: Overview of our multi-agent decentralized actor, centralized critic approach.

Idea: 通过确保 Env stationary .

如果已知 agents 动作，即使策略未知 π_i 确保环境是 stationary.

$$\text{状态转移 } P(s'|s, a_1, \dots, a_N) = P(s'|s, a_1, \dots, a_N, \pi'_1, \dots, \pi'_N)$$

①

Agent i 策略梯度

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s \sim p^\mu, a_i \sim \pi_i} [\underbrace{\nabla_{\theta_i} \log \pi_i(a_i | o_i)}_{\text{Actor}} \underbrace{Q_i^\pi(x, a_1, \dots, a_N)}_{\text{Critic}}].$$

Actor
Agent 找到
分步式

Critic
必须已知 Agent i's action
 $X = (0, \dots, 0_n)$ All agents 观察
也必须知道环境状态信息 S

② 集中式策略

$$\nabla_{\theta_i} J(\mu_i) = \mathbb{E}_{x, a \sim \mathcal{D}} [\underbrace{\nabla_{\theta_i} \mu_i(a_i | o_i)}_{\text{Actor}} \underbrace{\nabla_{a_i} Q_i^\mu(x, a_1, \dots, a_N)|_{a_i=\mu_i(o_i)}}_{\text{Critic+Max}}],$$

集中式 动机是通用的
产生后高激励

Actor 和 Single Agent

一样处理

求 $\nabla_{a_i} Q_i$ 必须已知 $a_i = \mu_i(o_i)$

即 All agent 都必须观测 (中心式)

Paper 认为在 Complex Env 下这不是
特别的限制 (改进)

MADDPG 只需知道 Oj 观测信息即可。

③ 估计 Other Agents 的策略

$$\text{Max } \mathcal{L}(\phi_i^j) = -\mathbb{E}_{o_j, a_j} [\log \hat{\mu}_i^j(a_j | o_j) + \lambda H(\hat{\mu}_i^j)], \text{ 近似估计 } \hat{\mu}_i^j$$

MADDPG 只需知道 O_j 观测信息即可。

④ Competitive Env Agent 和对手的策略变化敏感。

$$\nabla_{\theta_i^{(k)}} J_e(\boldsymbol{\mu}_i) = \frac{1}{K} \mathbb{E}_{\mathbf{x}, a \sim \mathcal{D}_i^{(k)}} \left[\nabla_{\theta_i^{(k)}} \boldsymbol{\mu}_i^{(k)}(a_i | o_i) \nabla_{a_i} Q^{\boldsymbol{\mu}_i}(\mathbf{x}, a_1, \dots, a_N) \Big|_{a_i = \boldsymbol{\mu}_i^{(k)}(o_i)} \right].$$

↓
提出 Policy Ensemble 模型

每个 Episode 7-10 是 Sub-optimal

MADDPG 是 MARL 领域里程碑 Paper 使用 A-C 架架。

采用 Critic 集中式训练，Actor 分布式执行。

Agent 间异构

ER 提高 Sample efficiency

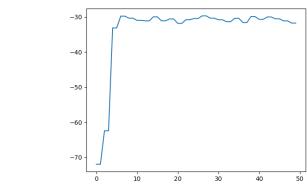
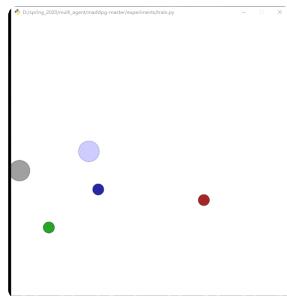
应用于 Cooperative Competitive Mix 各种场景。

但 $\mathbf{x} = (O_1, O_2, \dots, O_n)$ 随 Agent 数量增长会成几何增长。

如何划分 Agent 之间系统关系

⊕ 実験問題 Cooperative Competetive Mix

Speaker - listener



mean reward ~ -30 ドラク

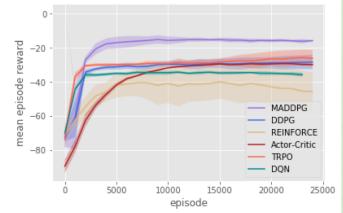


Figure 4: Agent reward on cooperative communication after 25000 episodes.

2.

AAIMAS - 2018 - Value - Decomposition Networks

ICML - 2018 - QMIX

ICML - 2019 - QTRAN

~~ただし MADDPG が遅い~~.