强化学习2022
第13节

涉及知识点：
目标导向的强化学习

目标导向的强化学习

张伟楠 – 上海交通大学

课程大纲

强化学习基础部分

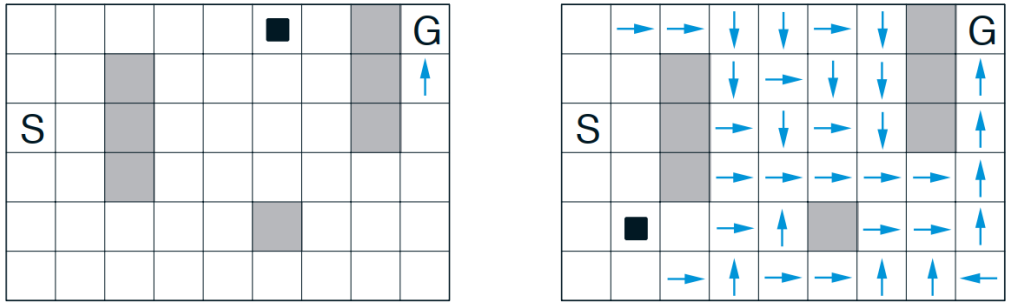
1. 强化学习、探索与利用
2. MDP和动态规划
3. 值函数估计
4. 无模型控制方法
5. 规划与学习
6. 参数化的值函数和策略
7. 深度强化学习价值方法
8. 深度强化学习策略方法

强化学习前沿部分

9. 基于模型的深度强化学习
10. 模仿学习
11. 离线强化学习
12. 参数化动作空间
13. 目标导向的强化学习
14. 多智能体强化学习
15. 强化学习大模型
16. 技术交流与回顾

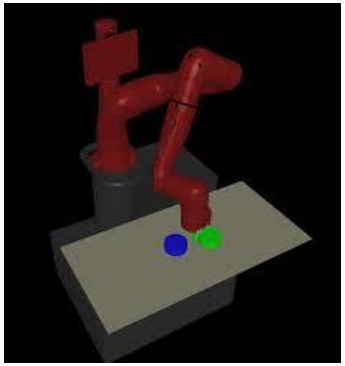
强化学习的任务复杂度

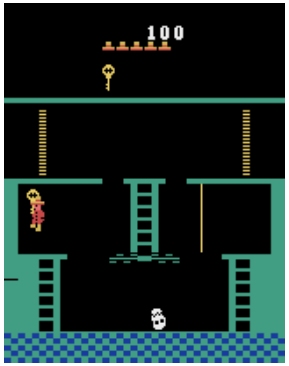
- 大部分的（深度）强化学习方法一般都只能完成特定任务
  - 例如走迷宫，需要固定地图、起点和重点，智能体学习到特定路线



  - 如果随意标定起点和终点，智能体不经过新的训练往往很难有效完成
  - 又例如，智能体需要走入到不同房间去拿物品，全都拿到了才算完成任务。此整体任务比较复杂，但可以分解成为几个简单的小任务。

机械臂抓取
多个物品



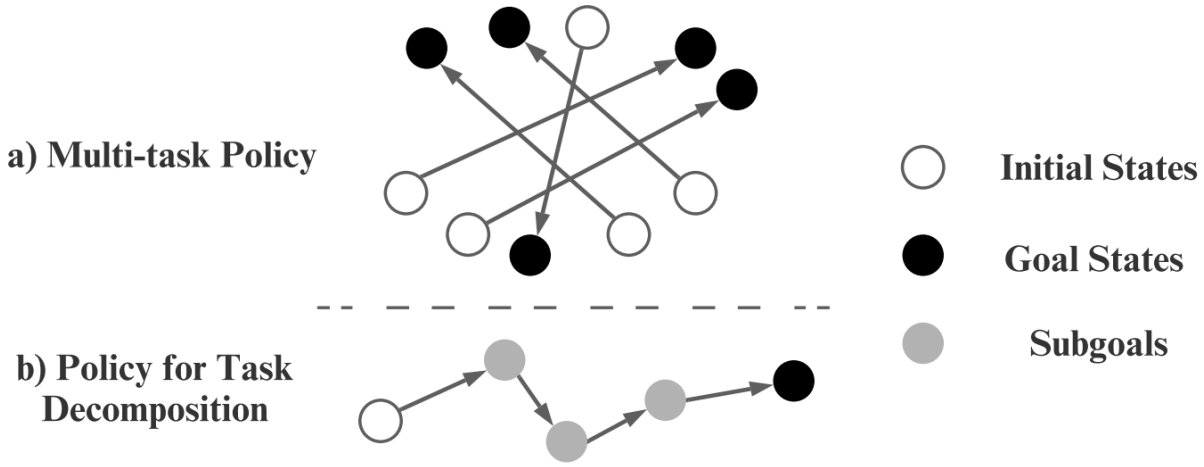


先拿到钥匙
才能去开门

目标导向的强化学习

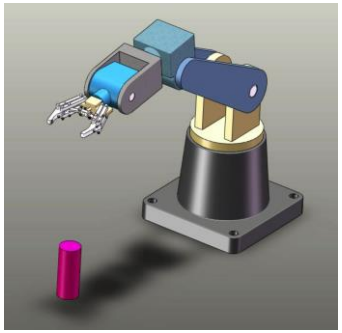
- 思路：如果把每次任务的终点以一种目标信息显式加到智能体的输入和奖励信号中，是否能训练出能完成不同终点的任务呢？

- 目标导向的强化学习（Goal-oriented RL，GoRL）
  - 在MDP环境中额外定义一个目标，将此目标显式地告诉智能体，并在训练过程中使用此目标信息，进而在测试阶段给定新的目标，使得智能体能直接泛化来解决此任务。



目标导向的强化学习的问题建模

- GoRL的问题通常被建模为一个MDP的扩充$(\mathcal{S}, \mathcal{A}, p, \gamma, r_g, \mathcal{G}, \phi)$
  - $\mathcal{S}$ 是环境状态集合
  - $\mathcal{A}$ 是智能体动作集合
  - $p: \mathcal{S} \times \mathcal{A} \to \Omega(\mathcal{S})$ 是环境转移的概率，其中$\Omega(\mathcal{S})$在$\mathcal{S}$的分布集合
  - $\gamma \in (0,1)$ 是奖励随时间步的衰减因子
  - $\mathcal{G}$ 是目标集合
  - $\phi: \mathcal{S} \to \mathcal{G}$ 是将状态映射成目标的函数
    - 有时直接就是等式转换 $g = s$
    - 有时是取s的几个维度作为g
  - $r_g: \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$ 为加入目标信息的新奖励函数
  - $\pi_i: \mathcal{S} \times \mathcal{G} \to \Omega(A_i)$ 为加入目标信息的策略



- 智能体目标是仍然是最大化期望累积奖励

$$J(\pi) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t,g), g \sim p_g, s_{t+1} \sim p(\cdot|s_t,a_t)} \left[\sum_{t=0}^{T} \gamma^t r_g(s_t, a_t, s_{t+1}) \right]$$

Liu, Minghuan, Menghui Zhu, and Weinan Zhang. Goal-Conditioned Reinforcement Learning: Problems and Solutions. IJCAI 2022.

目标导向的强化学习的具体模块定义

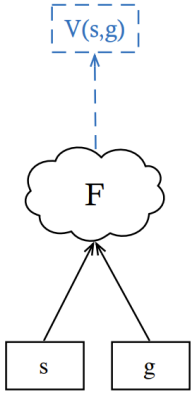
- 目标导向的奖励函数
  - $r_g: \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$ 为加入目标信息的新奖励函数

$$r_g\left(s_t, a_t, s_{t+1}\right) = \begin{cases} 0, & \left\|\phi\left(s_{t+1}\right) - g\right\|_2 \leq \delta_g \\ -1, & \text{otherwise} \end{cases}$$

（一个较小的阈值）

- 目标导向的策略
  - $\pi_i: \mathcal{S} \times \mathcal{G} \to \Omega(A_i)$ 为加入目标信息的策略
  - 最简单的定义可以直接即使原始策略 $\pi(a|s,g) = \pi(a|s)$
  - 也可以定义策略网络同时输入s和g

- 普适价值函数逼近（Universal Value Function Approximators）
  - $V: \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ 为加入目标信息的目标函数逼近网络
  - 最简单的实现可以直接在MLP输入层拼接 s和g



Schaul, Tom, et al. Universal value function approximators. International conference on machine learning. ICML, 2015.

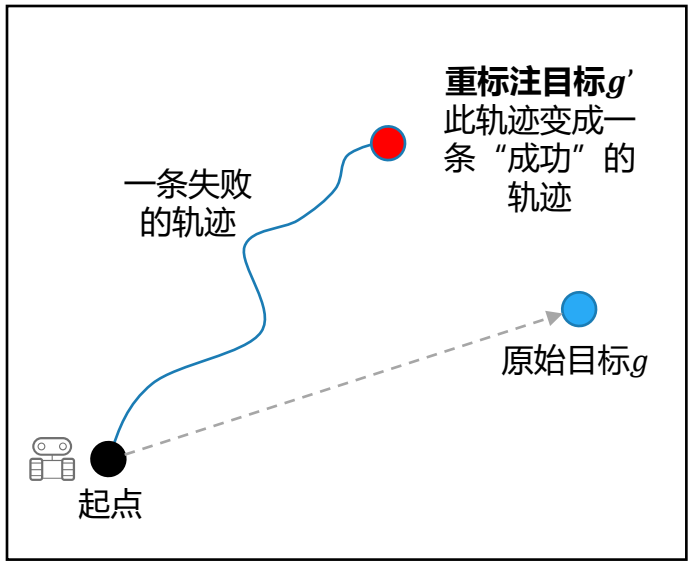
经典GoRL算法：HER

- 观察：可以发现目标导向的强化学习的奖励往往是非常稀疏的。智能体在训练初期难以完成目标而只能得到的奖励，从而使得整个算法的训练速度较慢。
- 事后经验回放（hindsight experience replay，HER）算法
  - HER使用重标注（relabeling）的方法，有效地利用“失败”的经验

HER算法流程

- 初始化策略π的参数θ，初始化经验回放池R
- **For** 序列 $e = 0 \to E$
  - 根据环境给予的目标g和初始状态s_0，使用π在环境中采样得到轨迹$\{s_0, a_0, r_0, \dots, s_T, a_T, r_T, s_{T+1}\}$，将其以$(s, a, r, s', g)$的形式存入$R$中
  - 从R中采样N个(s, a, r, s', g)元组
  - 对于这些元组，选择一个状态s''，将其映射为新的目标$g' = \phi(s'')$并计算新的奖励值$r' = r_{g'}(s, a, s')$，然后将新的数据(s, a, r', s', g')替换原先的元组
  - 使用这些新元组，对策略π进行训练
- **End for**

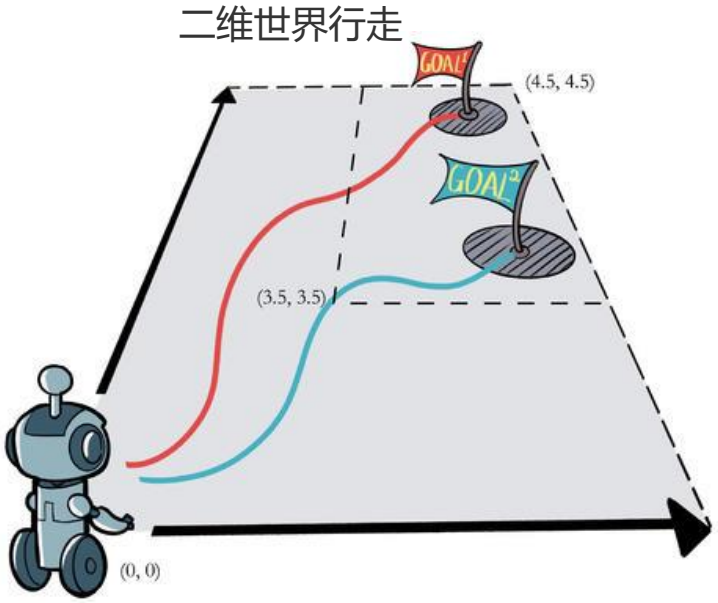
注：HER框架适合各种RL方法（DQN、DDPG、PPO）

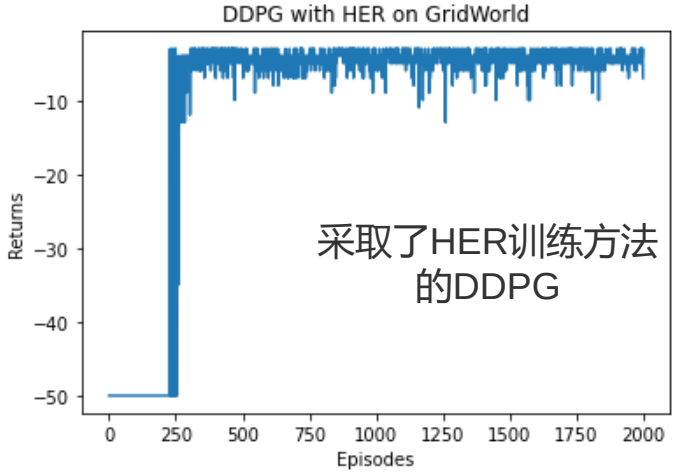


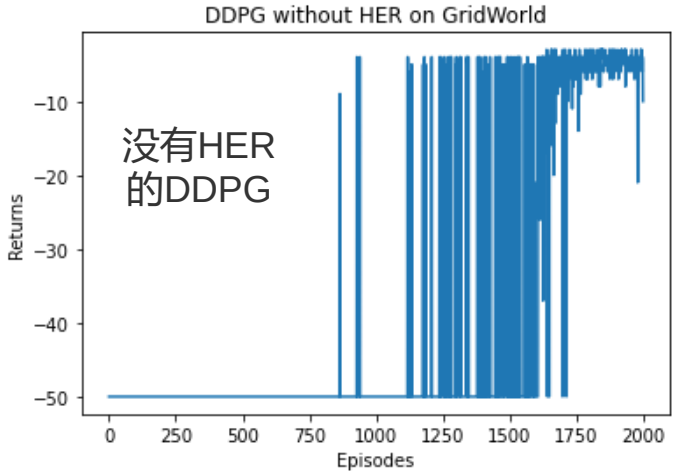
Andrychowicz, Marcin et al. Hindsight experience replay. NIPS 2017.

HER实验

- 实验环境：最基本的二维世界行走至目标点的任务







- 二维网格世界上，每维范围是，在每一个序列初始，智能体处于的位置，环境将自动从的一个矩形区域内生成一个目标。
- 每一个时刻智能体可以选择纵向和横向分别移动作为动作。当智能体距离目标足够近的时候，它将收到的奖励并结束任务，否则奖励为。每一条序列的最大长度为50。

THANK YOU