

Hindsight Experience Replay-hoho

论文试图解决什么问题？

奖励稀疏问题（sparse reward）：回合中长时间奖励不变，直到回合最后的时间步才有明显的奖励变化。如在整体轨迹中，目标g一直达不到的话（目标g也看作是一种状态s），奖励一直是-1.

这是否是一个新的问题？

不是一个新问题。本论文之前已经有研究次问题的解决方案，譬如奖励塑性方法，但这种方法需要很强的领域知识，不具备通用性。还有一些如基于计数的改善探索的方法，或者bootstrapped DQN，但这些方法没有解决本质问题：真正的问题不在于所探索的状态缺乏多样性，而是探索如此大的状态空间是不切实际的。

这篇文章要验证一个什么科学假设？

是否可以有不依赖领域知识的，改善奖励稀疏问题的强化学习方法。

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

当时的研究工作可以归为：

1. 基于经验回放技术，如优先级经验回放；
2. 同时面向多个任务的策略学习方法，如UVFA(Universal Value Function Approximators)，一种DQN的扩展
3. 课程学习。这也是一种解决奖励稀疏的方法，本文的HER方法可以看作是一种隐式的课程学习方法

论文中提到的解决方案之关键是什么？

本文提出了HER方法（Hindsight Experience Replay），核心思想是：这每个回合设置智能体要达到的多个目标而非只有一个。

具体来说：一般来说一个回合的轨迹只会告诉我们如何达到最后的状态 s_T ，而不会告诉我们如何达到某个目标（即某个状态 g ）。可以使用off-policy强化学习方法（因为需要经验回放的支持），在经验回放池中将目标 g 替换为 s_T 。即换一个目标来重新审视整条轨迹，虽然没达到目标 g ，但策略在探索过程中，完成了 s_1, s_2, \dots, s_T 等对应的目标 $(\phi(s_1), \phi(s_2), \dots, \phi(s_T))$ 。如果用这些目标将原先的目标替换成新的目标 g' ，重新计算轨迹中的奖励值，就能使策略从失败的经验中得到对训练有用的信息。

如此一来，至少一半的经验回放轨迹中的奖励都包含不同于-1的奖励（本文作者用了一个例子：直到回合结束前，每个时间步的奖励都是-1，最后的时间步奖励才不是-1）。

这就要是使用多目标强化学习方法。训练策略和价值函数时，输入要包含状态 s 和目标 g ，目标就是智能体要达到的某种状态 s ，我们可以定义一个映射 $f_g(s) = [s == g]$ ，所以目标 g 要满足 $f_g(s) = 1$ ，否则为0。那么，训练通用的策略可以基于一些分布通过采样目标和初始状态，让智能体与环境交互，这每个时间步，当智能体不能达到目标时，就设置负的奖励，如 $r_g(s, a) = -[f_g(s) = 0]$ 。

算法如下：

Algorithm 1 Hindsight Experience Replay (HER)

Given:

- an off-policy RL algorithm \mathbb{A} , ▷ e.g. DQN, DDPG, NAF, SDQN
- a strategy \mathbb{S} for sampling goals for replay, ▷ e.g. $\mathbb{S}(s_0, \dots, s_T) = m(s_T)$
- a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \rightarrow \mathbb{R}$. ▷ e.g. $r(s, a, g) = -[f_g(s) = 0]$

Initialize \mathbb{A} ▷ e.g. initialize neural networks

Initialize replay buffer R

for episode = 1, M **do**

 Sample a goal g and an initial state s_0 .

for $t = 0, T - 1$ **do**

 Sample an action a_t using the behavioral policy from \mathbb{A} :

$$a_t \leftarrow \pi_b(s_t || g)$$

▷ $||$ denotes concatenation

 Execute the action a_t and observe a new state s_{t+1}

end for

for $t = 0, T - 1$ **do**

$$r_t := r(s_t, a_t, g)$$

 Store the transition $(s_t || g, a_t, r_t, s_{t+1} || g)$ in R ▷ standard experience replay

 Sample a set of additional goals for replay $G := \mathbb{S}(\text{current episode})$

for $g' \in G$ **do**

$$r' := r(s_t, a_t, g')$$

 Store the transition $(s_t || g', a_t, r', s_{t+1} || g')$ in R

▷ HER

end for

end for

for $t = 1, N$ **do**

 Sample a minibatch B from the replay buffer R

 Perform one step of optimization using \mathbb{A} and minibatch B

end for

end for

论文中的实验是如何设计的？

本文自己搭建了一个机械臂操作环境，分别用机械臂操控进行三个动作：推动物体到目标点（pushing），扫动物体使其滑动到目标点（sliding），抓取物体最终移动到目标点（pick-and-place）

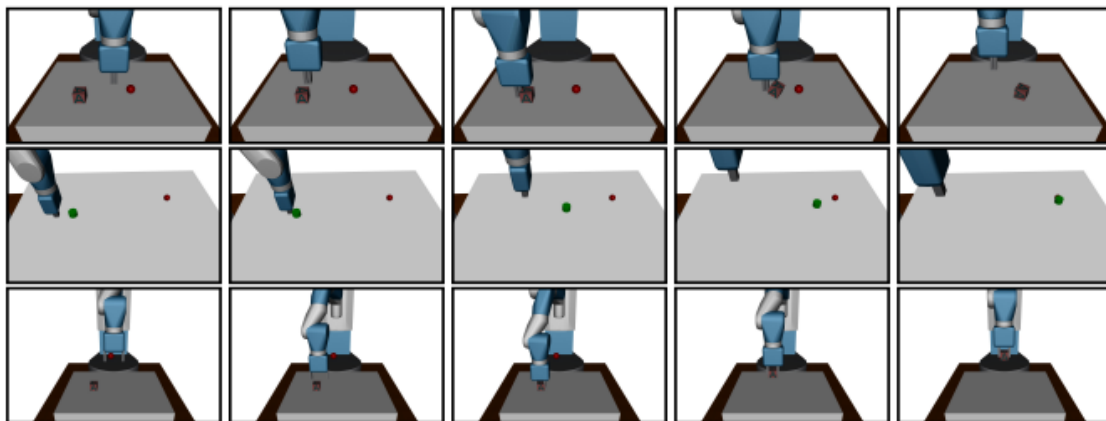


Figure 2: Different tasks: *pushing* (top row), *sliding* (middle row) and *pick-and-place* (bottom row). The red ball denotes the goal position.

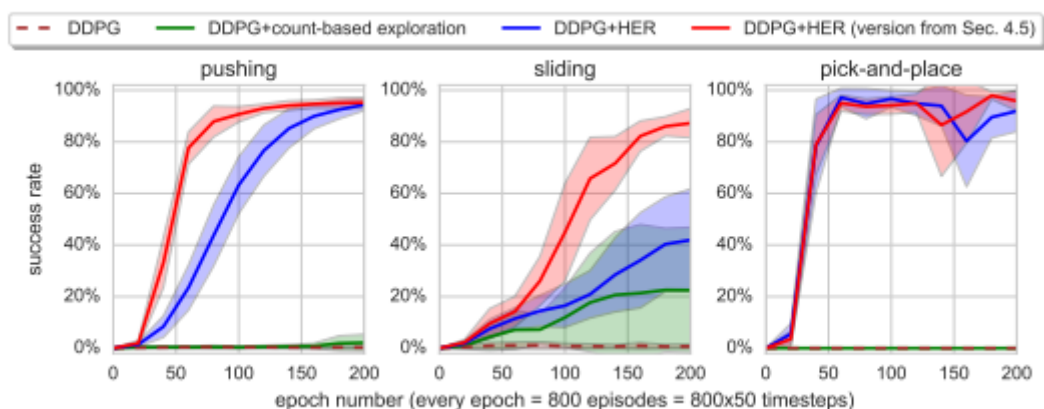
并且用加入HER和不加入HER的DDPG算法作对比。

用于定量评估的数据集是什么？代码有没有开源？

没数据集。没开源代码。

论文中的实验及结果有没有很好地支持需要验证的科学假设？

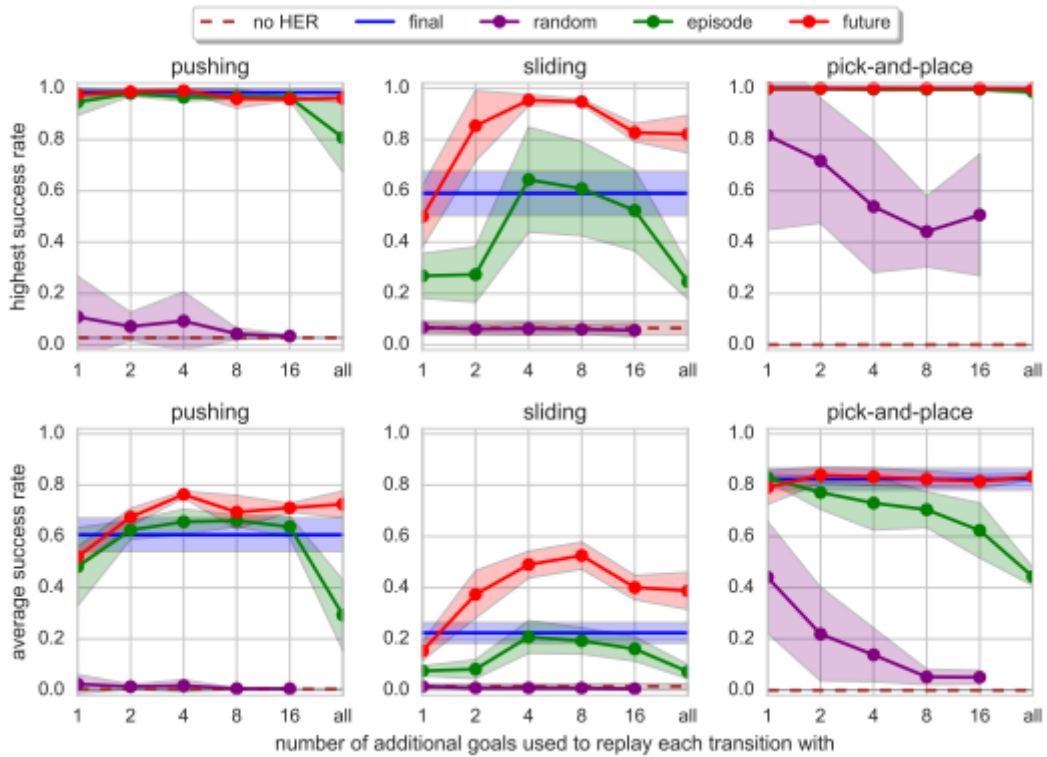
实验证明加入HER的DDPG算法可以持续提升机械臂算法。



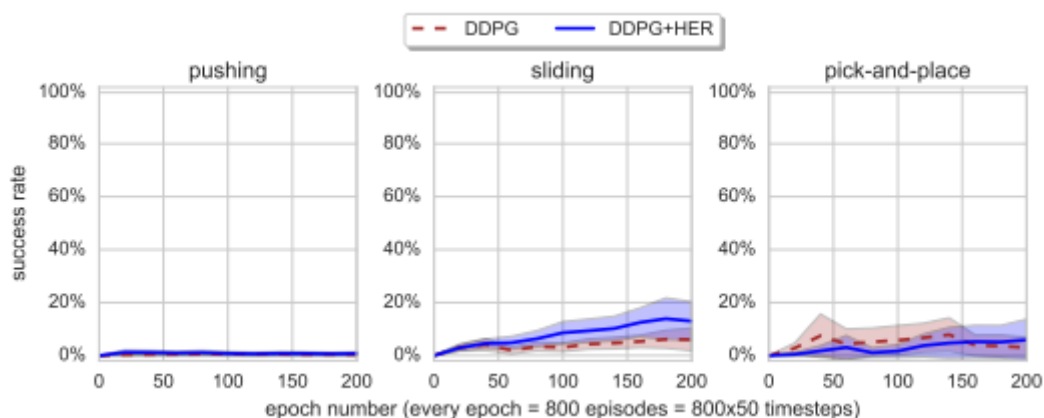
本文所使用的目标是最后的状态（称为final规划，strategy final），实验还对比了其他目标规划：

- future: 在同一个回合中，某个state之后的随机k个state作为goal；
- episode: 在同一个回合中，随机采样k个state作为goal，和future不同的是不需要从某个state之后随机采样；
- random: 在整个训练流程中，从多个回合的数据采样k个state作为goal

实验证明，使用future规划是目前最优的



本文还特别的对奖励塑形方法进行对比，将奖励函数塑形为 $r(s, a, g) = -|g - s'_{object}|^2$ ，其中 s'_{object} 是机械臂实验中物体在状态s下的位置。结果发现即使用HER，效果也是很差：



证明奖励塑形方法的弱点：十分依赖专家领域知识，不通用。

这篇论文到底有什么贡献？

提出了一种全新的解决奖励稀疏的比较通用方法：HER

下一步呢？有什么工作可以继续深入？

1. 如何改进奖励塑形方法（自己提的，非论文作者提）
2. HER只能结合off-policy RL方法，那on-policy呢？（自己提的，非论文作者提）
3. HER通过修改历史轨迹设置其目标，使奖励信号更加稠密，进而从原本失败的数据中学习使新任务成功的经验（根据下一步状态是否离目标足够近来修改当前步的奖励信号），改善每个时间片段带给智能体策略的学习价值。（总结）