1707.01495 - Hindsight Experience Replay

Hindsight Experience Replay

Marcin Andrychowicz*, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong,
Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel†, Wojciech Zaremba†
OpenAI

- **Yunqiu Xu**
- Focus on sparse reward:
  - HER: allows sample-efficient learning from rewards
  - Reward can be sparse and binary
  - Do not need complicated reward engineering
  - Experiment on robot arm manipulating
- I treat this work as further reading for imitation learning, following are some reference:
  - Another's notes
  - Two minute papers

1. Introduction

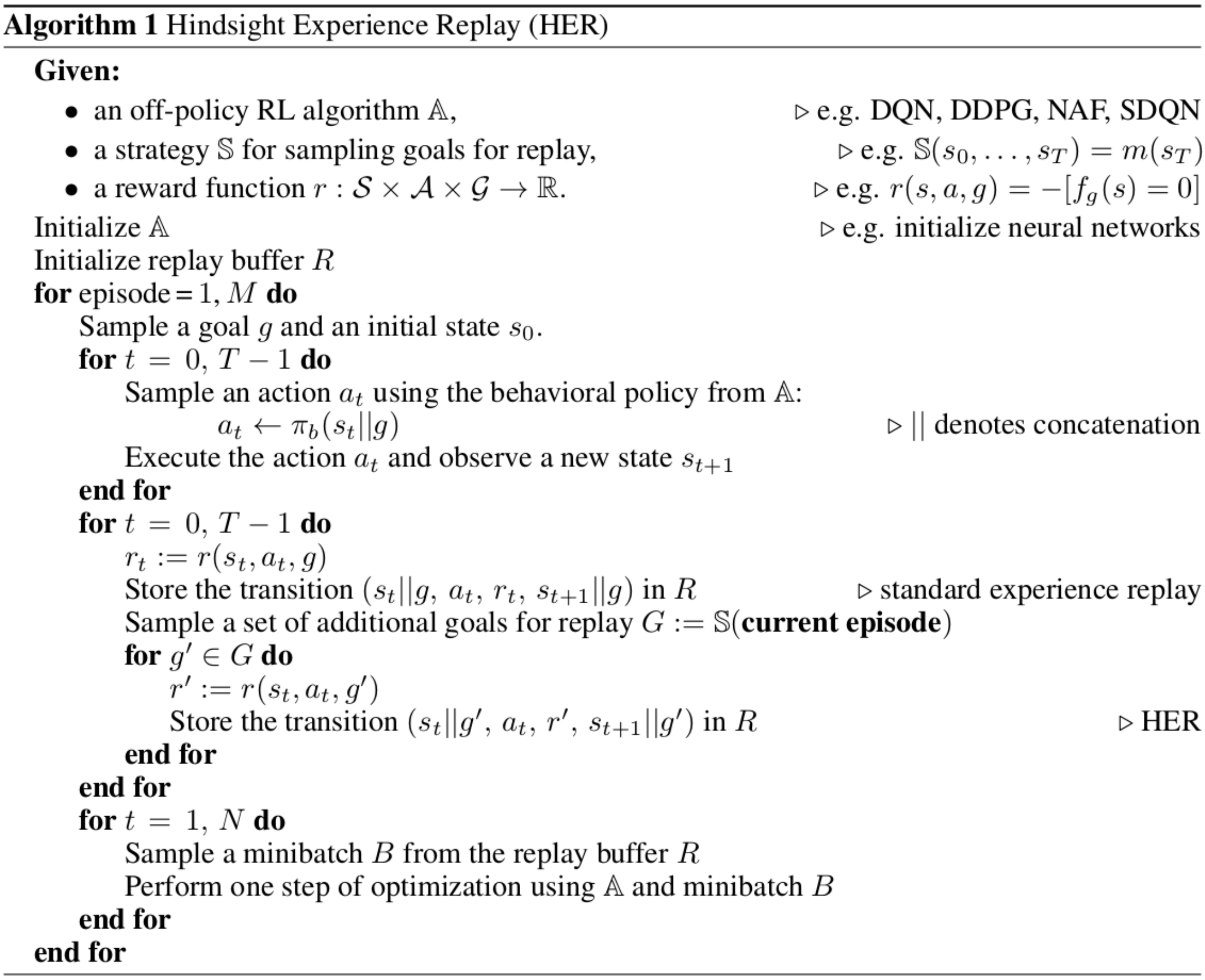
- Challenges: reward engineering → sparse reward
- Insight from human learning:
  - 人类从undesired outcome中学习到同desired ouecome中一样多的信息
  - 换言之, 人类可以从不好的结果中吸取教训, 而RL只能根据得到的reward学习
- 从另一份工作(Universal value function approximators, Schaul 2015)中得到的灵感: 每个episode都设定不同的目标 (疑问, 是否类似课程学习这样循序渐进的)
- Hindsight Experience Replay:
  - Suitable with off-policy RL (e.g. DQN)
  - Assumption: multiple goal can be achieved → 到达每个状态都会被给予不同的目标

2. Background

- DDPG
  - AC-like DQN for continuous action spaces
  - Actor:
    - $\pi : S \rightarrow A$
    - Target policy to choose action
    - Try to maximize action value with respect to policy's parameters
  - Critic:
    - $Q^\pi : S \times A \rightarrow R$
    - Action-value function to evaluate Q value
    - Try to minimize Bellman error
  - Learning: update C using Bellman, update A using PG
- UVFA: Universal Value Function Approximators
  - There are more than one goal we may try to achieve
  - Learning: for each episode sample a state-goal pair, so the "goal" stay fixed in this episode

3. HER

- Key idea: replay episodes with a different goal.
- Assumption: need multiple goals in an environment.
- HER can be combined with off-policy RL algorithms, so it doesn't replace them but can augment them
- Store an episode $(s_1, s_2, \ldots, s_T)$ in replay buffer twice:
  - One is with original goal
  - Another it with "final goal" in this episode: if the agent still fails at s_T, then set s_T as goal for this episode
- Simplest version
  - Store both final state s_T and original goal g per episode
  - Shape a mapping function $m(s_T)$ to represent state-goal pair



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. $\mathbb{S}(s_0, \ldots, s_T) = m(s_T)$
  • a reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}$.   ▷ 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 $G := \mathbb{S}($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
```

4. Experiment

- Robot arm manipulating tasks:

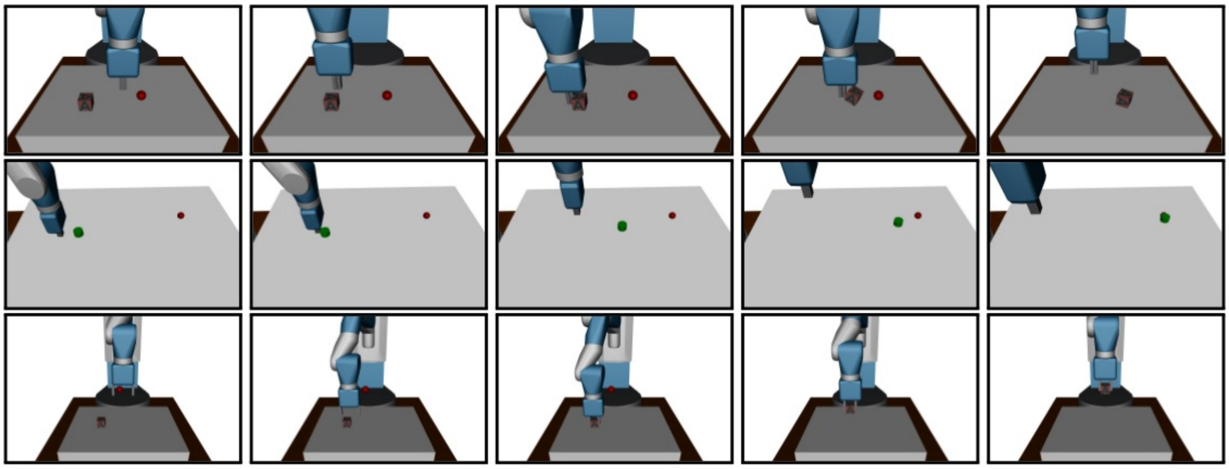


Figure 2: Different tasks: *pushing* (top row), *sliding* (middle row) and *pick-and-place* (bottom row). The red ball denotes the goal position.

- Does HER improve performance
    - Multiple goals

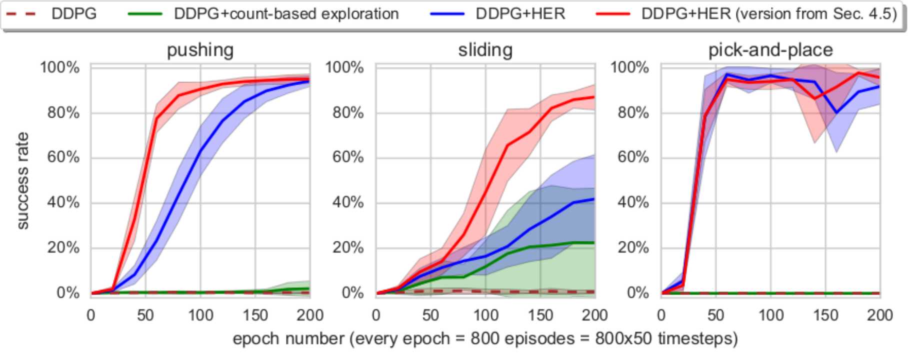


Figure 3: Learning curves for multi-goal setup. An episode is considered successful if the distance between the object and the goal at the end of the episode is less than 7cm for pushing and pick-and-place and less than 20cm for sliding. The results are averaged across 5 random seeds and shaded areas represent one standard deviation. The red curves correspond to the `future` strategy with $k = 4$ from Sec. 4.5 while the blue one corresponds to the `final` strategy.

    - Only one goal

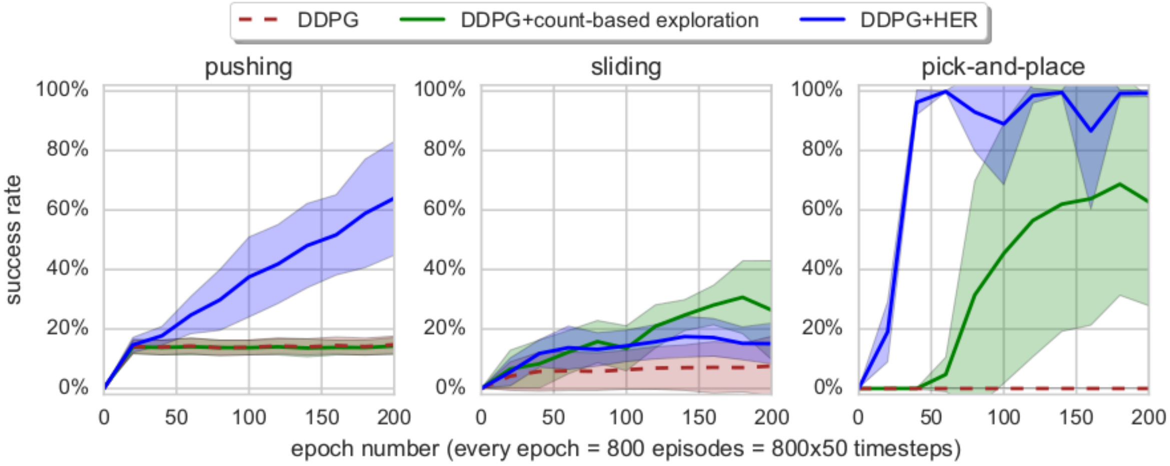


Figure 4: Learning curves for the single-goal case.

- How does HER interact with reward shaping (not only binary)

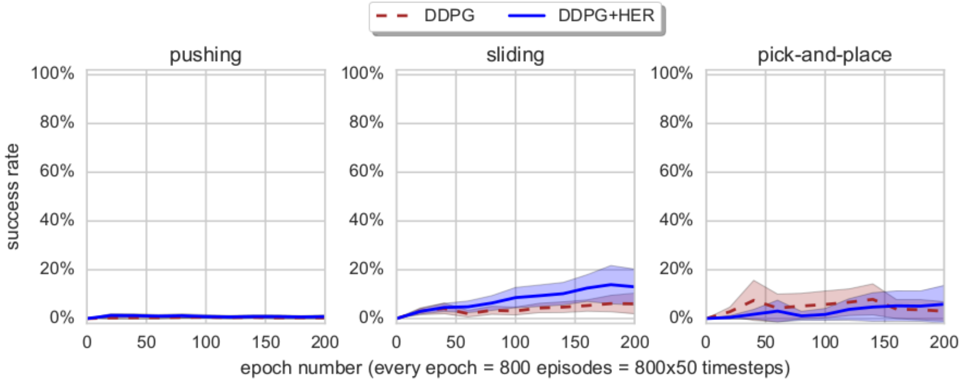


Figure 5: Learning curves for the shaped reward $r(s, a, g) = -|g - s'_{\textbf{object}}|^2$ (it performed best among the shaped rewards we have tried). Both algorithms fail on all tasks.

- How many goals should we replay each trajectory with and how to choose them

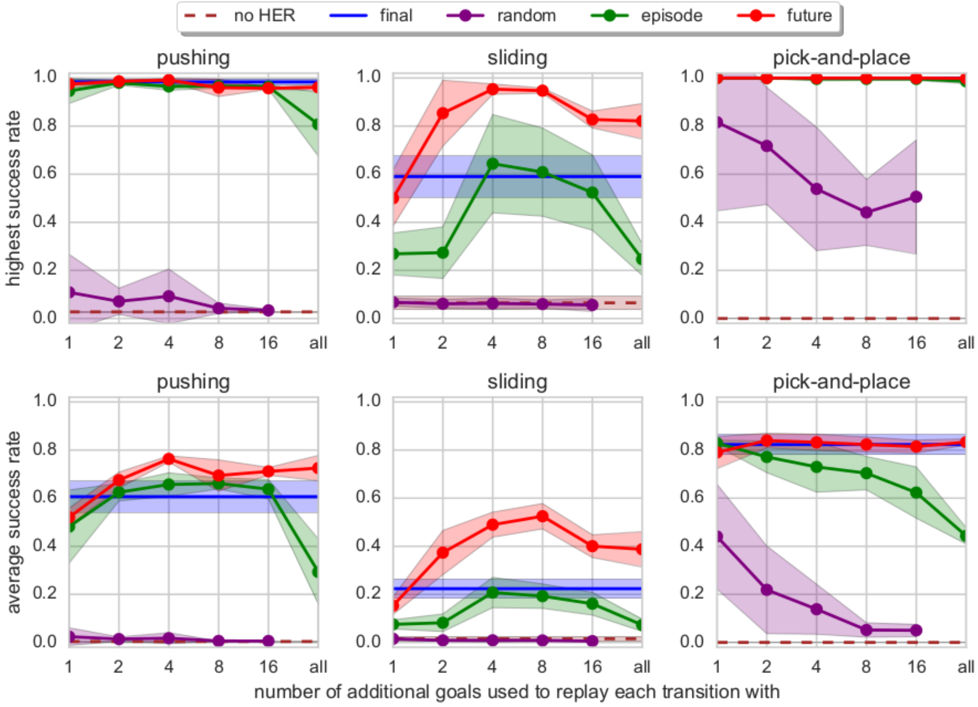


Figure 6: Ablation study of different strategies for choosing additional goals for replay. The top row shows the highest (across the training epochs) test performance and the bottom row shows the average test performance across all training epochs. On the right top plot the curves for `final`, `episode` and `future` coincide as all these strategies achieve perfect performance on this task.

5. Summary

- Try to handle sparse reward
- If the original goal can not be achieved in this episode, set final state as goal
- An implementation of HER in imitation learning: 1709.10089 - Overcoming Exploration in Reinforcement Learning with Demonstrations