HER: Hindsight Experience Replay
AAC: Advantage Actor Critic

Hindsight Experience Replay

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

NeurIPS 2017

https://papers.nips.cc/paper/7090-hindsight-experience-replay.pdf

universal policy:

also conditioned on goals

$$\pi : \mathcal{S} \times \mathcal{G} \rightarrow \mathcal{A}$$

$$r_t = r_g(s_t, a_t)$$

A motivating example: bit-flipping

$$\mathcal{S} = \{0, 1\}^n$$

$$\mathcal{A} = \{0, 1, ..., n-1\}$$

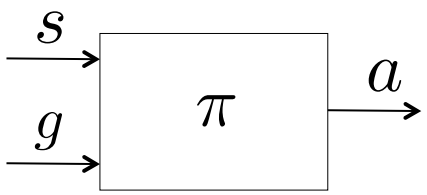
$$a_i : \text{flip bit } i$$

$$r = 1 \text{ if } s = g$$

1 0 0 1 1 1 0

⇩

0 1 0 1 0 1 0



challenge: exponentially difficult to stumble upon reward

initial state and goal are randomly selected for each episode

1 0 0 1 1 1 0 2
1 1 0 1 1 1 0 6
1 1 0 1 1 0 1 4
1 1 0 0 1 0 1

Original transitions

s_t	g	a	s_{t+1}	g	r
1 0 0 1 1 1 0	1 1 1 1 1 1 1	2	1 1 0 1 1 1 0	1 1 1 1 1 1 1	0
1 1 0 1 1 1 0	1 1 1 1 1 1 1	6	1 1 0 1 1 0 1	1 1 1 1 1 1 1	0
1 1 0 1 1 0 1	1 1 1 1 1 1 1	4	1 1 0 0 1 0 1	1 1 1 1 1 1 1	0

Example hindsight replay transition

s_t	g'	a	s_{t+1}	g'	r'
1 0 0 1 1 1 0	1 1 0 1 1 1 0	2	1 1 0 1 1 1 0	1 1 0 1 1 1 0	1

https://www.youtube.com/watch?v=Dz_HuzgMxzo

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

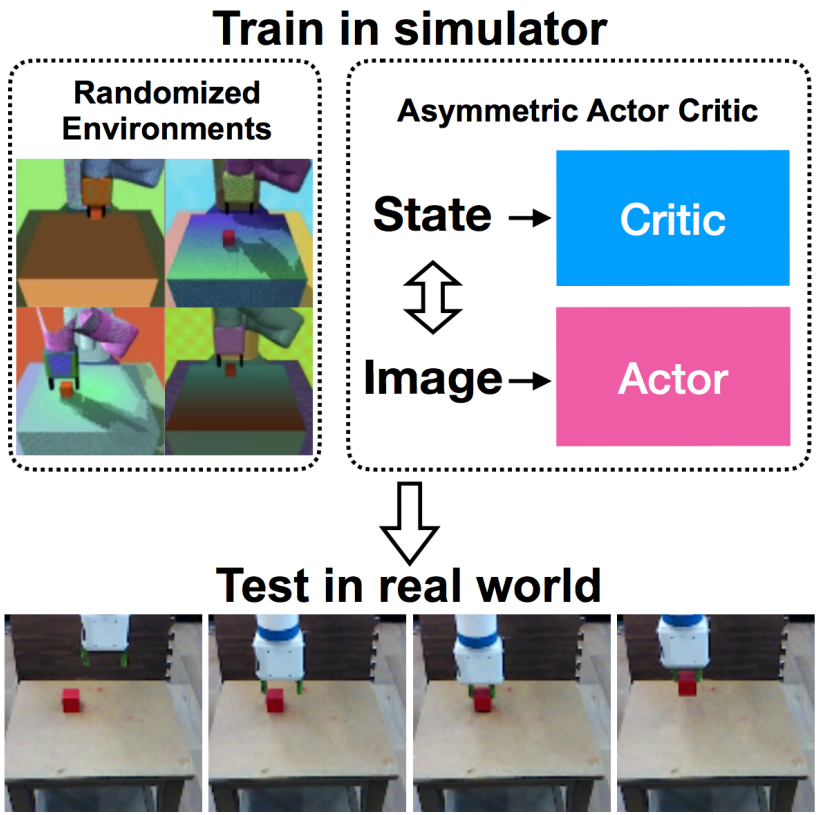
collect an epsiode

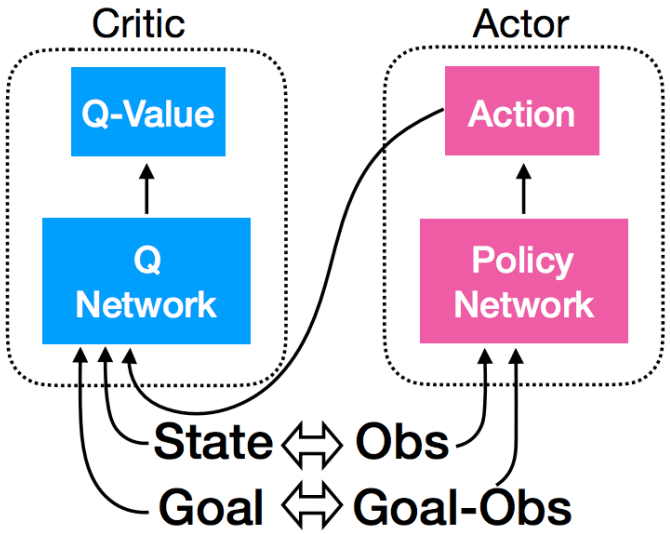
store in replay buffer:
- original experiences
- hindsight experiences

off-policy training

Asymmetric Actor Critic for Image-Based Robot Learning

Lerrel Pinto[1,2] Marcin Andrychowicz[1] Peter Welinder[1] Wojciech Zaremba[1,†] Pieter Abbeel[1,†]





Can be used with any actor-critic algorithm; paper uses DDPG