Hindsight Experience Replay

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

***Abstract*—The paper introduces a technique for efficiently learning in environments with sparse and binary rewards without the need for reward shaping and can be used for real world applications where there is lack of domain-specific knowledge for engineering rewards. The algorithm can be combined with any off-policy algorithm and can be considered as an implicit curriculum learning. Results have been shown using DQN and DDPG in robotic tasks.**

I. Summary of the paper

Motivation - " Learn almost as much from achieving an undesired outcome as from the desired one"

In *multi-goal* RL setup, the goal along with the state is provided as input to the value and policy networks; essentially training it for one task instead of multiple tasks. This approach combined with powerful function approximators (e.g., deep neural networks) allows the agent to learn how to achieve the goal g even if it has never observed it during training. A map is defined $m : \mathcal{S} \rightarrow \mathcal{G}$ s.t $\forall_{s \in \mathcal{S}} f_{m(s)} = 1$ such that the agent aims to achieve any state s that satisfies $f_g(s) = 1$. This gives the flexibility of specifying full or part of a system (ie., $\mathcal{S} = \mathcal{G}$ s.t $f_g(s) = [s = g] \implies m = s$ or
$\mathbb{S} = \mathbb{R}^2, \mathbb{G} = \mathbb{R}, f_g(x, y) = [x = g] \implies m = x$).

The agent is penalised for every time step where the goal is not achieved ($r_g(s, a) = -[f_g(s) = 0]$). In the HER framework, the replay buffer contains all transitions in the episode not just with the original goal but also with a set of other goals. (This shift in goals from simpler to more difficult ones in HER can be viewed as curriculum learning.) Replaying with the goal corresponding to the final state in each episode is referred to as *final strategy*. For choosing goals for HER, the following strategies were also suggested :

- *future* : replay with k random states coming from the same episode as that of the transition being replayed and were observed after it
- *episode* : replay with k random states coming from the same episode as the transition being replayed
- *random* : replay with k random states come across so far in the entire training procedure

It was observed that *future* strategy with $k = 4$ or 8 gave the best results in the experiments conducted where k is the ratio of HER data to data coming from normal experience replay in the replay buffer.

II. Experiments

They used a 7-DOF Fetch Robotics arm which has a two-fingered parallel gripper in all experiments and simulated it using *MuJoCo*. 3 robotic tasks were considered :

- *Pushing* The task is to move the box placed in front of the robot without grasping; using pushing and rolling.
- *Sliding* The task is to hit the puck with enough force to slide and reach the target(out of arm's reach)
- *Pick-and-place* The task is to pick up the target and bring it to a specified location in space using the gripper

For the 3rd task, half of the training episodes were started from a recorded single state in which the box was grasped, to make exploration easier. The states comprises of angles and velocities of all robot joints along with positions, rotations and velocities (linear and angular) of all objects. The goal is defined as $G = \mathbb{R}^3$ and $f_g(s) = [|g - s_{\text{object}}| \leq \epsilon]$ where s_{object} is the position of the object and ϵ is the tolerance. The rewards are sparce and binary ($r(s, a, g) = [f_g(s') = 0]$; s' is the next state). Replay is used with the goal corresponding to the final state in each episode.

The evaluations showed that DDPG without HER is unable to solve any of the tasks while DDPG with count-based exploration made some progress on the sliding task, whereas DDPG with HER solved all the tasks, confirming that HER is required for learning from sparse and binary rewards. It was also noted that HER learns faster if training episodes contain multiple goals. Hence, it is advised to train the agent for a wide variety of goals although only 1 goal is desired.

They also compared the performance of the algorithms in the presence of reward shaping; $r(s, a, g) = \lambda |g s_{\text{object}}|^p - |g - s'_{\text{object}}|^p$ where s' is next state, $\lambda \in \{0, 1\}, p \in \{1, 2\}$. DDPG with and without HER was not able to solve any of the tasks. This suggests the difficulty of shaping rewards; which require a lot of domain knowledge which can be as difficult as manually finding the optimal policy, stressing the necessity for HER to be used with sparse rewards. The poor performance of the algorithms in the presence of reward shaping could be because of hindered exploration due to additional penalization or because of ambiguity in the reward specification(object within some radius of the goal; episode ending) w.r.t to the optimization objective.

The *pick-and-place* task was deployed on a real robotic arm (target positions identified from raw images from the camera, using a CNN) and it was found that the trained policies perform well on the physical robot without the requirement of further finetuning (Gaussian noise added to observations to make it robust to differences between real and simulated environments).

This paper was the first one to introduce an agent that successfully learned to execute complicated tasks using only sparse and binary rewards.

References

[1] Marcin Andrychowicz et al. "Hindsight experience replay". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5048–5058.