Hindsight Experience Replay

Manan Tomar : ED14B023

Abstract

This work (Andrychowicz et al., 2017) proposes a technique for improving exploration in Reinforcement Learning by learning from what happened in "hindsight". The proposed approach is called HER and is applied on top of a model based deep RL algorithm. The experiments are shown both in simulation and in real robot settings.

1. Approach

Having a goal parameterized representation of the policy $\pi(s, g)$ and the Q value function $Q(s, a, g)$ (Schaul et al., 2015), allows us to learn generalized policies for a given goal distribution. HER uses such representations and makes the assumption that the goal can be described as a unique state that can be reached in order to achieve the goal. In the case of manipulation tasks, this goal state can therefore be the 3d coordinate vector of the target.

The basic intuition behind HER is that given a policy fails to achieve the desired goal but achieves another goal instead, the policy learning loop can positively reward such instances if the goal is replaced by whatever the policy actually achieved. HER is therefore applied over a base off policy Reinforcement Learning algorithm, it being Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015) in this case. HER considers sparse reward problems where there is only a binary reward available, i.e. 1 if the task is solved and 0 if not. Such a formulation is essential as engineering a complex dense reward function hurts the generality of an algorithm.

After sampling transitions from the replay buffer for learning, HER replaces a fraction of the sampled goals with goals achieved in the future trajectory states. The authors propose different strategies for goal replacement such as using end of trajectory goal states, randomly selecting from the trajectory, etc.

2. Experiments

The simulation based experiments involve the reaching, pushing, sliding and pick and place tasks using the Fetch robot.

2.1. Reaching, Pushing and Sliding

HER is able to solve these tasks efficiently whereas DDPG fails to solve the latter two. Note that the gripper however always starts over the table and from a single start state. Moreover, for some of the episodes, the object is instantiated at the target location itself, thus providing a richer signal to learn from.

2.2. Pick and Place

For solving this task, the target is sampled on the table for half of the episodes and in the air for the other half. Given this assumption, HER is able to learn pick and place. This is extended to the real robot setting, however the agent starts with the object in the gripper.

3. Critique and Future Work

- As mentioned, it is not discussed how HER performs if the above conditions are not maintained. Moreover, the authors test different strategies for sampling goals, however do not provide a clear justification of why they choose to go with only one of these strategies. In the experiments section, it can be seen that the sampling strategy as a significant effect on the performance as well, thus arguing for a better explanation in choosing one.
- In the initial stages of training when the robot does not hit the object consistently, HER still rewards all actions irrespective of if they reach towards the object or not. This begs the question whether the initial intuition provided to propose HER is valid in such training phases.

References

Andrychowicz, Marcin, Wolski, Filip, Ray, Alex, Schneider, Jonas, Fong, Rachel, Welinder, Peter, McGrew, Bob, Tobin, Josh, Abbeel, OpenAI Pieter, and Zaremba, Wojciech. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pp. 5048–5058, 2017.

Lillicrap, Timothy P, Hunt, Jonathan J, Pritzel, Alexander, Heess, Nicolas, Erez, Tom, Tassa, Yuval, Silver, David, and Wierstra, Daan. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Schaul, Tom, Horgan, Daniel, Gregor, Karol, and Silver, David. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320, 2015.