Goal-conditioned Reinforcement Learning without Goal Rewards

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Vikas Dhiman[1], Shurjo Banerjee[1], Jeffrey M. Siskind[2], Jason J Corso[1]
[1]EECS, University of Michigan, Ann Arbor, MI. [2]ECE, Purdue University, West Lafayette, IN

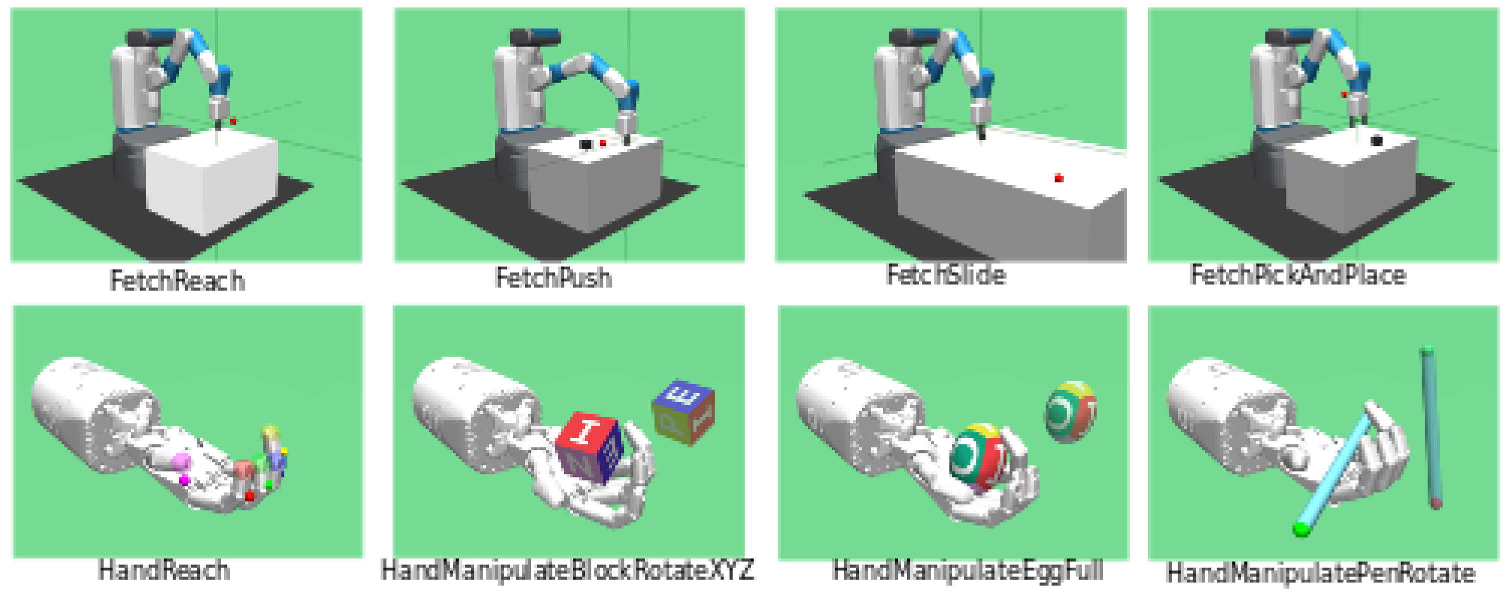
Introduction

Goal-conditioned reinforcement learning (GCRL) addresses tasks where the desired goal can change for every trial. State-of-the-art algorithms model these problems such that the reward formulation depends on the goal rewards, to associate goals with high reward.

This dependence introduces additional goal reward resampling steps in algorithms like Hindsight Experience Replay (HER) that reuse trials in which the agent fails to reach the goal by recomputing rewards as if reached states were psuedo-desired goals.

We propose a reformulation of goal-conditioned value functions for GCRL that yields a similar algorithm, while removing the dependence of reward functions on the goal. Our formulation thus obviates the requirement of reward-resampling that is needed by HER and its extensions.

We also extend a closely related algorithm, Floyd-Warshall Reinforcement Learning, from tabular domains to deep neural networks for use as a baseline. Our results are competitive with HER while substantially improving sampling efficiency in terms of reward samples.



Loss terms

1. DDPG Loss

$$\mathcal{L}_{DDPG}(\theta_Q, \theta_\pi) = \mathbb{E}[(Q_m(s_t, a_t; \theta_Q) - y_t)^2]$$
$$y_t = R(s_t, a_t) + \gamma Q_{\text{tgt}}(s_{t+1}, \pi_{\text{tgt}(s_{t+1}, \theta_\pi); \theta_{Q_{\text{tgt}}}})$$

2. Step Loss

$$\mathcal{L}_{\text{step}}(\theta_Q) = (Q^P_*(s_{l-1}, a_{l1}, g_l; \theta_Q) - R(s_{l-1}, a_{l-1}))$$

3. FWRL Loss upper bound

$$\mathcal{L}_{\text{lo}} = \text{ReLU}[Q_{\text{tgt}}(s_t, a_t, g_w) + Q_{\text{tgt}}(s_w, \pi_t(s_w, g_{t+f}; \theta_\pi), g_{t+f}) - Q_m(s_t, a_t, g_{t+f})]^2$$

4. FWRL Loss lower bound

$$\mathcal{L}_{\text{up}} = \text{ReLU}[Q_m(s_t, a_t, g_w) + Q_{\text{tgt}}(s_w, \pi_t(s_w, g_{t+f}; \theta_\pi), g_{t+f}) - Q_{\text{tgt}}(s_t, a_t, g_{t+f})]^2.$$

Methods

1. Hindsight Experience Replay (HER)[1]: $\mathcal{L}_{DDPG}$
2. Path Reward Reinforcement Learning (Ours): $\mathcal{L}_{DDPG} + \mathcal{L}_{\text{step}}$
3. Floyd-Warshall RL (FWRL [2]): $\mathcal{L}_{DDPG} + \mathcal{L}_{\text{up}} + \mathcal{L}_{\text{lo}}$

Reward formulations

1. With Goal Rewards: R(s, a, g) = (0 if s == g else -1)
2. Without Goal Rewards: R(s, a, g) = -1

Results

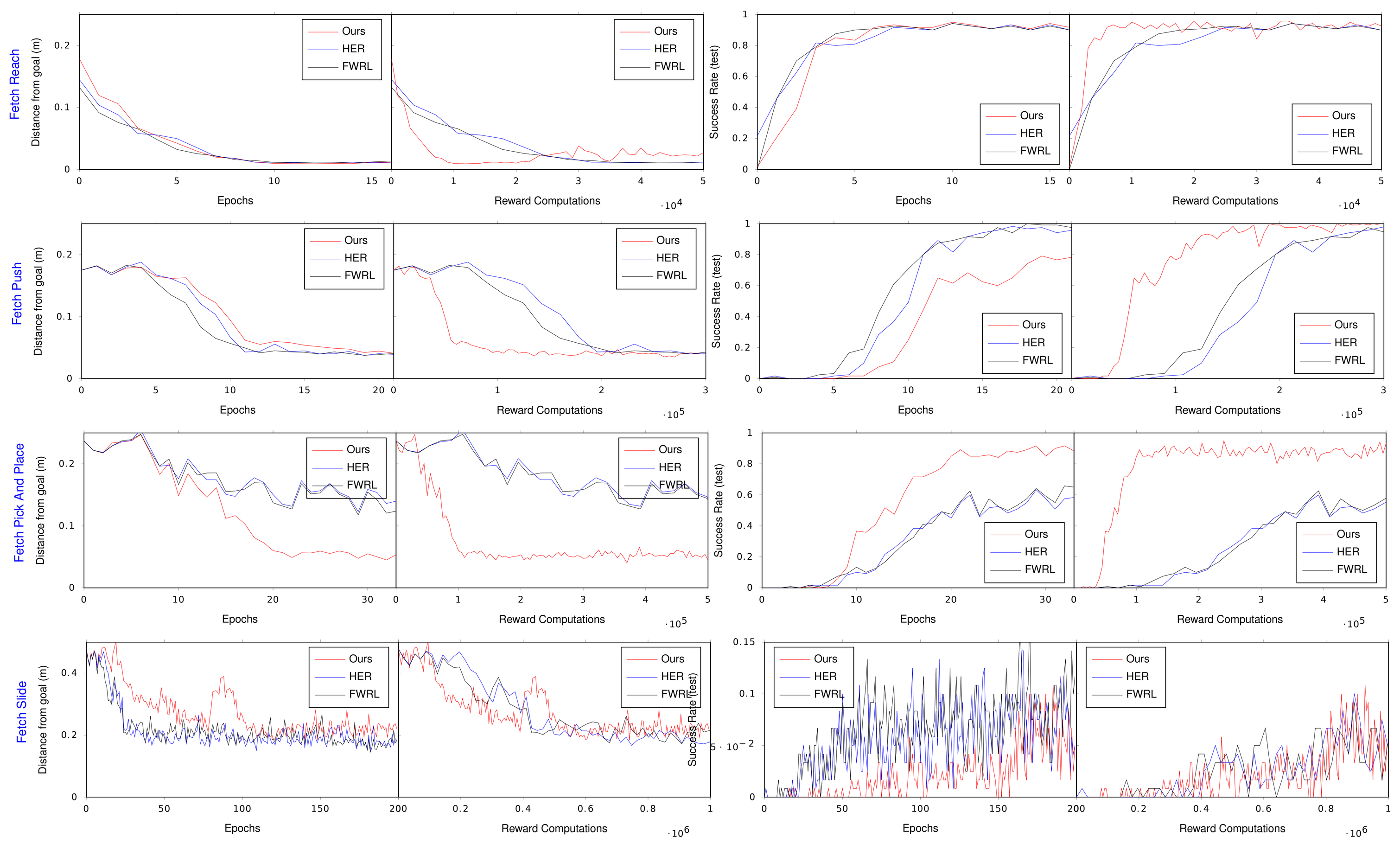


Figure 2: For the Fetch tasks, we compare our method (red) against HER (blue)(Andrychowicz et al., 2016) and FWRL (green) (Kaelbling, 1993) on the distance-from-goal and success rate metrics. Both metrics are plotted against two progress measures: the number of training epochs and the number of reward computations. Except for the Fetch Slide task, we achieve comparable or better. performance across the metrics and progress measures.

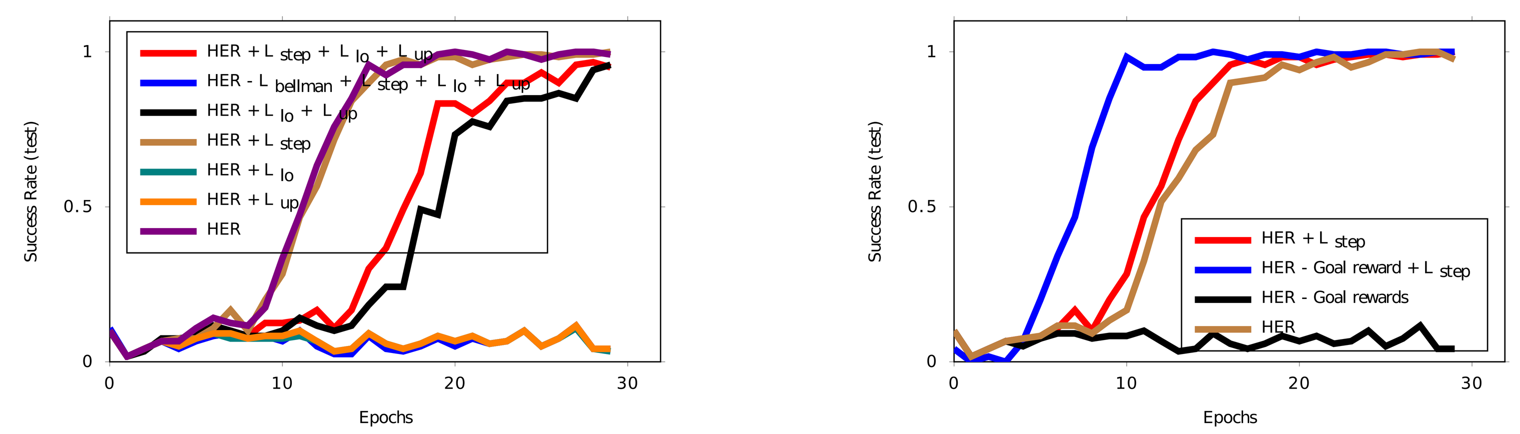
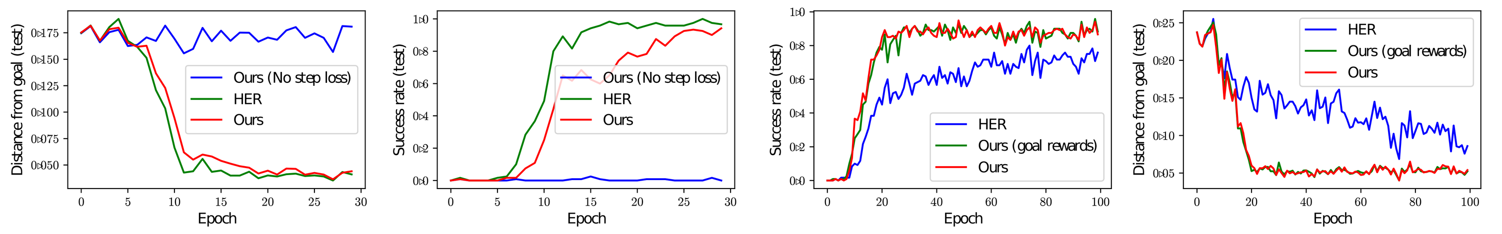
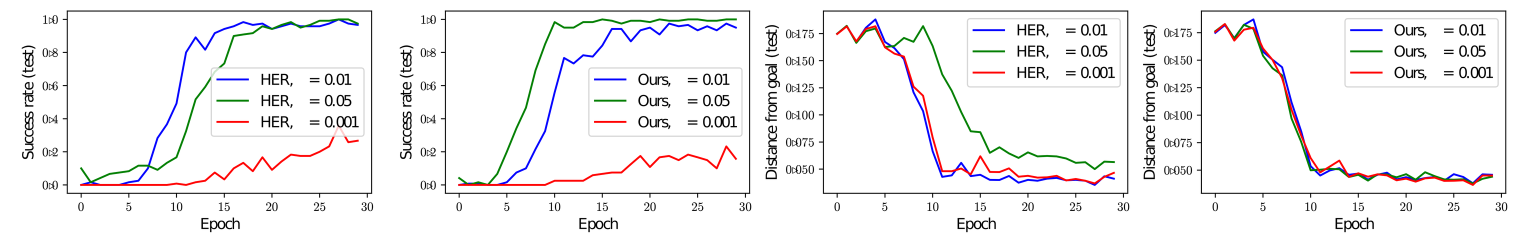


Figure 5 (left) Ablation on loss functions for Fetch Push task. The Floyd-Warshall inspired loss functions L lo and L up do not help much. L step helps a little but only in conjunction with HER [1]. (right) Even when the Goal rewards are removed from HER [1] training, the HER is able to learn only if the L step is added again.



(a) Do we really need the step-loss? (b) Effect of goal-rewards

Figure 3: (a) Effects of removing the step-loss from our methods. Results show that it is a critical component to learning in the absence of goal-rewards. (b) Adding goal-rewards to our algorithm that does have an effect further displaying how they are avoidable.



(a) Sucesss rate (b) Distance from goal

Figure 4: We measure the sensitive of HER and our method to the dsitance-threshold () with respect to the success-rate and distance-from-goal metrics. Both algorithms success-rate is sensitive the threshold while only HER's distance-from-goal is affected by it.

Conclusions

In this work we pose a reinterpretation of goal-conditioned value functions and show that under this paradigm learning is possible in the absence of goal reward. This is a surprising result that runs counter to intuitions that underlie most reinforcement learning algorithms. In future work, we will augment our method to incorporate the distance-threshold information to make the task easier to learn when the threshold is high. We hope that the experiments and results presented in this paper lead to a broader discussion about the assumptions actually required for learning multi-goal tasks.

References

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

[2] Leslie Pack Kaelbling. Learning to achieve goals. In IJCAI, pp. 1094–1099. Citeseer, 1993.