Exploration via Hindsight Goal Generation

Zhizhou Ren, Kefan Dong, Yuan Zhou, Qiang Liu, Jian Peng

October 23, 2019

Sparse Reward Robotic Manipulation Tasks

- Multi-goal Reinforcement Learning
- Challenges
  - sparse indicator reward
  - reward shaping causes suboptimal behavior
- Advance Approaches
  - Hindsight Experience Replay (HER)
  - Automatic Curriculum Generation

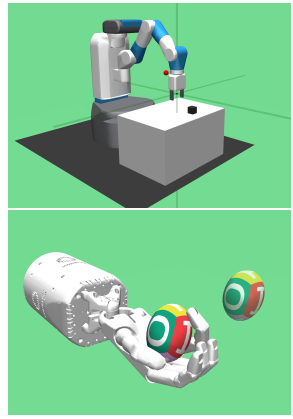


Figure 1: Tasks from OpenAI Gym

Hindsight Experience Replay

- Embedding in off-policy RL algorithm
- Standard experience replay

$$\mathcal{B} = \{(s_t, g, a_t, r_t, s_{t+1})\}$$

- Hindsight experience replay

$$\begin{aligned} \mathcal{B}^H &= \{(s_t, g', a_t, r'_t, s_{t+1})\} \\ s.t. \quad & g' = \phi(s_{t+k}) \\ & r' = R_{g'}(s_t, a_t, s_{t+1}) \end{aligned}$$

  - replay achieved imaginary goals

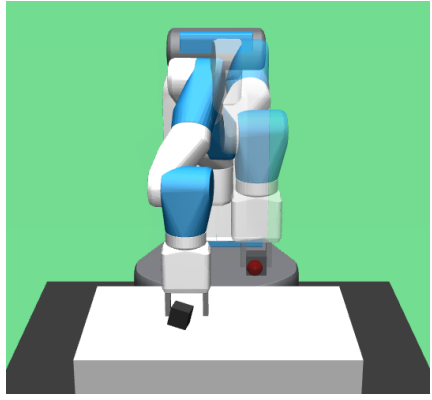


Figure 2: Hindsight goal

Motivation: Hindsight Goal Generation as Automatic Curriculum

Hindsight Experience Replay

- enrich reward signal
- replay past success tasks
- imaginary goals for exploitation
- easy to implement

Automatic Curriculum Generation

- self-paced training
- evaluate current policy
- imaginary goals for exploration
- using GAN or hand-craft features

Common Assumption:
The learned policy has generalizability to similar tasks.

Characterization of Policy Generalizability

Generalizability of Value Function

Assume that the value function $V^\pi(s, g)$ satisfies the Lipschitz continuity:

$$|V^\pi(s, g) - V^\pi(s', g')| \leq L \cdot d((s, g), (s', g'))$$

where $d(\cdot, \cdot)$ is a predefined metric, e.g.

$$d((s, g), (s', g')) = c\|\phi(s) - \phi(s')\|_2 + \|g - g'\|_2$$

in which $\phi(\cdot)$ is a state abstraction.

- For two task distributions $\mathcal{T}, \mathcal{T}^*$,

$$V^\pi(\mathcal{T}^*) \geq V^\pi(\mathcal{T}) - L \cdot D(\mathcal{T}, \mathcal{T}^*)$$

Optimize Surrogate Lower Bound

- Alternatively optimize $\pi, \mathcal{T}$,

$$\max_{\pi, \mathcal{T}} \; V^{\pi}(\mathcal{T}) - L \cdot D(\mathcal{T}, \mathcal{T}^*)$$

- $\mathcal{T}$ is a high-level controller to guide exploration
- Apply hindsight heuristics to select $\mathcal{T}$
  - construct intermediate task distribution $\mathcal{T}$ around hindsight goals

$$\text{supp}(\mathcal{T}) \subseteq \mathcal{B}^H$$

    where $\mathcal{B}^H$ is the replay buffer $\mathcal{B}$ with hindsight experience replay
  - reduce to simple combinatorial optimization problem

Alternative Optimization Framework

Algorithm 1 Exploration via Hindsight Goal Generation (HGG)

Initialize π

for iteration $= 1, 2, \ldots, N$ **do**

  Sample $\widehat{\mathcal{T}}^* = \{(\hat{s}_0^i, \hat{g}^i)\}_{i=1}^K \sim \mathcal{T}^*$

  Select K hindsight task instances to construct $\mathcal{T}$

$$\max_{\mathcal{T}} V^\pi(\mathcal{T}) - L \cdot D(\mathcal{T}, \widehat{\mathcal{T}}^*)$$
$$s.t. \;\; \text{supp}(\mathcal{T}) \subseteq \mathcal{B}^H$$

  Collect K trajectories $\{\tau_i\}_{i=1}^K \sim \mathcal{T} \times \pi$

  Store $\{\tau_i\}$ into replay buffer $\mathcal{B}$

  Perform minibatch update on value and policy network

Request of Regularizer

- How to construct intermediate $\mathcal{T}$?

$$\max_{\mathcal{T}} V^{\pi}(\mathcal{T}) - L \cdot D(\mathcal{T}, \widehat{\mathcal{T}}^{*})$$
$$s.t. \;\; \text{supp}(\mathcal{T}) \subseteq \mathcal{B}^{H}$$

- Use a greedy solver?
  - Simply select best instances in the replay buffer $\mathcal{B}^{H}$
- Easy to be stuck
  - value estimation is noisy
  - task-specific bad examples

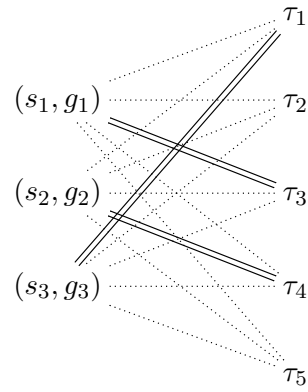
Figure 3: A noisy trajectory with high estimated value

Additional Constraint on Diversity

$(s_i, g_i) \in \widehat{\mathcal{T}}^*$ $\tau_j \in \mathcal{B}$

- Select K task instances from distinct trajectories
- Maximum Weight Bipartite Matching (MWBM)
- weight: $w((s_i, g_i), \tau_j = \{s_k^{\tau_j}\}_{k=0}^{T})$

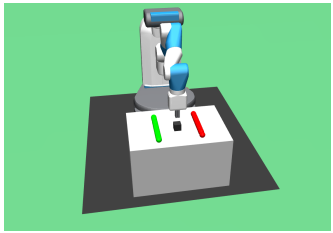
  $= \max_k V^{\pi}(s_i, \phi(s_k^{\tau_j})) - d((s_i, g_i), (s_0^{\tau_j}, \phi(s_k^{\tau_j})))$
- construct $\mathcal{T}$
  - immediate initial states s_i
  - hindsight goals $g_i \leftarrow \phi(s_k^{\tau_j})$

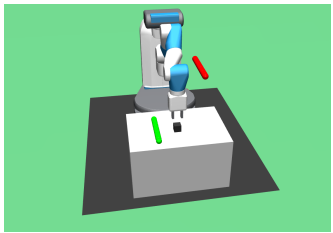


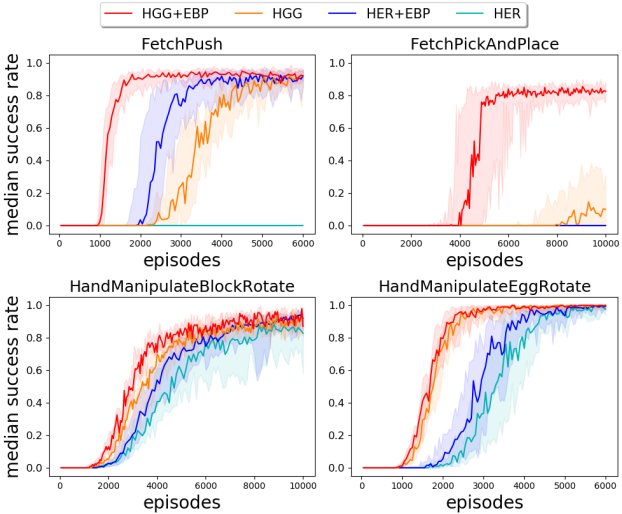
Experiment: Visualization

Figure 4: HGG vs. HER

Experiment: Improvement on Sample Efficiency







Thank you!