Pre-Training Goal-based Models for Sample-Efficient Reinforcement Learning

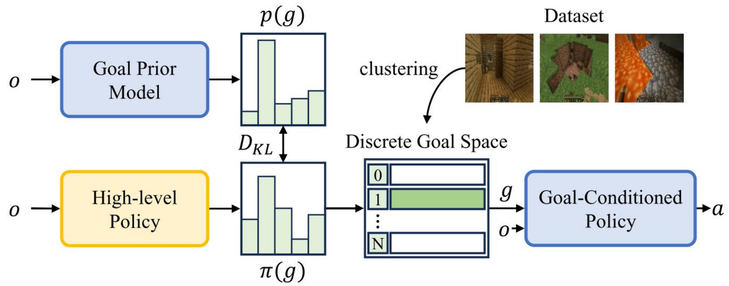
Haoqi Yuan, Zhancun Mu, Feiyang Xie, Zongqing Lu
Peking University, Beijing Academy of Artificial Intelligence

Introduction

- Pre-training on **task-agnostic datasets** can accelerate RL for downstream tasks. One promising approach is **pre-training low-level skills** to provide temporal abstractions.
- Existing methods have not yet **scaled** to high-dimensional, complex open-world environments or large datasets:
  (1) Fail to accurately generate long action sequences in large action spaces.
  (2) Downstream RL in continuous latent action spaces tends to be inefficient.

Method

- **Motivation**: recent works show that goal-conditioned behavior cloning on large datasets can effectively model diverse skills in challenging open-world environments. We propose **PTGM** that pre-trains such goal-based models to accelerate downstream RL.



- **Pre-Training a Goal-Conditioned Policy**
  - **Hindsight relabeling**: for a k-step subsequence $\tau = (s_t, a_t, \dots, s_{t+k}, a_{t+k})$, we label each sample $(s_i, a_i), i = t, \dots, t+k$ with the goal state $s^g = s_{t+k}$.
  - **Goal-conditioned behavior cloning**

$$\mathcal{L}(\phi) = \mathbb{E}_D \left[-\log P_\phi(a_i | s_i, s^g) \right]$$

- **Clustering in the Goal Space**
  - High-dimensional, continuous goals $s^g \in S$ as high-level actions make RL inefficient.
  - We use t-SNE and K-Means to cluster goals in the dataset and use the N cluster centers to represent all goals.
  - Downstream RL uses a discrete action space $A^h = [N]$.

- **The Goal Prior Model**
  - We propose learning a prior about 'how to select the goal' to effectively guide the high-level RL policy.

$$a^h = \arg\max_{i \in [N]} \left(\frac{s_i^g \cdot s^g}{\|s_i^g\| \cdot \|s^g\|} \right)$$

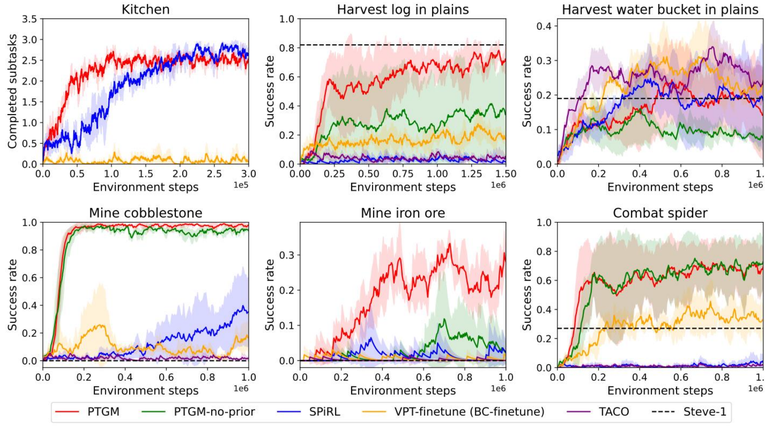
$$\mathcal{L}(\psi) = \mathbb{E}_D \left[-\log \pi_\psi^p(a^h | s_t) \right]$$

- **Downstream RL**
  - Train a high-level policy $\pi_\theta(a^h|s)$ to maximize a combination of the task return and a goal prior regularization term.

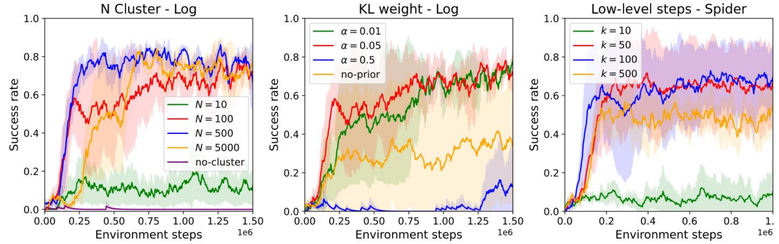
$$J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \left(\sum_{i=kt}^{(k+1)t} R(s_i, a_i) - \alpha D_{\mathrm{KL}} \left(\pi_\psi^p(a^h|s_{kt}) \| \pi_\theta(a^h|s_{kt}) \right) \right) \right]$$

Results

- **Kitchen and Minecraft tasks**:
  - PTGM outperforms baselines (SPiRL, TACO, BC-finetune) in sample efficiency and task performance.
  - PTGM does not suffer from forgetting in long-horizon tasks (Iron ore).
  - PTGM enhances the capabilities of its low-level policy, Steve-1.

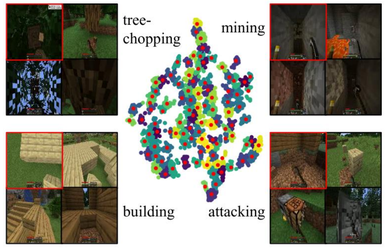


- **Ablation study**:
  - PTGM maintains stable performance as the number of clusters increases.
  - The goal prior model stabilizes RL training.
  - Temporal abstraction improves sample efficiency.



Qualitative Study

- Visualization of the goal clusters in Minecraft: each cluster represents an interpretable behavior of human players. Samples within the same cluster exhibit similar behavior.



- Capacity of the discrete goal space: a goal can induce varying behaviors depending on the context.

Test task	Sheep	Pig	Chicken
Success rate	0.82	0.36	0.94
Test task	Place	Water	Wool
Success rate	0.65	0.16	0.44

Summary

- PTGM holds advantages in the sample efficiency, learning stability, interpretability, and generalization of the low-level skills.
- The proposed clustering and goal prior techniques improve sample efficiency.
- Experiments demonstrate PTGM's capability to learn on diverse domains and solve the challenging Minecraft tasks efficiently.

Acknowledgment

This work was supported by NSFC under grant 62250068.