Efficient Skill Transfer for Simulated Agents

Ademi Adeniji
ademi@stanford.edu
Stanford University Department of Computer Science

Overview

Problem

- One major roadblock in the way of achieving general purpose robots is task generality of learned representations
- In this project I investigate how **learned high-level skills can be transferred between similar tasks to accelerate learning of new tasks**

Significance

- Intelligent robots should exhibit increasingly flexible behaviors with the **same or more facility as humans**
- Transform manufacturing, environmental cleanup, search and rescue, food and drug delivery, elderly care, etc.

Existing Approaches

- Model Agnostic Meta-Learning (MAML)
  - General purpose approach for **fast adaptation of deep neural networks**
  - Attempts to find a parameter assignment in the **observed space** that is a few gradient steps away from optimal on multiple tasks
  - **Does not use latent representations**

Setup

- MuJoCo simulated HalfCheetah-v2 agent with fully observable states as well as a **state-action based reward specification**
- **Input:** Randomly initialized policy ***π_rand*** for some base environment ***E_pre***
- **Output:** Learned policy ***π**** for target environment ***E_tar***
- **Model-free, off-policy** learning on {s, a, s', r} transitions

Task

*Learn **π*** that achieves high reward on E_tar given learned information from E_pre*

Approach

Project Phases

1. Deep RL Algorithms - Provide model baseline and **learn base policy on *E_pre***
2. Naive Transfer - Initial approach that **fine-tunes base policy** from *E_pre* on *E_tar*
3. Latent Soft Actor-Critic - Latent transfer using **conditional variational autoencoder**

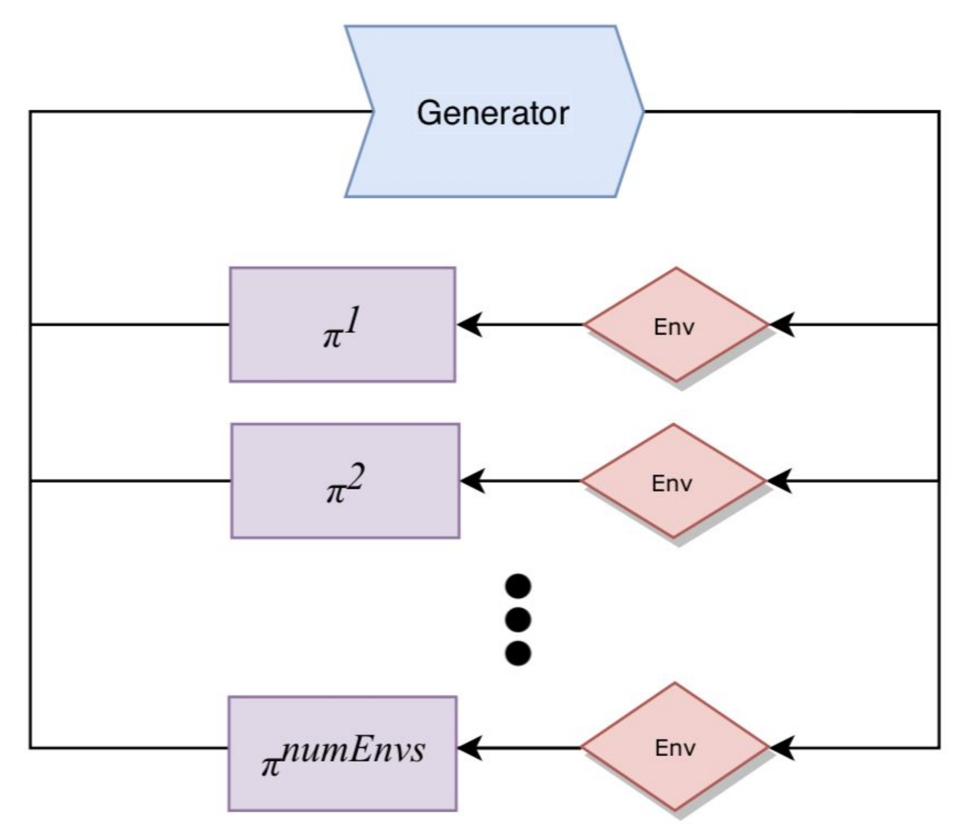

Figure 1. Multi-Task Framework. Action generator trained to **capture high-level skills general to multiple environments**. Policies trained to output environment-specific low-level dynamics

Training

$$\pi^* = argmax_\pi E_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t(R(s_t, a_t, s_{t+1}) + \alpha H(\pi(s_t)))]$$

$$\log p(a|s;\theta_g) \geq E_{\pi(z|s,a;\phi_g)}[\log g(a|s,z;\theta_g)] - D_{KL}(\pi(z|s,a;\phi_g)||p(z)) = J$$

Figure 2. Soft-Actor Critic with Variational Inference. Train SAC to output latent codes instead of actions. Separately train action generator to map latent codes to actions in the environment using **standard conditional variational inference**

Inference

$$z \sim \pi(z|s,a;\phi_g)$$
$$a \sim g(a|s,z;\theta_g)$$

Figure 3. Ancestral Sampling. To generate actions at test time, sample a latent code z from the learned latent distribution and then sample an action a from the learned action generator conditioned on z

Results and Analysis

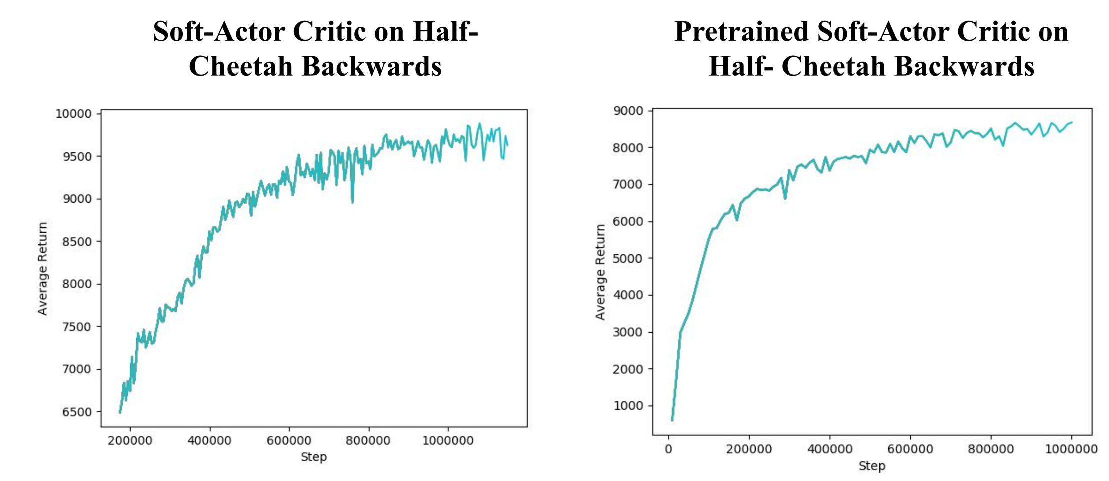

Figure 4. Naive Transfer Results. Naively transferring the pretrained SAC policy from the forwards to backwards task **does not yield efficiency benefits**

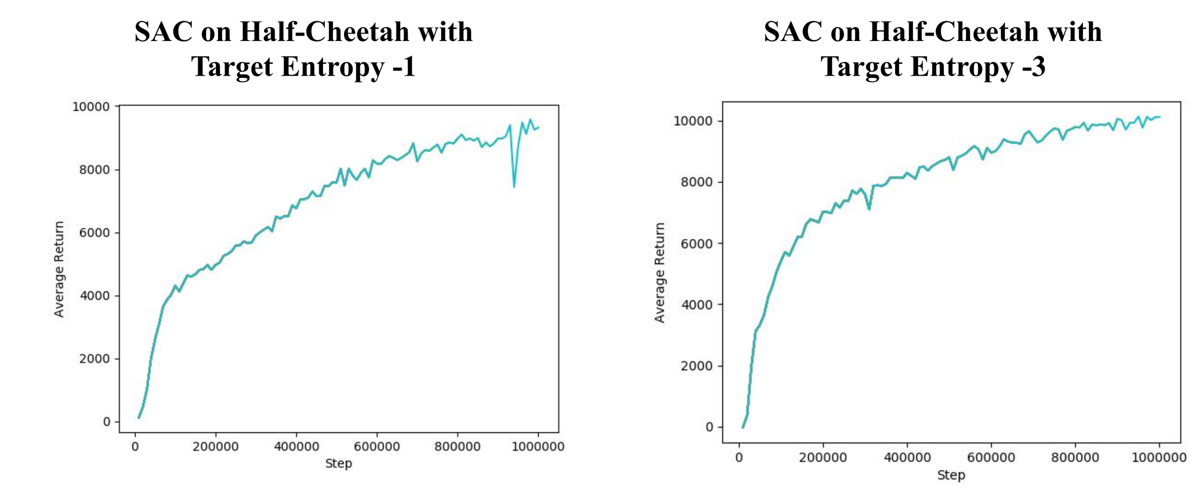

Figure 6. Target Entropy Ablation. Modifying the target entropy from its default value of -6 on the Half-Cheetah environment does not change performance in a significant way

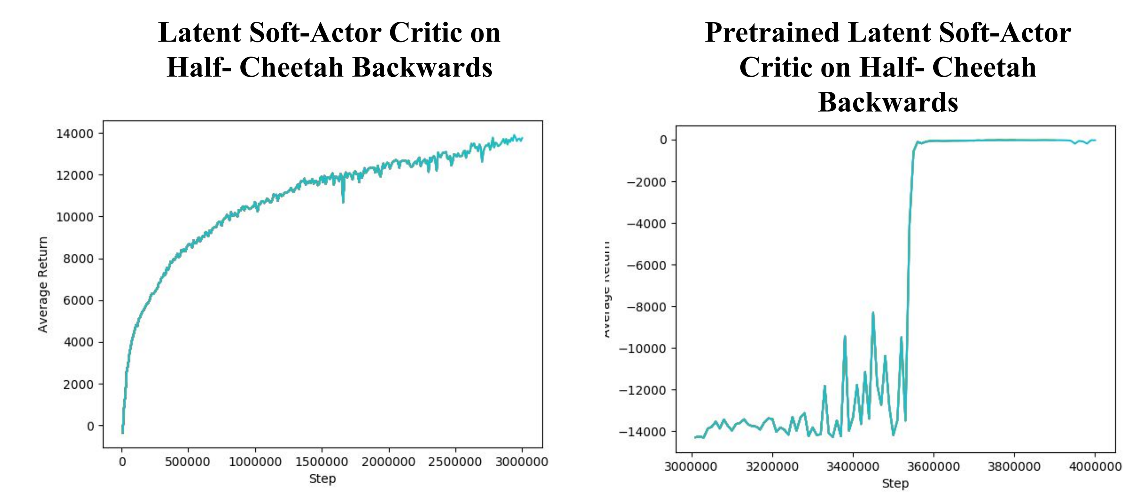

Figure 5. Latent Transfer Results. The latent SAC model **is not able to transfer the learned representations from the forwards running task**. Directly training the target backwards tasks yields much better performance

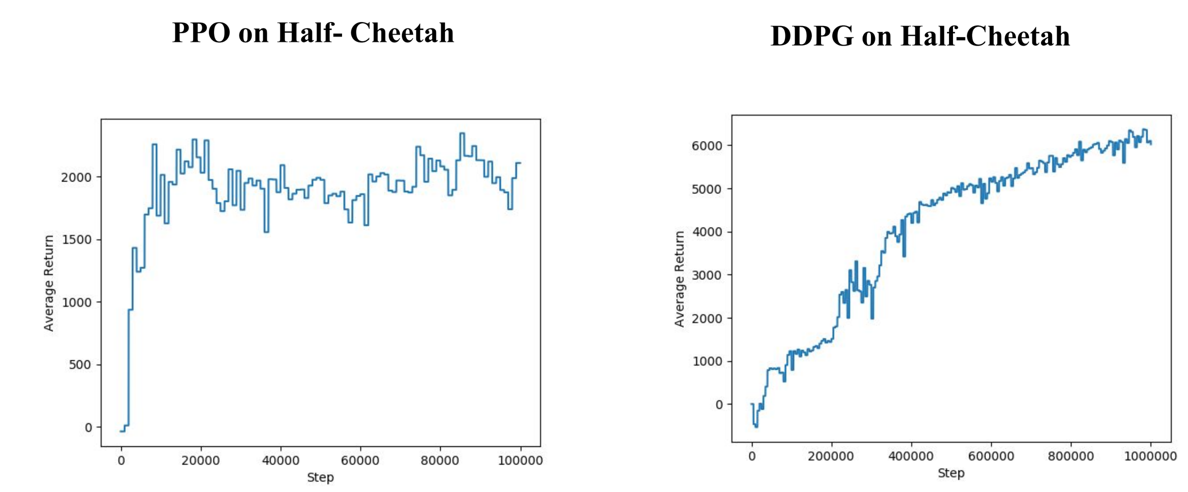

Figure 7. Deep RL Algorithms. Other deep RL algorithms such as Deep Deterministic Policy Gradients and Proximal Policy Optimization underperform SAC

Conclusion

- From our quantitative results, we conclude that our **the latent Soft Actor-Critic Model is unable to efficiently transfer learned representations** as evidenced by its slow learning on the new task
- The latent model's ability to achieve high-rewards of up to 14k on the primary task suggest **that further modifications to the architecture or hyperparameters may yield better results**
- We were impressed by the naive model's **ability to relatively quickly readjust to the new task specification** with an old policy and achieve reasonable reward
- We were surprised to find that changes to the target entropy parameter had such a small effect on performance

Future Work

- Further ablation with the **target entropy parameters** as well as the **variational prior** may improve performance
- Model Agnostic Meta-Learning in a latent space may be a better suited approach for skill transfer learning
- Other **more expressive latent variable models** such as normalizing flows or energy based models may allow for better low-level representations

References

[1] Dynamics Learning with Cascaded Variational Inference for Multi-Step Manipulation
[2] Deep Compositional Question Answering with Neural Module Networks
[3] Learning Synergies between Pushing and Grasping with Self-supervised Deep Reinforcement Learning
[4] Learning Task-Oriented Grasping for Tool Manipulation from Simulated Self-Supervision
[5] Mechanical Search: Multi-Step Retrieval of a Target Object Occluded by Clutter
[6] Deep Visual Foresight for Planning Robot Motion
[7] Learning Sampling Distributions for robot motion planning
[8] Self-consistent Trajectory Autoencoder: Hierarchical Reinforcement Learning with Trajectory Embeddings
[9] Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks