Latent Skill Planning for Exploration and Transfer

Kevin Xie*, Homanga Bharadhwaj*, Danijar Hafner, Animesh Garg, Florian Shkurti

University of Toronto
Vector Institute
Google Brain
University of Toronto Robotics Institute

Motivation

- Quickly solve new tasks in complex environments
- Agents need to build up reusable knowledge
- Learned world model captures environment
- Skills capture reusable behaviour

Overall Research Question

How can we efficiently integrate reusable skill learning with learned world models for control?

Previous works

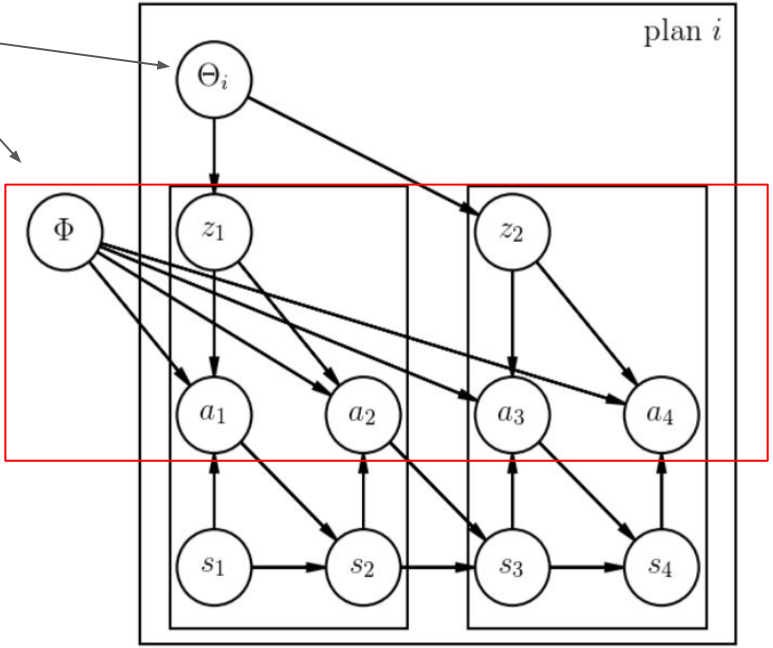
- Learned world models for control
  - observation space (Chua et al., 2018; Sharma et al., 2019; Wang & Ba, 2019) or in a
  - latent space (Hafner et al., 2019, 2018).
- Online planning methods
  - f.e. PETS (Chua et al., 2018)
  - only learn the dynamics (and reward) model
  - online search f.e. Cross-Entropy Method (CEM; Rubinstein, 1997) for control
- Amortized policy methods
  - f.e. Dreamer (Hafner et al., 2019)
  - train a reactive policy with many imagined rollouts

Previous Works

- Amortized policy method benefits
  - Improve with experience
  - Policies execute faster
- On the other hand, poor generalization with a different reward function.

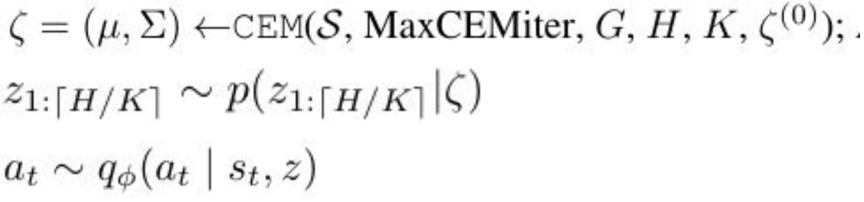
Partial Amortization: Planning over Learned Skills

- Plan latent skills z
- Learn skill-conditioned policy



LSP (Latent Skill Planning)

- Skills sampled from CEM planning distribution
- Low level policy conditioned on skills
- Train with imagination rollouts



$$\zeta = (\mu, \Sigma) \leftarrow \mathrm{CEM}(\mathcal{S}, \mathrm{MaxCEMiter}, G, H, K, \zeta^{(0)});$$

$$z_{1:\lceil H/K \rceil} \sim p(z_{1:\lceil H/K \rceil} | \zeta)$$

$$a_t \sim q_\phi(a_t \mid s_t, z)$$

Mutual Information Skill Objective

- Incentivizes the skill-conditioned policy to pay attention to skill variable
- Make skills predictable given resulting trajectory

$$\mathcal{MI}(z, \{s\}|s_0) \geq \int p(z, \{s\}, s_0) \log \frac{q(z|\{s\}, s_0)}{p(z|s_0)}$$

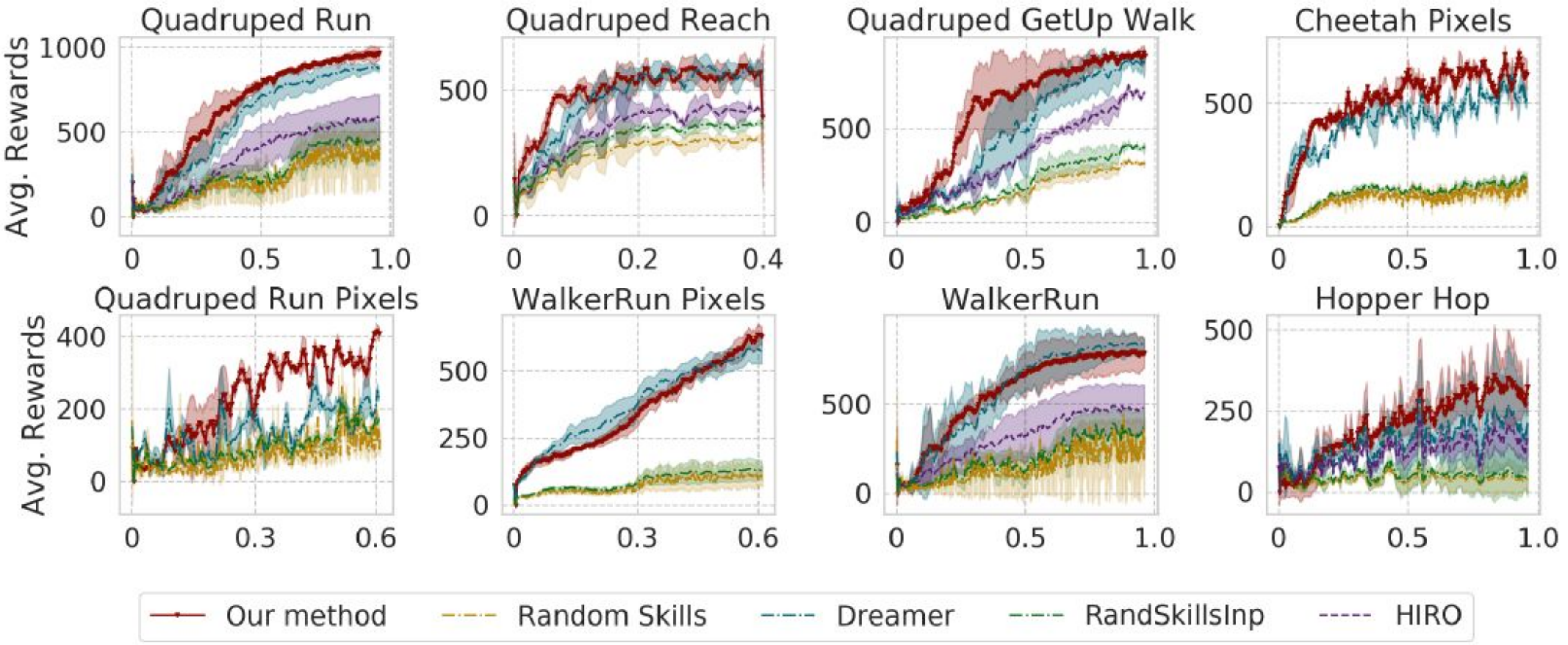
$$= \mathbb{E}_{p(z,\{s\},s_0)}[\log q(z|\{s\}, s_0)] + \mathbb{E}_{s_0}[\mathcal{H}[p(z|s_0)]]$$

Experiments

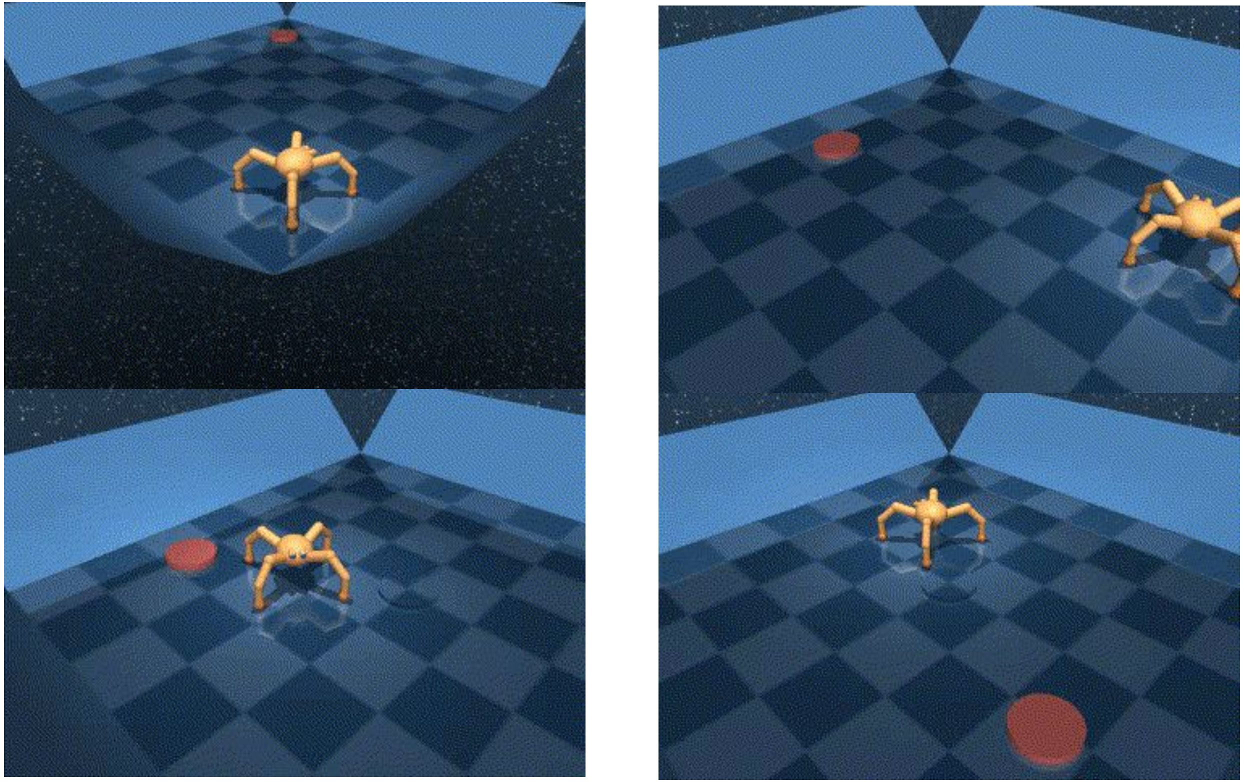
We consider the following baselines

- Dreamer
- HIRO (Nachum et al., 2018)
    - hierarchical RL, high level policy
- Random Skills
    - LSP but with random skills

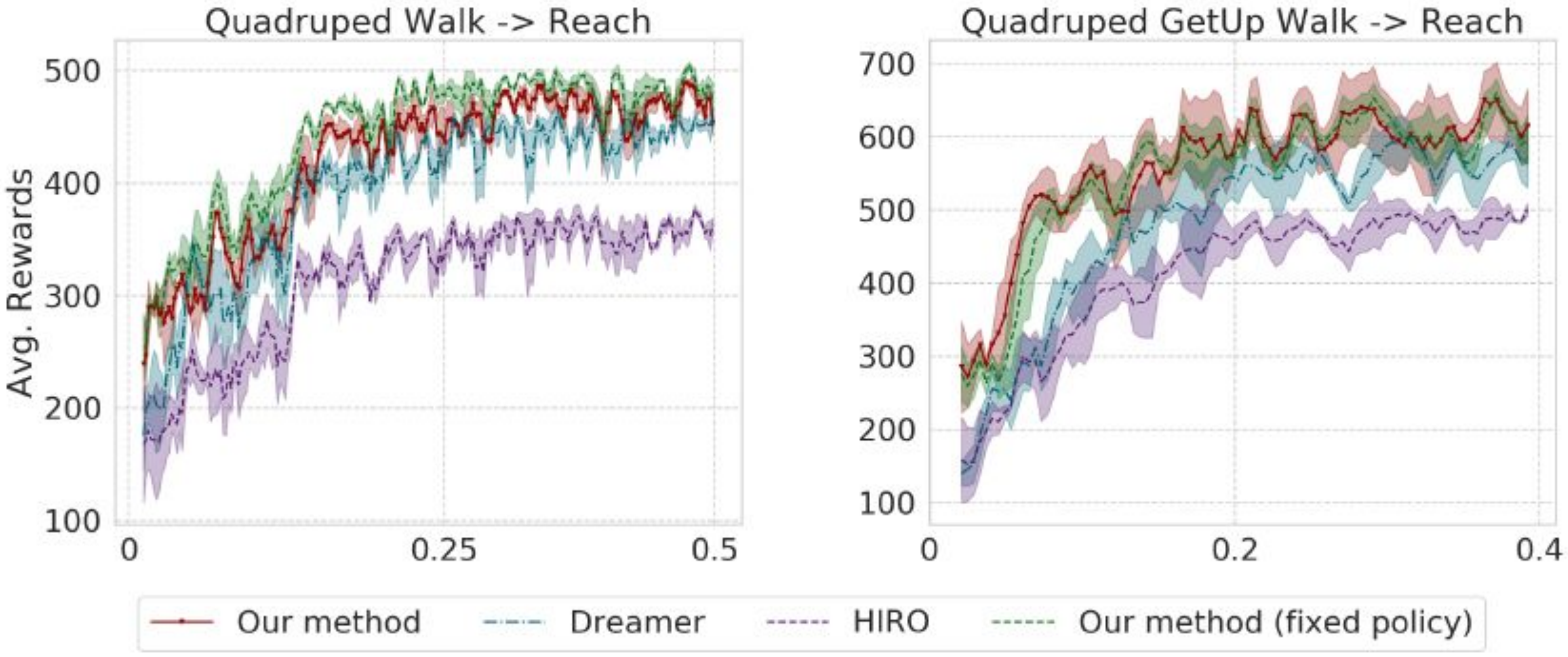
Results - Single task performance



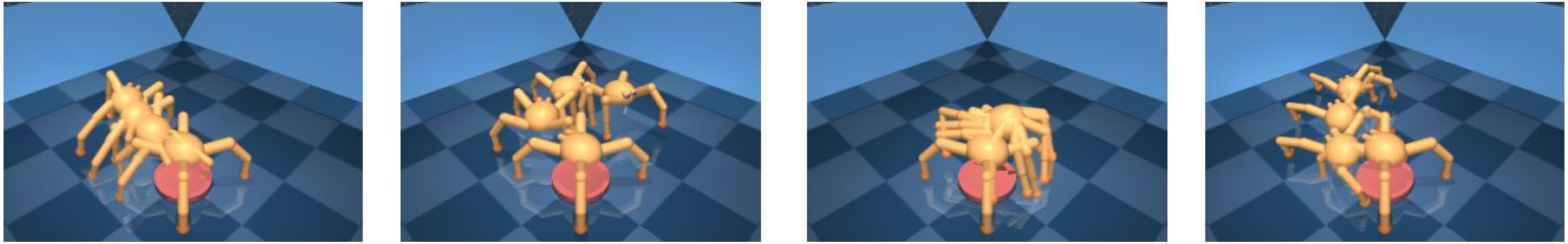
Transfer from Quadruped Walk to Quadruped Reach



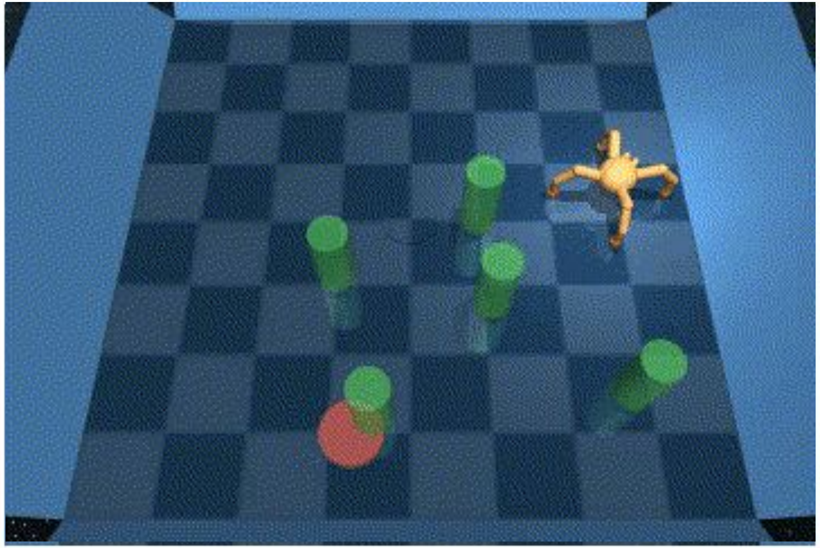
Results - Transfer performance

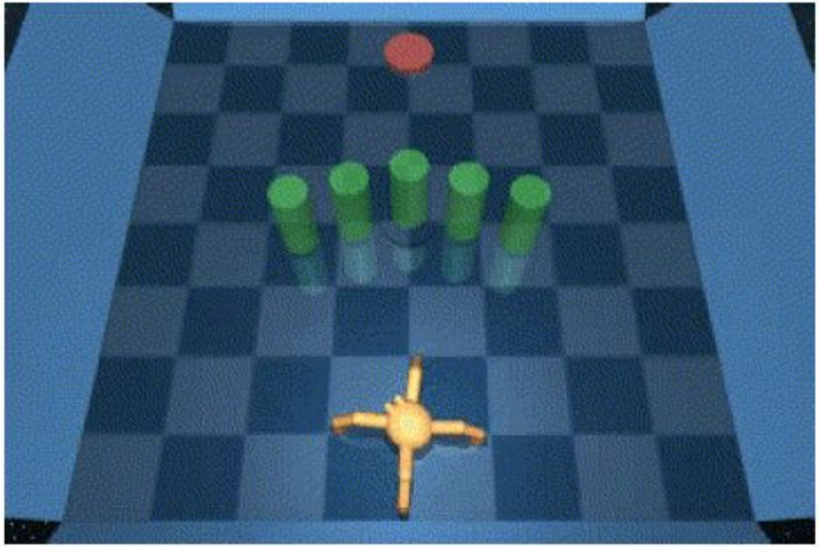


Results - Skills visualization for Quadruped Walk

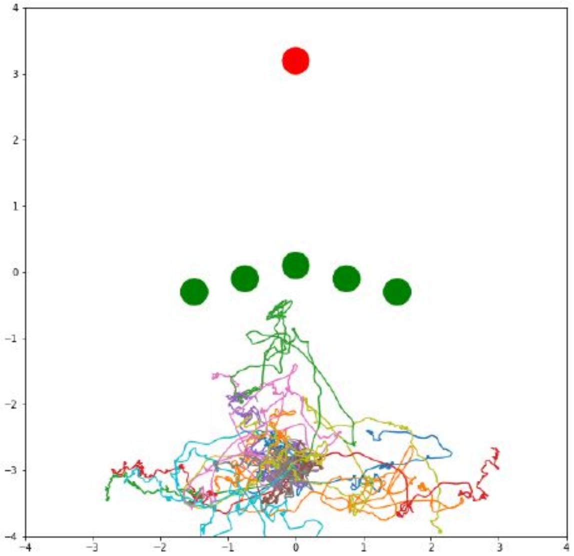


Results - Quadruped reach with obstacles

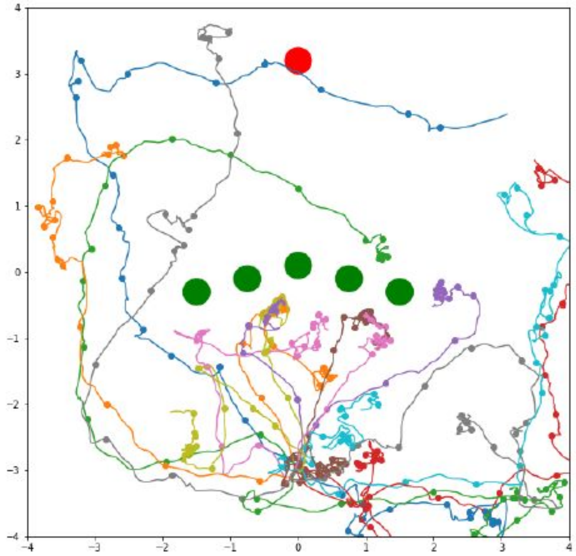




Results - Quadruped reach with obstacles

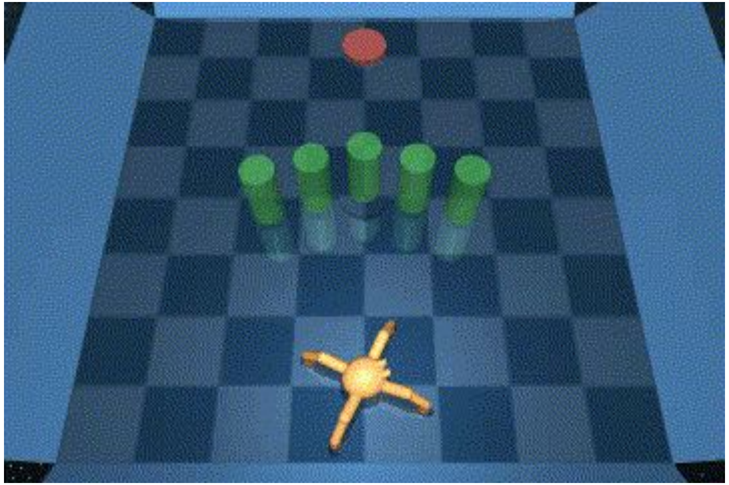


Dreamer

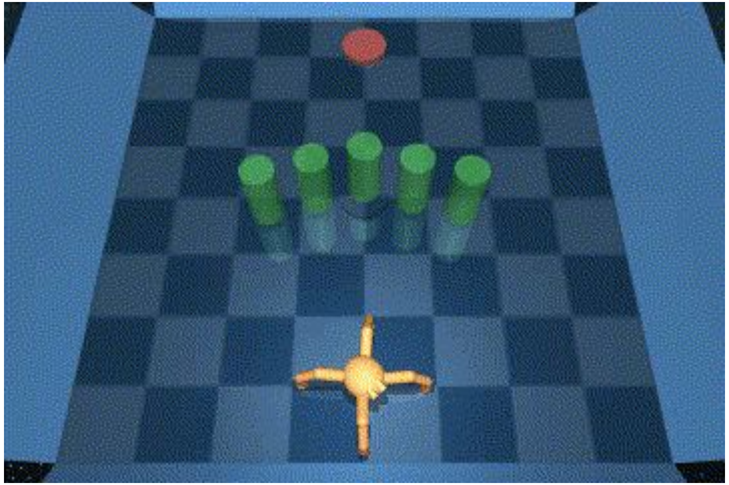


LSP (fixed policy)

Models after 80 episodes on sparse target task



Dreamer



LSP (fixed policy)

Conclusion

We present LSP a Model-based RL method that combines online action planning and amortized policy optimization through learning temporally extended skills with mutual information.

We demonstrate improved sample efficiency and exploration in single task and task transfer settings.

Thank you

Please check our paper for more details