Intrinsic Motivation & RL

Introduction

- Extrinsically Motivated Behavior

Behavior undertaken to achieve some externally supplied reward.

eg: a prize, a high grade, or a high-paying job

- Intrinsically Motivated Behavior

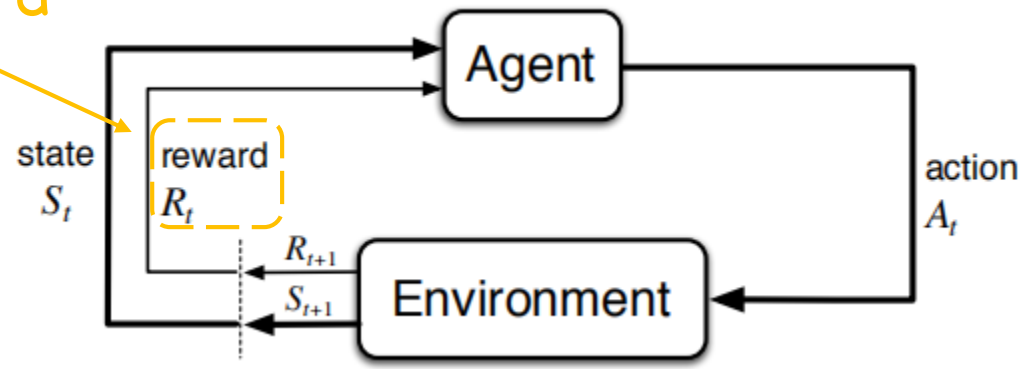
Behavior done for its own sake.

eg: interest, curiosity

Introduction

- MDP $\langle S, A, P, R, \gamma \rangle$

extrinsic reward



- S the set of possible states;
- A the set of possible actions;
- P the transition function $P : S \times A \times S \to \mathbb{P}(S'|S, A)$
- R the reward function $R : S \times S \times A \to \mathbb{R}$;
- $\gamma \in [0, 1]$ the discount factor;

goal

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, s_{t+1}, \pi(s_t))\right]$$

- Intrinsic Motivation (IM) & RL ➡ exploration policy

Rewards extrinsic to the agent are extremely sparse, or absent altogether. In such cases, how to enable the agent to explore its environment and learn skills that might be useful later in its life.

Introduction

- Exploration Policy

1. Curiosity-driven

   ICM [Pathak et al., 2017]

2. Learn skills by exploring

   DIAYN [Eysenbach et al., 2019]

Curiosity-driven Exploration by Self-supervised Prediction

Deepak Pathak [1] Pulkit Agrawal [1] Alexei A. Efros [1] Trevor Darrell [1]

ICML 2017

Intrinsic Curiosity Module (ICM)

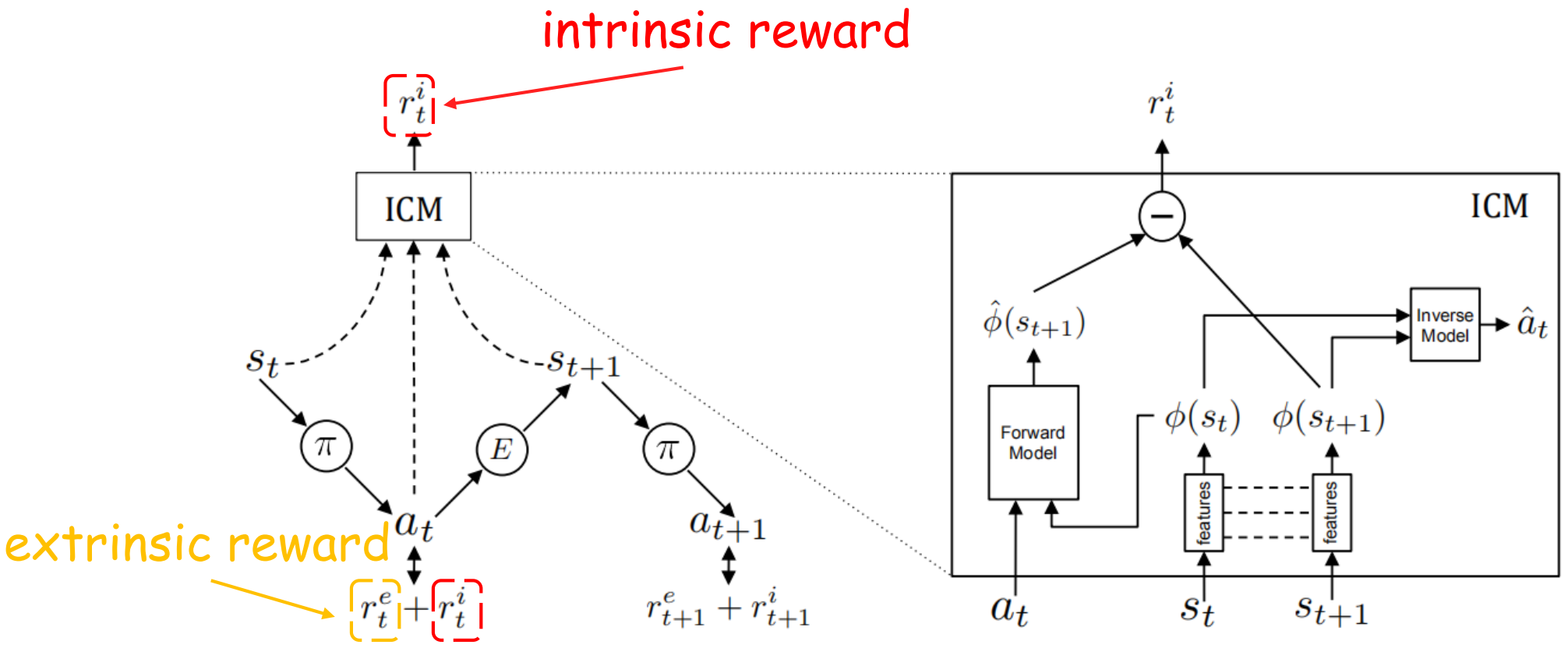
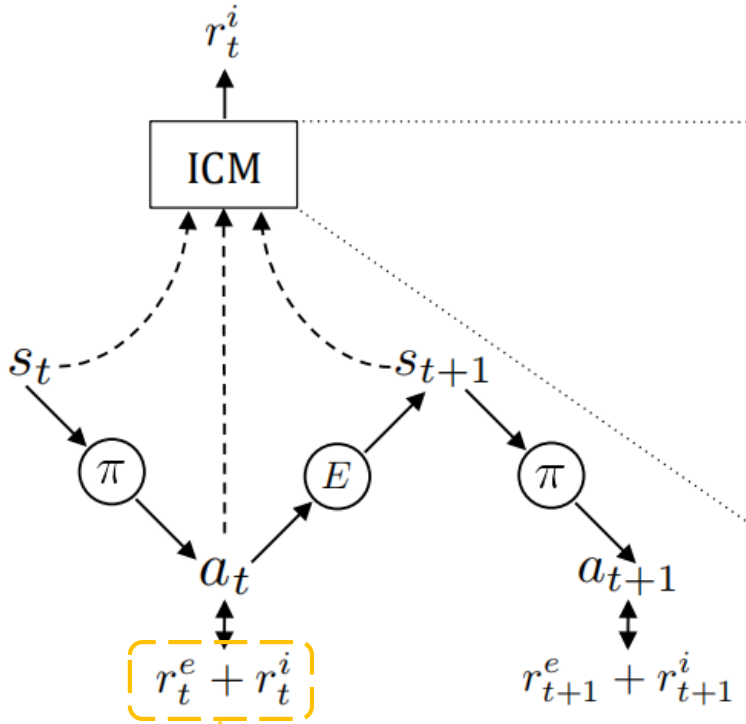


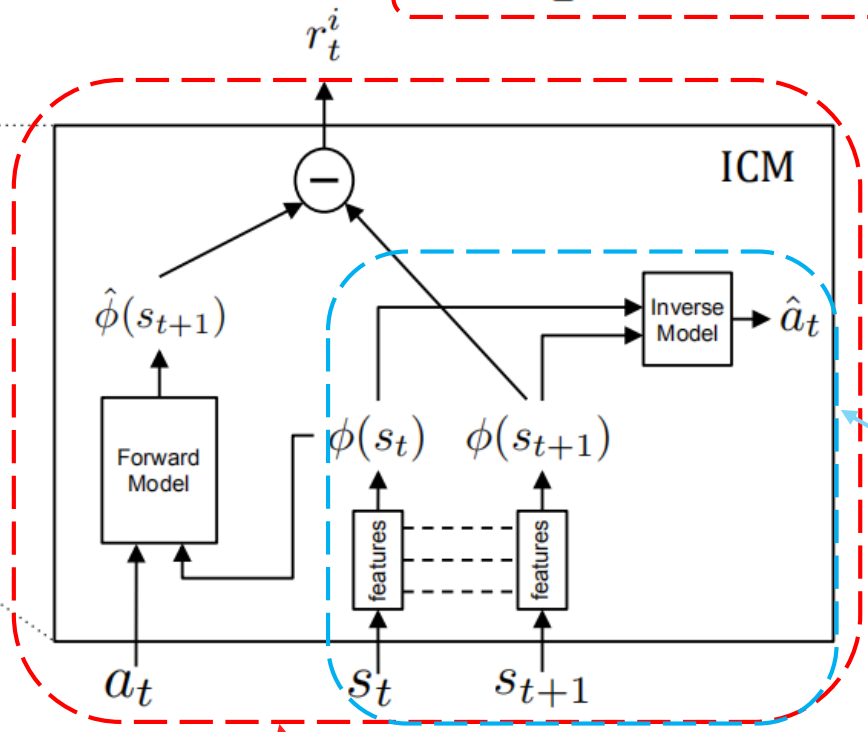
Figure 2. The agent in state s_t interacts with the environment by executing an action a_t sampled from its current policy π and ends up in the state s_{t+1}. The policy π is trained to optimize the sum of the extrinsic reward (r_t^e) provided by the environment E and the curiosity based intrinsic reward signal (r_t^i) generated by our proposed Intrinsic Curiosity Module (ICM). ICM encodes the states s_t, s_{t+1} into the features $\phi(s_t), \phi(s_{t+1})$ that are trained to predict a_t (i.e. inverse dynamics model). The forward model takes as inputs $\phi(s_t)$ and a_t and predicts the feature representation $\hat{\phi}(s_{t+1})$ of s_{t+1}. The prediction error in the feature space is used as the curiosity based intrinsic reward signal. As there is no incentive for $\phi(s_t)$ to encode any environmental features that can not influence or are not influenced by the agent's actions, the learned exploration strategy of our agent is robust to uncontrollable aspects of the environment.

Intrinsic Curiosity Module (ICM)

$$\min_{\theta_P,\theta_I,\theta_F}\left[-\lambda\mathbb{E}_{\pi(s_t;\theta_P)}[\Sigma_t r_t]+(1-\beta)L_I+\beta L_F\right]$$

$$r_t^i=\frac{\eta}{2}\|\hat{\phi}(s_{t+1})-\phi(s_{t+1})\|_2^2 \quad \eta>0$$





$$\min_{\theta_I} L_I(\hat{a}_t,a_t)$$

$$\max_{\theta_P}\mathbb{E}_{\pi(s_t;\theta_P)}[\Sigma_t r_t]$$

$$L_F\Big(\phi(s_t),\hat{\phi}(s_{t+1})\Big)=\frac{1}{2}\|\hat{\phi}(s_{t+1})-\phi(s_{t+1})\|_2^2$$

Diversity is all you need :Learning skills without a reward function

Benjamin Eysenbach*
Carnegie Mellon University
beysenba@cs.cmu.edu

Abhishek Gupta
UC Berkeley

Julian Ibarz
Google Brain

Sergey Levine
UC Berkeley
Google Brain

ICLR 2019

Introduction

Motivation

Intelligent creatures can explore their environments and learn useful skills without supervision.

Contribution

diverse & distinguishable

"Diversity is All You Need"(DIAYN)
A method for learning useful skills without a reward function.

$\pi_\theta(a_t \mid s_t, z)$

extrinsic reward

A skill is a latent-conditioned policy that alters that state of the environment in a consistent way.

Ideas

1. Different skills should visit different states, and hence be distinguishable.

2. we want to use states, not actions, to distinguish skills, because actions that do not affect the environment are not visible to an outside observer.

3. Encourage exploration and incentivize the skills to be as diverse as possible by learning skills that act as randomly as possible, but remain discriminable.

Fomalization

- Objective

for states and actions, respectively; $Z \sim p(z)$ is a latent variable, on which we condition our policy; we refer to a the policy conditioned on a fixed Z as a "skill"; $I(\cdot;\cdot)$ and $\mathcal{H}[\cdot]$ refer to mutual

latent variable: $Z \sim p(z)$

$\pi_\theta(a_t \mid s_t, z)$

maximize $\mathcal{F}(\theta) \triangleq I(S;Z) + \mathcal{H}[A \mid S] - I(A;Z \mid S)$

- Analysis

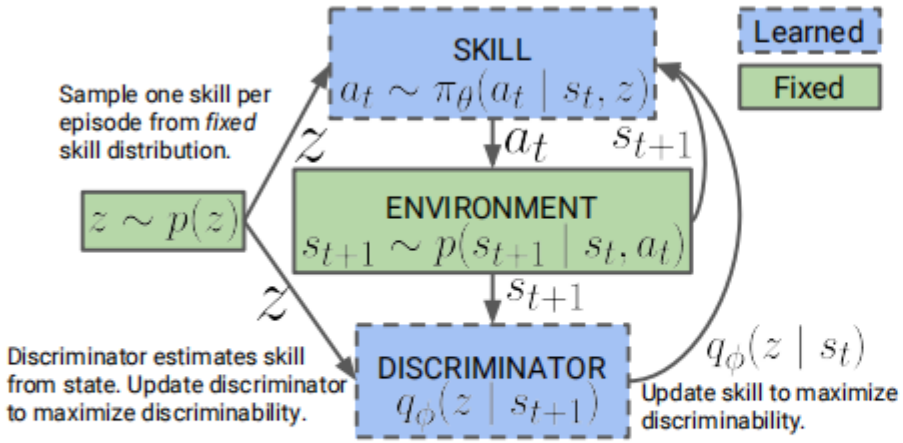
$$\begin{aligned}\mathcal{F}(\theta) &\triangleq I(S;Z) + \mathcal{H}[A \mid S] - I(A;Z \mid S)\\ &= (\mathcal{H}[Z] - \mathcal{H}[Z \mid S]) + \mathcal{H}[A \mid S] - (\mathcal{H}[A \mid S] - \mathcal{H}[A \mid S, Z])\\ &= \mathcal{H}[Z] - \mathcal{H}[Z \mid S] + \mathcal{H}[A \mid S, Z]\\ &= H(A \mid S, Z) + \mathbb{E}_{s\sim\pi(z), z\sim p(z\mid s)}[\log p(z \mid s)] - \mathbb{E}_{z\sim p(z)}[\log p(z)]\end{aligned}$$

Lemma 5.1 *For random variables X, Y and function $f(x, y)$ under suitable regularity conditions:* $\mathbb{E}_{x\sim X, y\sim Y\mid x}[f(x,y)] = \mathbb{E}_{x\sim X, y\sim Y\mid x, x'\sim X\mid y}[f(x', y)]$.

$$\begin{aligned}\mathcal{F}(\theta) &= H(A \mid S, Z) + \mathbb{E}_{s\sim\pi(z), z\sim p(z)}[\log p(z \mid s)] - \mathbb{E}_{z\sim p(z)}[\log p(z)]\\ &= H(A \mid S, Z) + \mathbb{E}_{s\sim\pi(z), z\sim p(z)}\left[\log \frac{p(z \mid s)}{q(z \mid s)} + \log q(z \mid s)\right] - \mathbb{E}_{z\sim p(z)}[\log p(z)]\\ &= H(A \mid S, Z) + D_{KL}[p(z \mid s) \| q(z \mid s)] + \mathbb{E}_{s\sim\pi(z), z\sim p(z)}[\log q(z \mid s)] - \mathbb{E}_{z\sim p(z)}[\log p(z)]\\ &\geq H(A \mid S, Z) + \mathbb{E}_{s\sim\pi(z), z\sim p(z)}[\log q(z \mid s) - \log p(z)] \triangleq \mathcal{G}(\theta, \phi)\end{aligned}$$

DIAYN

$$\mathcal{H}[A \mid S, Z] + \mathbb{E}_{z \sim p(z), s \sim \pi(z)}[\log q_\phi(z \mid s) - \log p(z)]$$

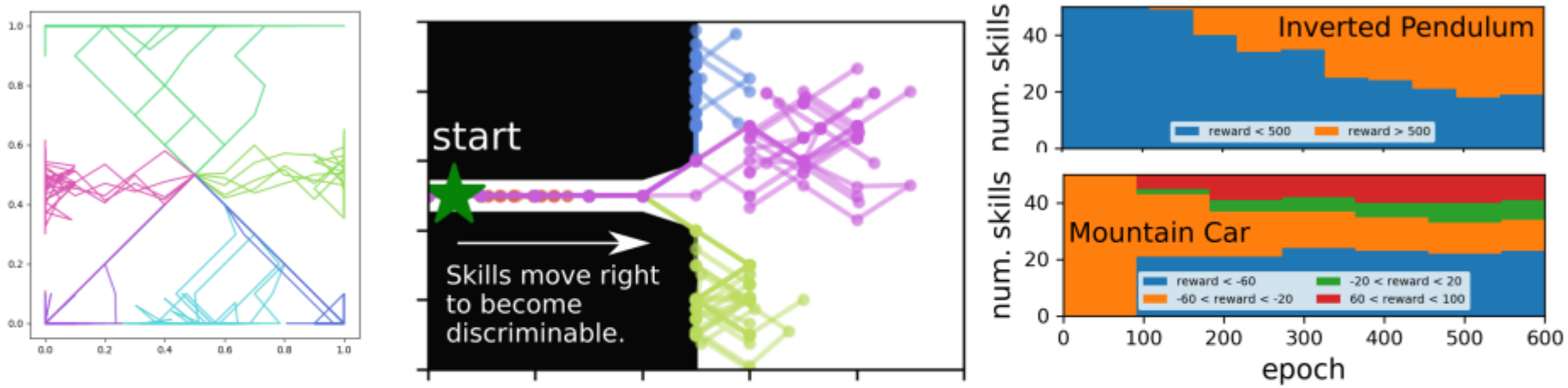


Algorithm 1: DIAYN

while *not converged* **do**
  Sample skill $z \sim p(z)$ and initial state $s_0 \sim p_0(s)$
  for $t \leftarrow 1$ **to** *steps_per_episode* **do**
    Sample action $a_t \sim \pi_\theta(a_t \mid s_t, z)$ from skill.
    Step environment: $s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$.
    Compute $q_\phi(z \mid s_{t+1})$ with discriminator.
    Set skill reward $r_t = \log q_\phi(z \mid s_{t+1}) - \log p(z)$
    Update policy (θ) to maximize r_t with SAC.
    Update discriminator (ϕ) with SGD.

Figure 1: **DIAYN Algorithm**: We update the discriminator to better predict the skill, and update the skill to visit diverse states that make it more discriminable.

Experiments-Analysis of learned skills



(a) 2D Navigation (b) Overlapping Skills (c) Training Dynamics

Figure 2: *(Left)* DIAYN skills in a simple navigation environment; *(Center)* skills can overlap if they eventually become distinguishable; *(Right)* diversity of the rewards increases throughout training.

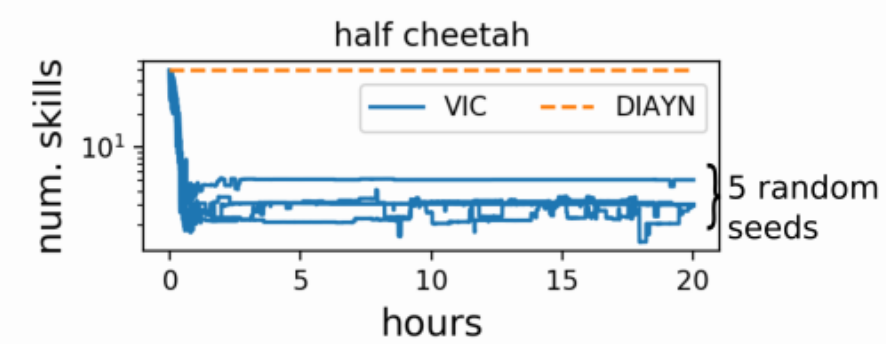
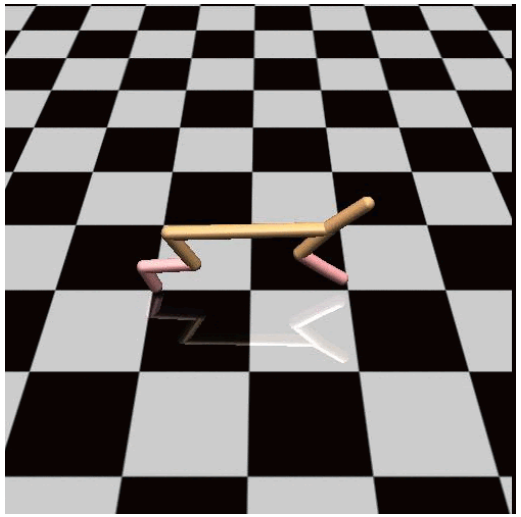
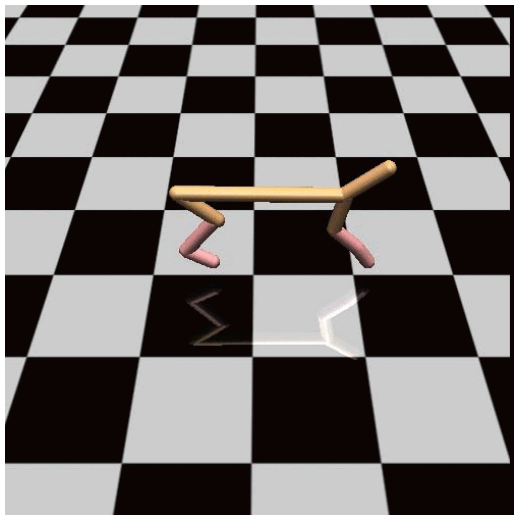
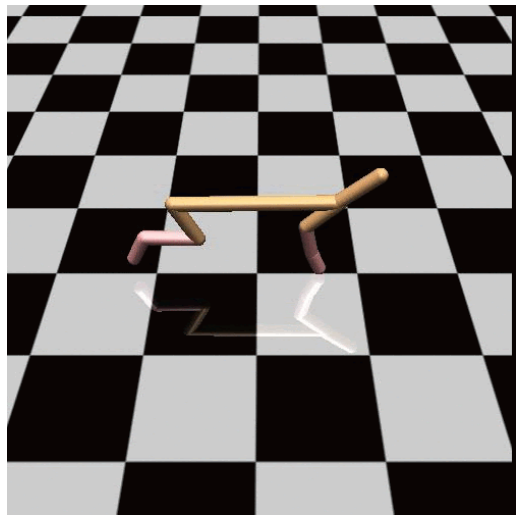
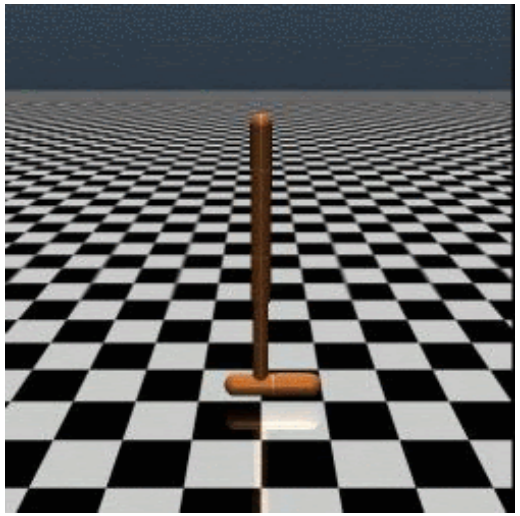
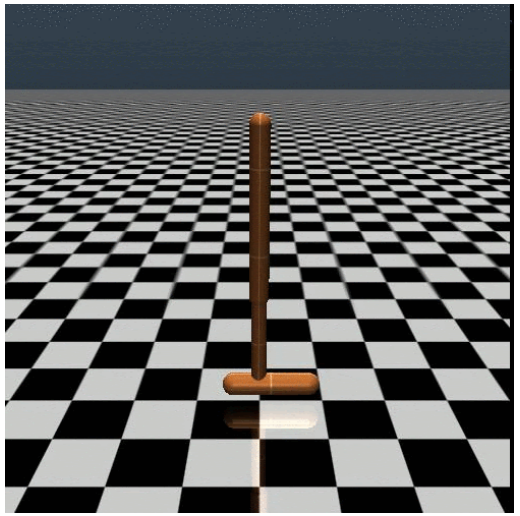
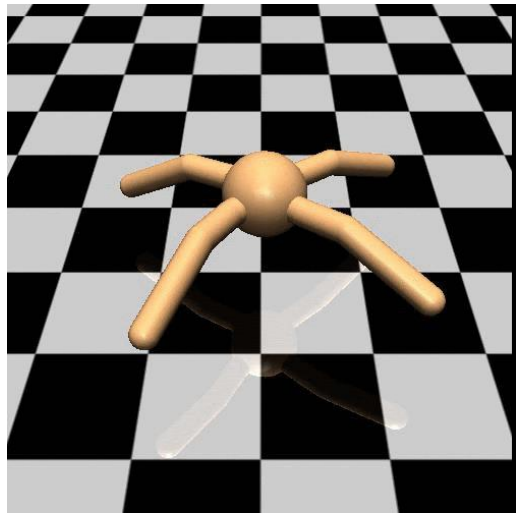


Figure 4: **Why use a fixed prior?** In contrast to prior work, DIAYN continues to sample all skills throughout training.

Experiments-Analysis of learned skills

Experiments- Downstream tasks

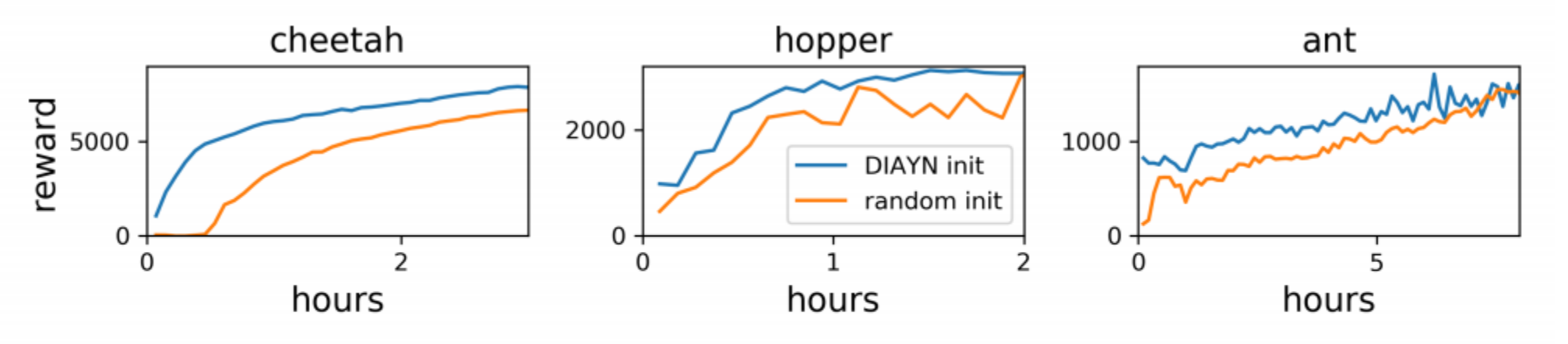


Figure 5: **Policy Initialization**: Using a DIAYN skill to initialize weights in a policy accelerates learning, suggesting that pretraining with DIAYN may be especially useful in resource constrained settings. Results are averages across 5 random seeds.

Experiments-application

expert trajectory: $\tau^* = \{(s_i)\}_{1 \le i \le N}$

skill: $\hat{z} = \arg\max_z \Pi_{s_t \in \tau^*} q_\phi(z \mid s_t)$ → $\pi_\theta(a_t \mid s_t, \hat{z})$

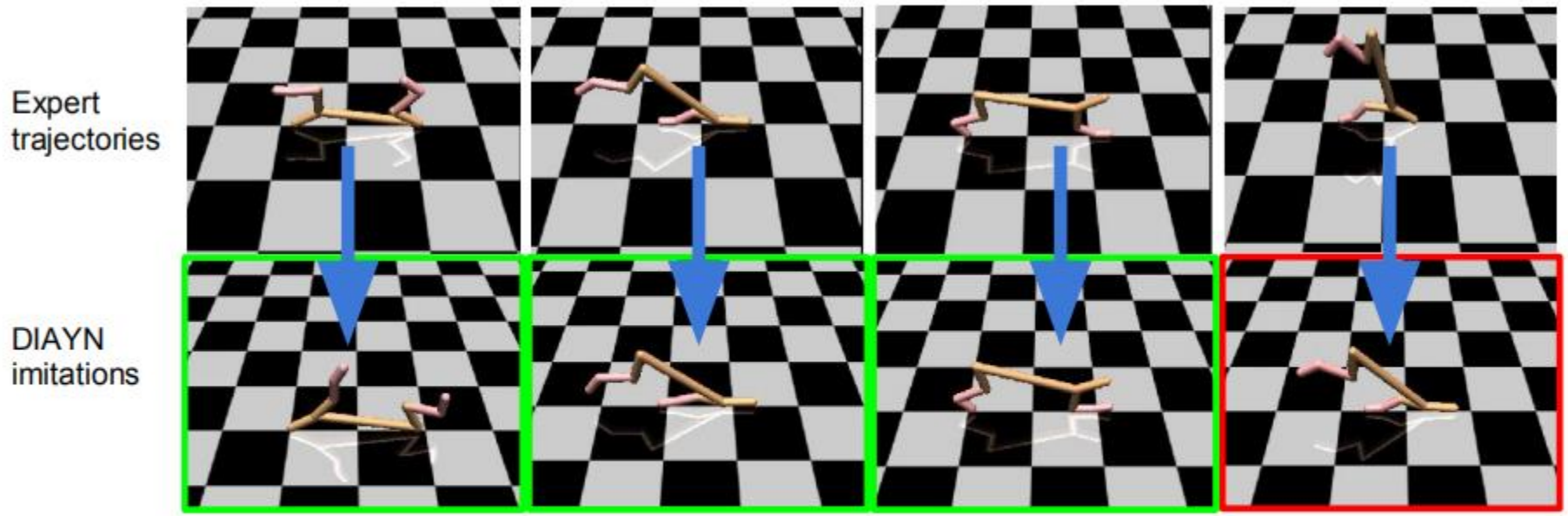


Figure 9: **Imitating an expert**: DIAYN imitates an expert standing upright, flipping, and faceplanting, but fails to imitate a handstand.

Thanks