Learn Goal-Conditioned Policy with Intrinsic Motivation for Deep Reinforcement Learning

Presenter: Jinxin Liu (liujinxin@westlake.edu.cn)

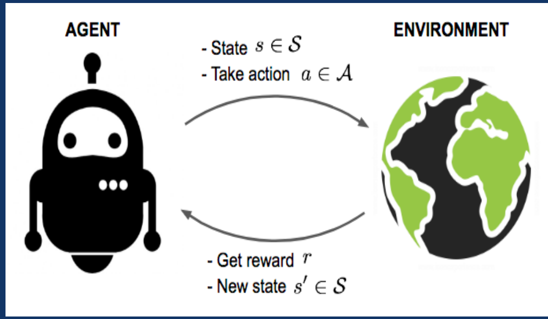
Co-Authors: Donglin Wang; Qiangxing Tian; Zhengyu Chen

CONTENTS

CHAPTER 1

Deep Reinforcement Learning (RL)

A brief explanation of deep reinforcement learning.

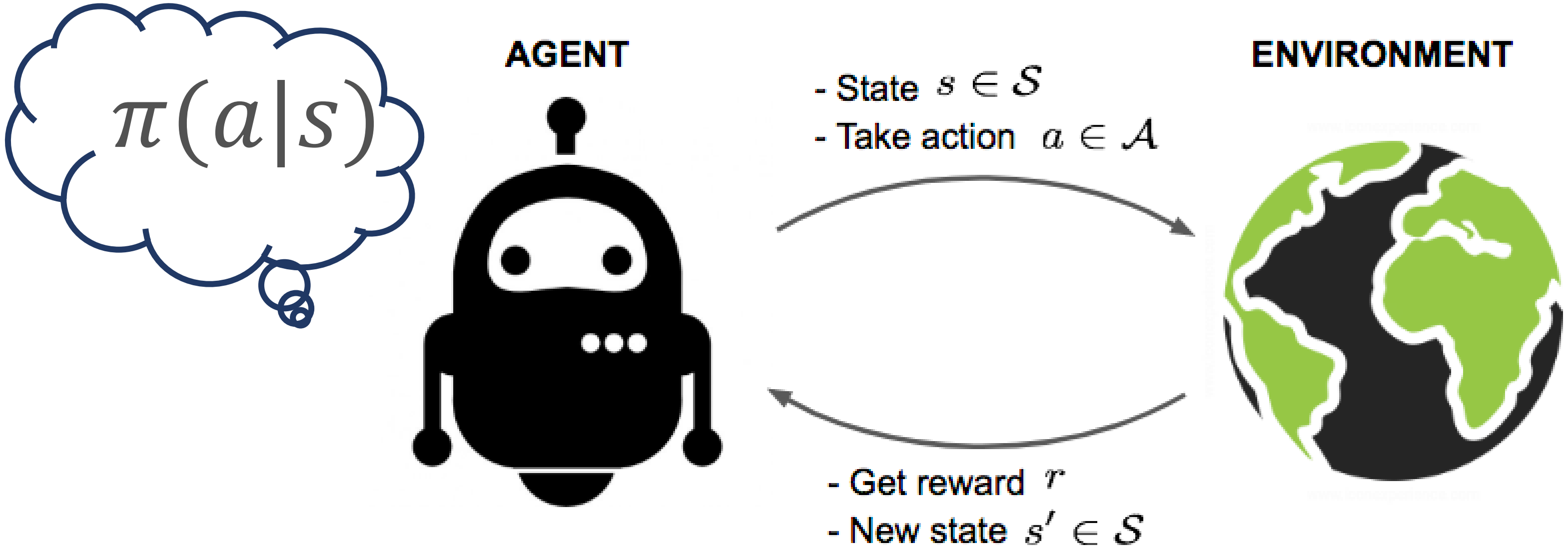


西湖大學
WESTLAKE UNIVERSITY

Reinforcement Learning (RL)

Mathematical formalism:
Reinforcement Learning

Representation:
Deep Networks



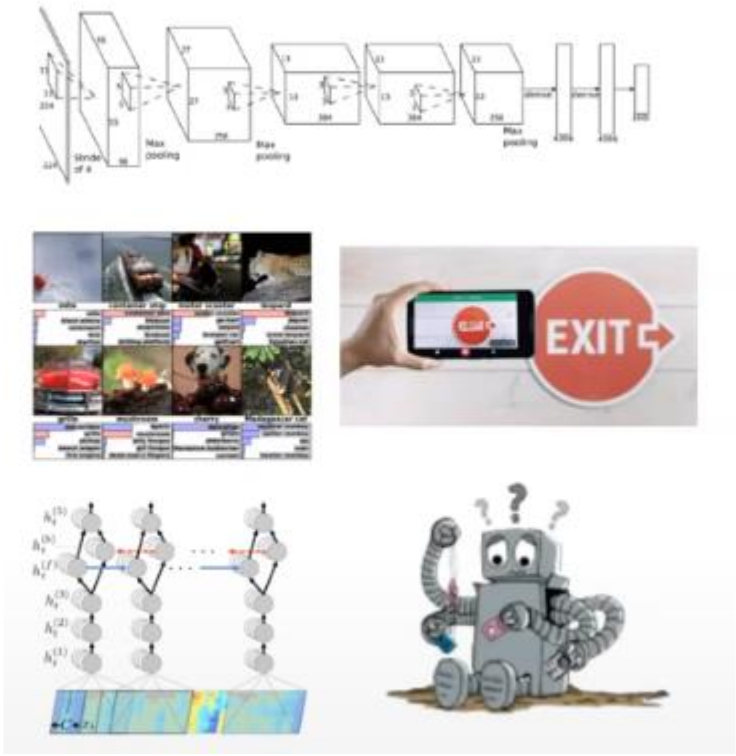


Image credit: RAIL

$$\max_{\pi} \mathrm{E}_{\pi}(\Sigma_t[r_t|\pi])$$

where reward $r_t := r(s_t, a_t)$

Image Credit:
https://lilianweng.github.io/lil-log/2018/02/19/a-long-peek-into-reinforcement-learning.html

Reinforcement Learning (RL)

It (more or less) works

Stack a Lego block

Image Credit: Google

Scalable Deep Reinforcement Learning for Robotic Manipulation

Image credit: *Push a mug onto a coaster*
https://bair.berkeley.edu/blog/2019/05/20/solar/

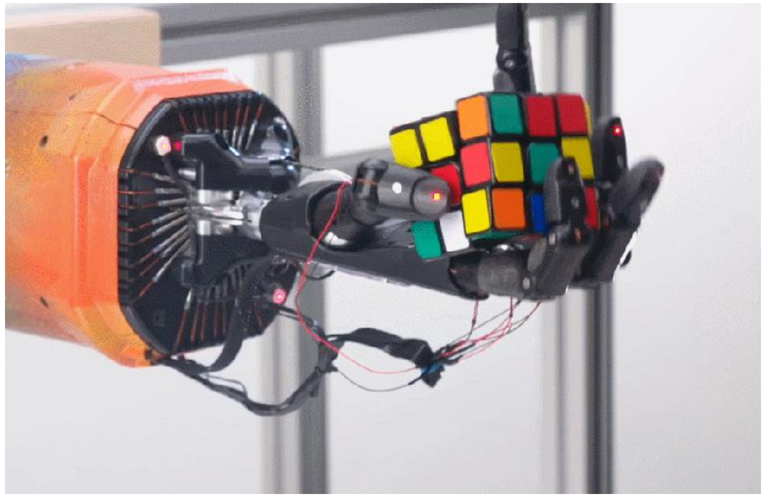
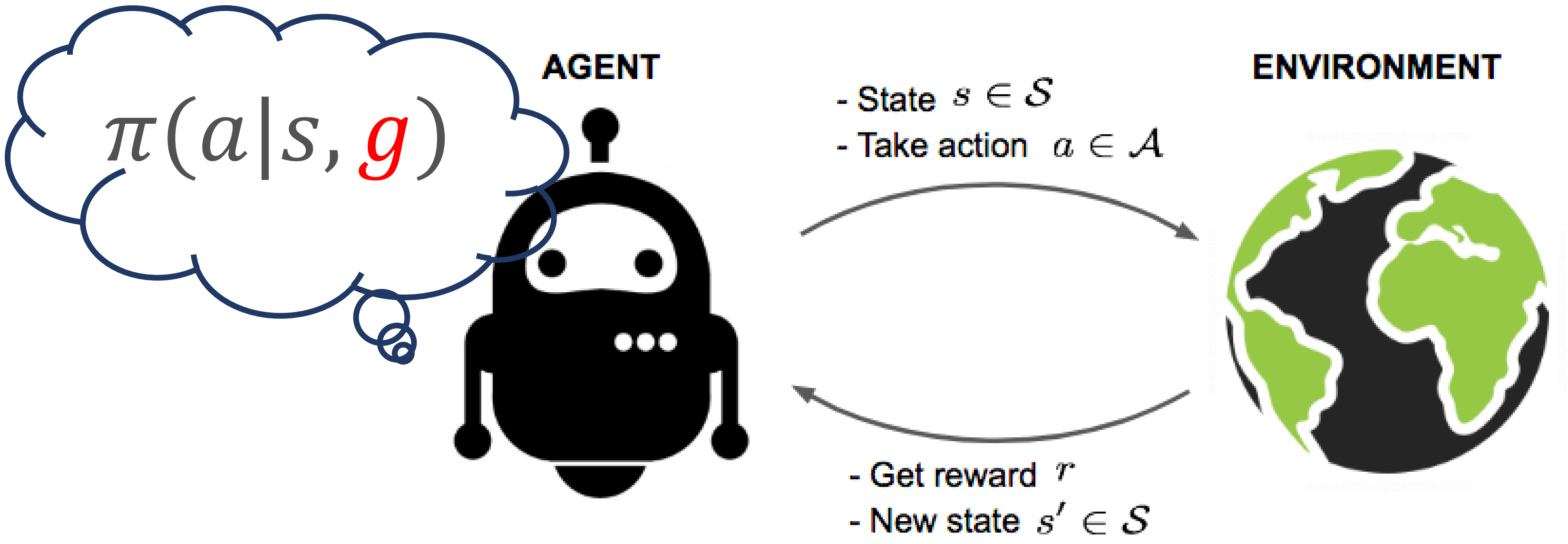


Image Credit: OpenAI

Solving Rubik's Cubewith a Robot Hand

Multi-Goal Reinforcement Learning

Mathematical formalism:
Multi-Goal Reinforcement Learning



$$\max_{\pi} \mathrm{E}_g \mathrm{E}_\pi (\Sigma_t [r_t | \pi])$$

where reward $r_t := r(s_t, a_t, g)$

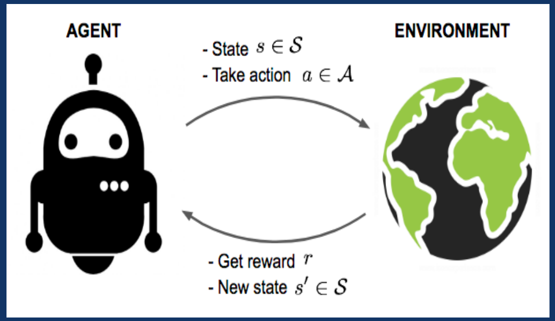
Image Credit:
https://lilianweng.github.io/lil-log/2018/02/19/a-long-peek-into-reinforcement-learning.html

CONTENTS

CHAPTER 1

Deep Reinforcement Learning (RL)

A brief explanation of deep reinforcement learning.



CHAPTER 2

Unsupervised RL

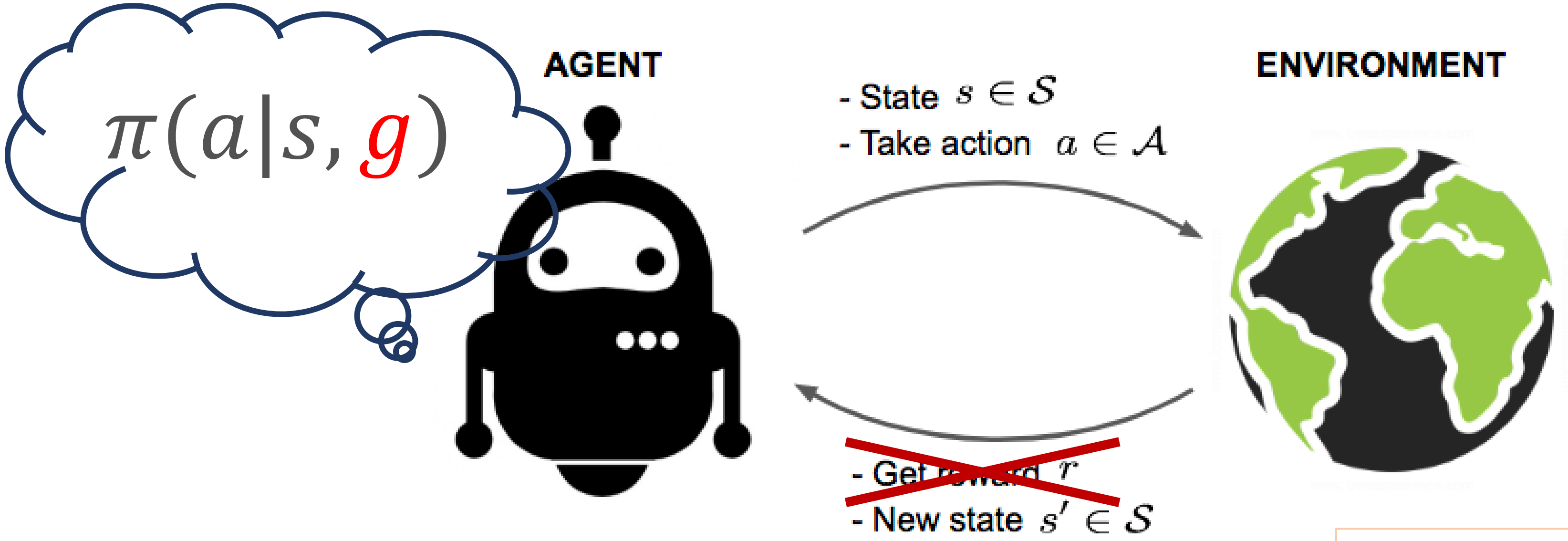
A brief explanation of unsupervised reinforcement learning.

西湖大學
WESTLAKE UNIVERSITY

Unsupervised/Self-supervised RL

Mathematical formalism:

Unsupervised Reinforcement Learning



$$\max_{\pi} \mathrm{E}_g \mathrm{E}_\pi (\Sigma_t [r_t | \pi])$$

where reward $r_t := r(s_t, a_t, s_{t+1}, g)$ ✗

1. Learn reward function

$$r_t := r(s_t, a_t, s_{t+1}, g)$$

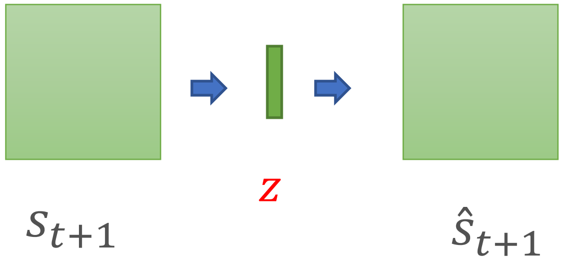
2. Learn goal distribution

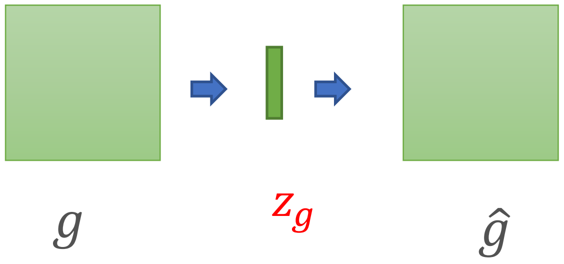
$$\mathrm{E}_g$$

Unsupervised/Self-supervised RL

1. Learn reward function

$$r_t := r(s_t, a_t, s_{t+1}, g)$$
$$= r(s_{t+1}, g)$$





$$r(s_{t+1}, g) \coloneqq -\cos(z, z_g)$$

2. Learn goal distribution

Assume the spaces of perceptual goals and states are same.

Sample goals from the historical trajectories of the policy to be trained.

perceptually-specific goal based approach

Unsupervised/Self-supervised RL

Latent-variable based approach

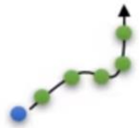
Sharma, Archit, et al.
"Dynamics-aware unsupervised discovery of skills."

Eysenbach, Benjamin, et al.
"Diversity is all you need: Learning skills without a reward function."

$$\max_{\pi,p(\mathcal{T})} \mathcal{H}[\mathbf{s}_H]-\mathcal{H}[\mathbf{s}_H|\mathcal{T}] = \mathcal{I}[\mathbf{s}_H;\mathcal{T}]$$

$$\max_{\pi,p(\mathcal{T})} \mathcal{H}[\tau]-\mathcal{H}[\tau|\mathcal{T}] = \mathcal{I}[\tau;\mathcal{T}]$$

$$\max_{\pi,p(\mathcal{T})} \mathcal{H}[\mathbf{s}]-\mathcal{H}[\mathbf{s}|\mathcal{T}] = \mathcal{I}[\mathbf{s};\mathcal{T}]$$



Unsupervised/Self-supervised RL

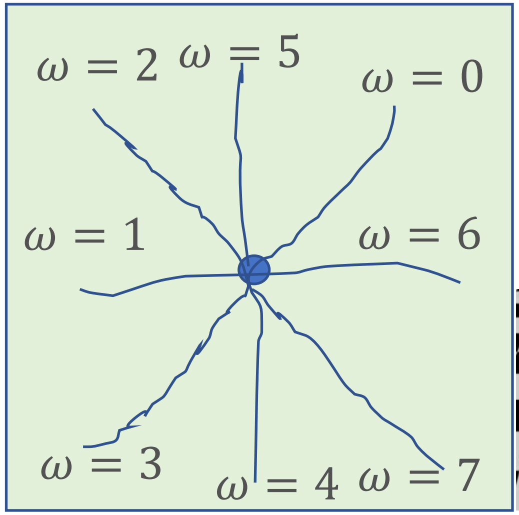
Eysenbach, Benjamin, et al.
"Diversity is all you need: Learning skills without a reward function."

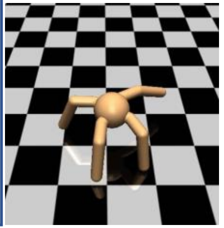
DIAYN: $\max \; I(\omega; s)$

$$\pi(a|s,\omega), \omega \sim p(\omega)$$

$$\max_{\pi,p(s)} I(\omega; s) = H[\omega] - H[\omega|s]$$

$$I(\omega;\tau) \geq \underbrace{H[\omega]}_{\textit{prior (fixed)}} + \underbrace{\mathbb{E}_{\omega,s}[\, q_\phi(\omega|s)]}_{\textit{reward function: } r}$$





● *Initial State* *Agent*

Latent-variable based approach

Unsupervised/Self-supervised RL

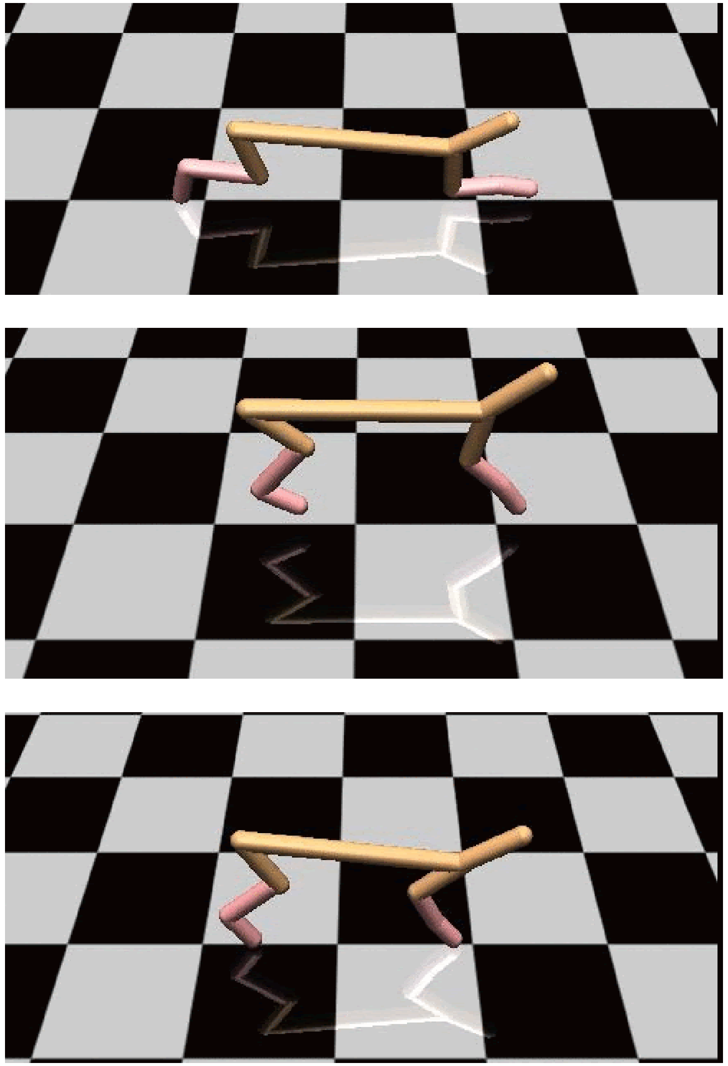
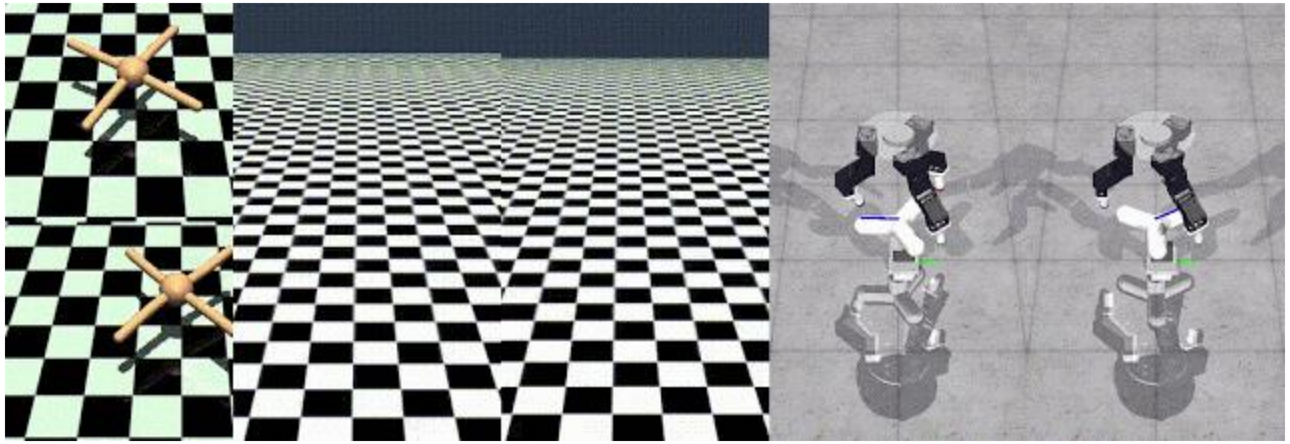
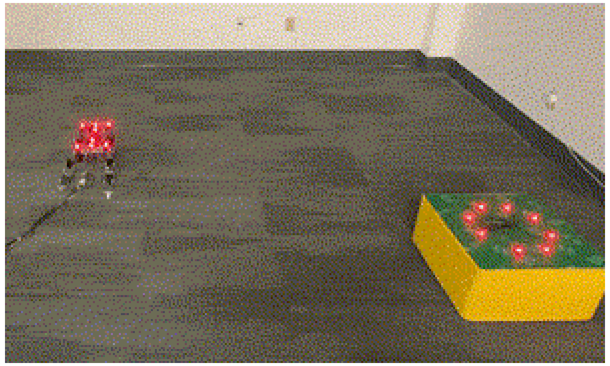


Image Credit: Eysenbach et al., 2018





Learned diverse skills.

Image Credit: GOOGLE

Unsupervised/Self-supervised RL

perceptually-specific goal based approach

$$r(s_{t+1}, g) \coloneqq -\cos(z, z_g)$$

Prior non-parametric measure function may limit the repertoires of behaviors and impose manual engineering burdens.

Latent-variable based approach

$$\max_{\pi(a|s,\omega),\ \omega \sim p(\omega)} I(\omega; s)$$

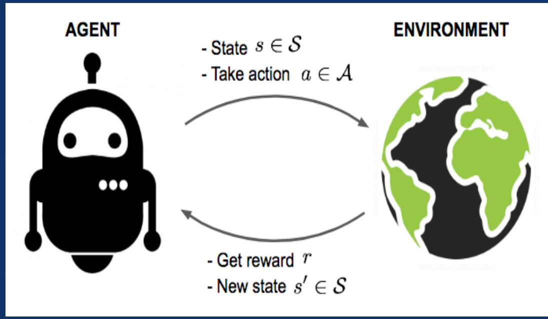
Such policy is conditioned on latent variables rather than perceptually-specific goals.

CONTENTS

CHAPTER 1

Deep Reinforcement Learning (RL)

A brief explanation of deep reinforcement learning.



CHAPTER 2

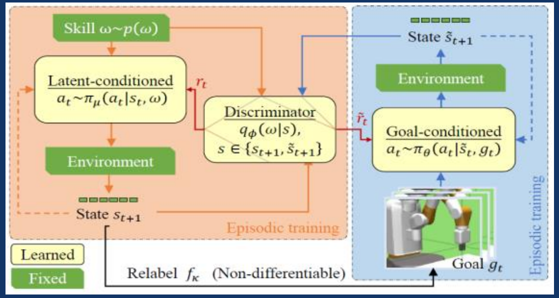
Unsupervised RL

A brief explanation of unsupervised reinforcement learning.

CHAPTER 3

Method: GPIM

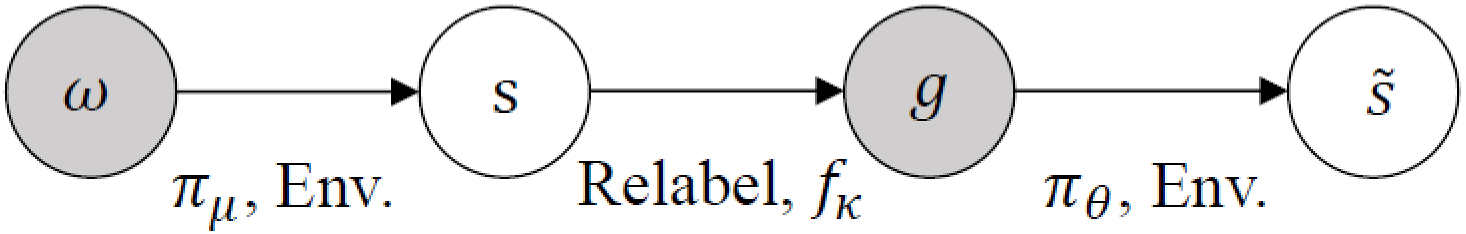
Learn Goal-Conditioned Policy with Intrinsic Motivation for Deep Reinforcement Learning



西湖大學
WESTLAKE UNIVERSITY

Proposed GPIM method

Learn Goal-conditioned Policy with Intrinsic Motivation



Generate goal distribution
Learn reward function

Relabel states s as goals g for training π_θ

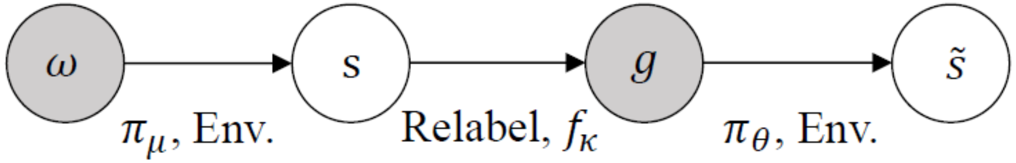
Learn policy

$$\mathcal{F}(\mu, \theta) = \mathcal{I}(s; \omega) + \mathcal{I}(\tilde{s}; g)$$

$$\begin{aligned} \mathcal{F}(\mu, \theta) &\geq \mathcal{I}(s; \omega) + \mathcal{I}(\tilde{s}; \omega) \\ &= 2\mathcal{H}(\omega) + \mathbb{E}_{p_m(\cdot)}\left[\log p(\omega|s) + \log p(\omega|\tilde{s})\right] \end{aligned}$$

Proposed GPIM method

Learn Goal-conditioned Policy with Intrinsic Motivation



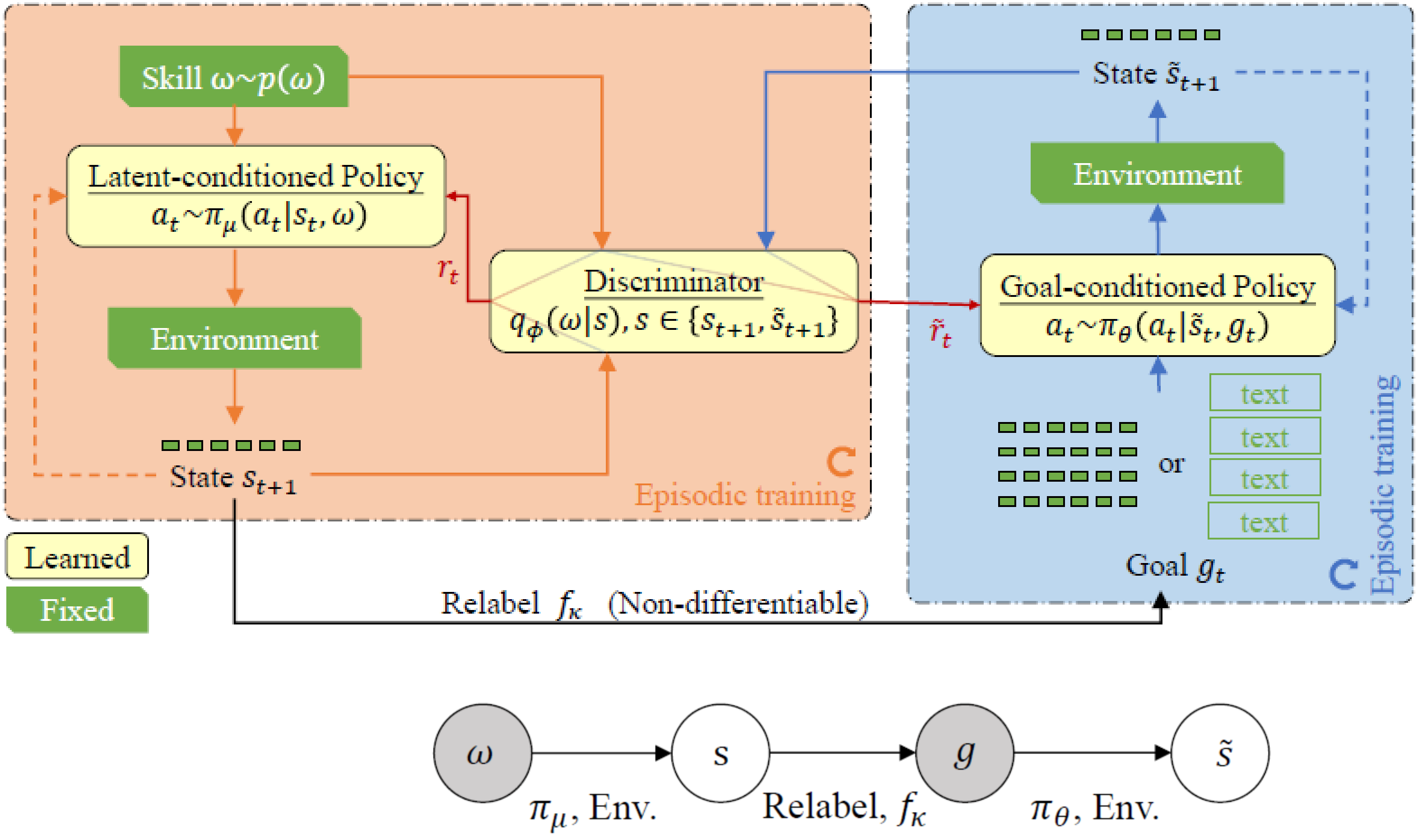
$$\begin{aligned}\mathcal{F}(\mu,\theta) &\geq \mathcal{I}(s;\omega) + \mathcal{I}(\tilde{s};\omega)\\ &= 2\mathcal{H}(\omega) + \mathbb{E}_{p_m(\cdot)}\left[\log p(\omega|s) + \log p(\omega|\tilde{s})\right]\\ &\geq 2\mathcal{H}(\omega) + \mathbb{E}_{p_m(\cdot)}\left[\underline{\log q_\phi(\omega|s)} + \underline{\log q_\phi(\omega|\tilde{s})}\right]\end{aligned}$$

Reward function (for both policies):

$$r_t = \log q_\phi(\omega|s_{t+1}) - \log p(\omega)$$

Proposed GPIM method

Learn Goal-conditioned Policy with Intrinsic Motivation

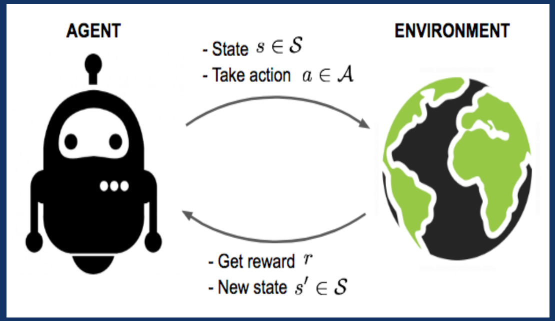


CONTENTS

CHAPTER 1

Deep Reinforcement Learning (RL)

A brief explanation of deep reinforcement learning.



CHAPTER 2

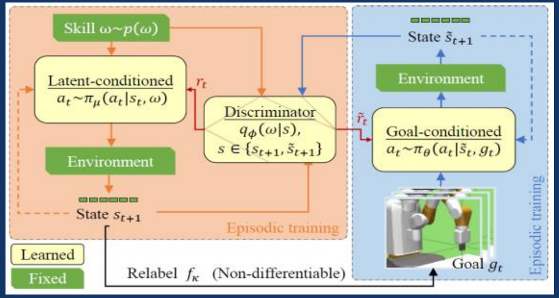
Unsupervised RL

A brief explanation of unsupervised reinforcement learning.

CHAPTER 3

Method: GPIM

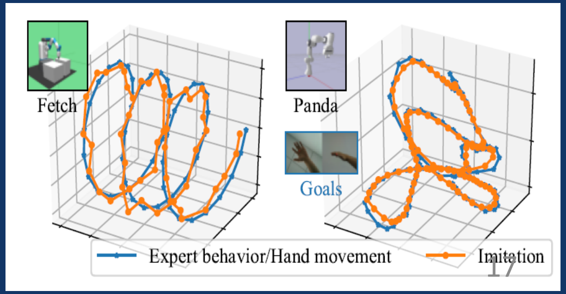
Learn Goal-Conditioned Policy with Intrinsic Motivation for Deep Reinforcement Learning



CHAPTER 4

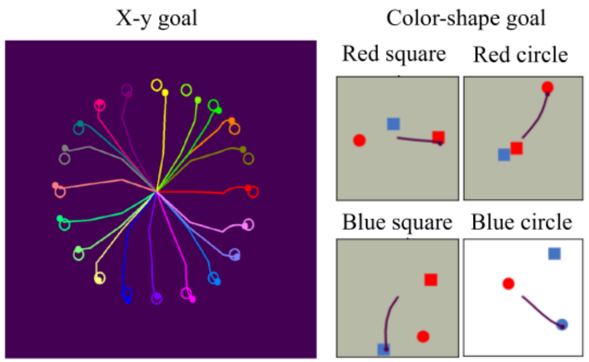
Experiments

Briefly introduce the experiments in GPIM.

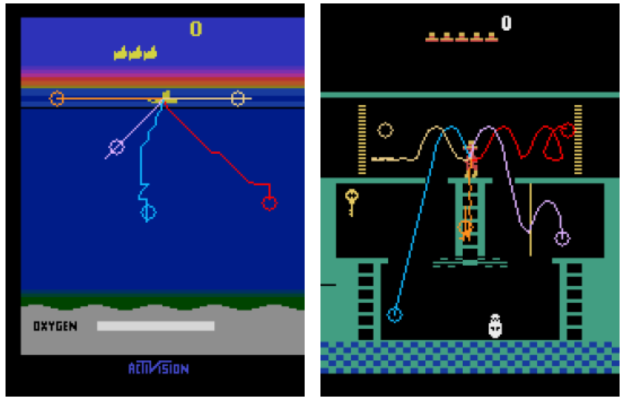


西湖大學
WESTLAKE UNIVERSITY

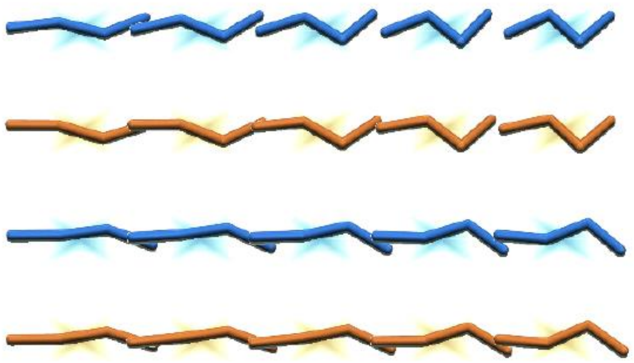
Experiments: Visualization



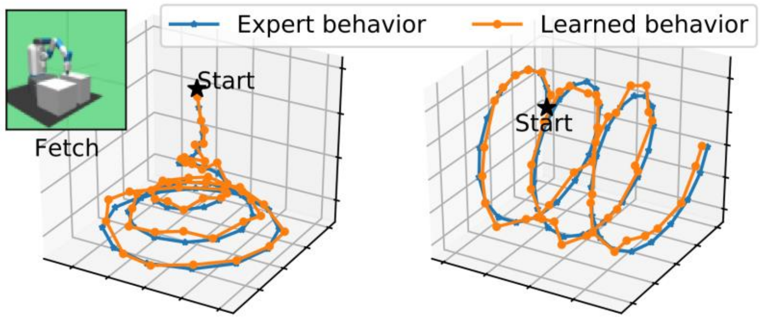
Fixed goals



High dimensional goals

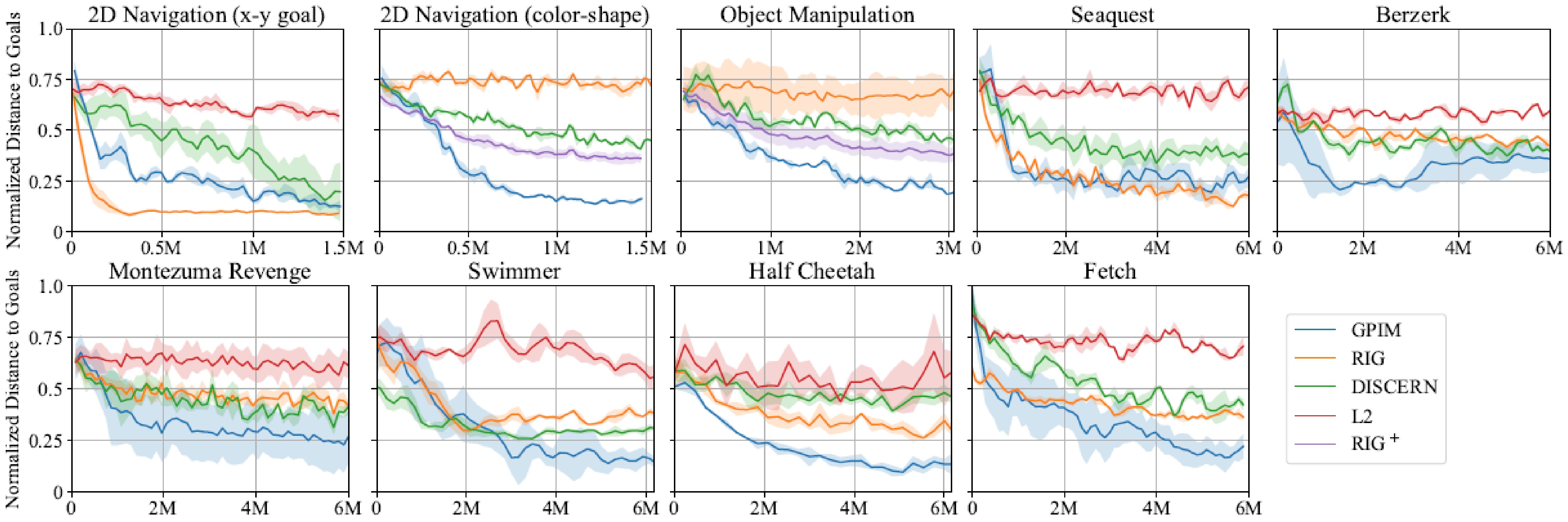


Dynamic goals



Temporally-extended goals

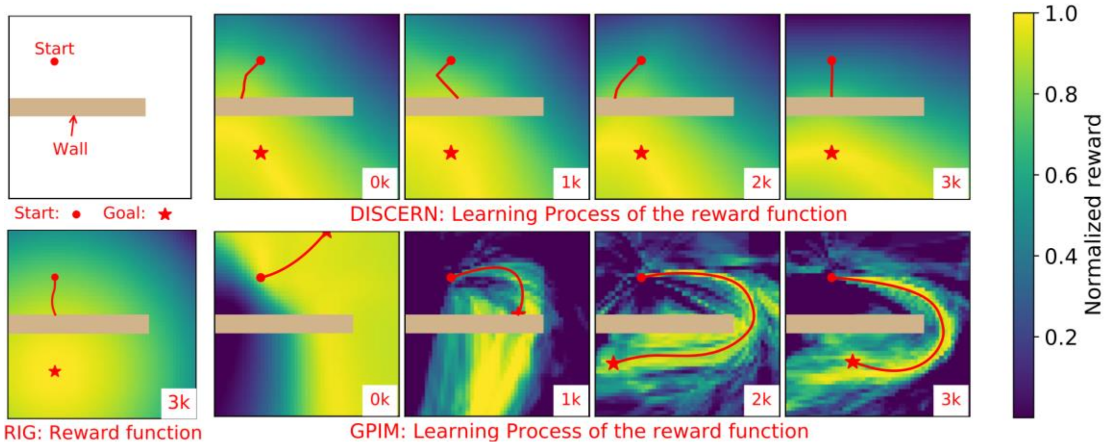
Experiments: Compared with Baselines



Performance (normalized distance to goals vs. actor steps) of our GPIM and baselines.

Experiments: Dynamical Distance Learning

The performance of unsupervised RL methods depends on the diversity of autonomously generated goals and the expressiveness of the learned reward function, which is conditioned on the generated goals.



Our method (GPIM) builds up the reward function after exploring the environment, *the dynamic of which itself further shapes the reward function*. We can see that our model provides the reward function better expressiveness of the task by compensating for the dynamic.

Conclusion

1. We introduce a latent-conditioned policy with a procedural relabeling function to generate tasks for training the goal-conditioned policy.

2. We theoretically describe the performance guarantee of our (unsupervised) objective compared with the standard multi-goal RL.

3. We also conduct extensive experiments on a variety of robotic tasks to demonstrate the effectiveness and efficiency of our method, which outperforms prior unsupervised methods.

Thank you