Planning for Empowerment in Visual Environments

Minghan Li

Overview

Empowerment

Contributions

Background

Methods

Experiments

Future work

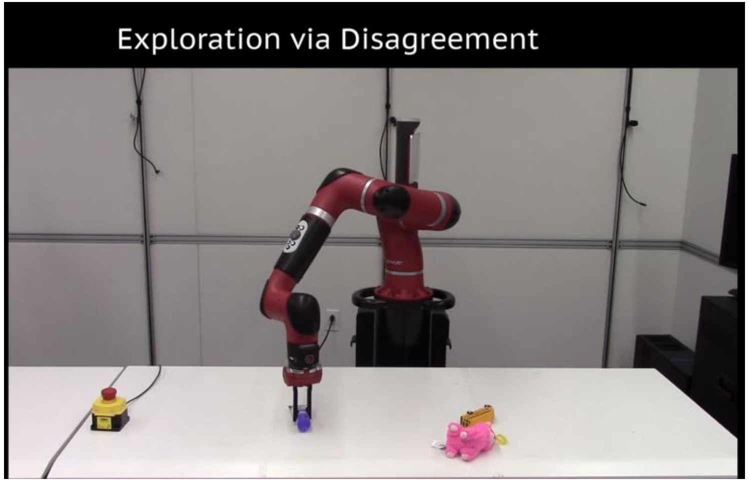
Intrinsic Motivation

Providing external rewards is often difficult or expensive in reinforcement learning, especially in applications in the physical world.

Intrinsic motivation, on the other hand, encourages the agent to learn general knowledge or skills in its environment to later solve more difficult tasks.

(a) learn to explore in Level-1 (b) explore faster in Level-2

Figure 1. Discovering how to play *Super Mario Bros* **without rewards**. (a) Using only curiosity-driven exploration, the agent



Many Intrinsic objectives

Information gain e.g. Lindley 1956, Sun 2011, Houthooft 2017

Prediction error e.g. Schmidhuber 1991, Bellemare 2016, Pathak 2017

Empowerment e.g. Klyubin 2005, Tishby 2011, Gregor 2016

Skill discovery e.g. Eysenbach 2018, Sharma 2020, Co-Reyes 2018

Surprise minimization e.g. Schrödinger 1944, Friston 2013, Berseth 2020

Bayes-adaptive RL e.g. Gittins 1979, Duff 2002, Ross 2007

Empowerment Objective

Defined as mutual information between agent's future actions and inputs.

Measures degree of control over future inputs. In each state, choose precise action. Across all states, use all actions.

Empowerment Objective

Defined as mutual information between agent's future actions and inputs.

Measures degree of control over future inputs. In each state, choose precise action. Across all states, use all actions.

Prior methods (Jung et al., Gregor et al., Karl et al.) use tractable models or variational methods to optimize empowerment in discrete toy environments.

Empowerment Objective

Defined as mutual information between agent's future actions and inputs.

Measures degree of control over future inputs: In each state, choose precise action. Across all states, make use of all actions.

Prior methods (Jung et al., Gregor et al., Karl et al.) use tractable models or variational methods to optimize empowerment in discrete toy environments.

We aim to scale empowerment to complex visual environments.

- Requires scalable MC estimators that let us estimate empowerment using flexible deep neural networks.
- Trivial to achieve diverse pixel inputs (e.g. spin around). Need a meaningful representation of the high-dimensional input that the agent can control.

Summary of Contributions

1. We leverage a world model learned from pixels to infer a latent state about the environment that we apply empowerment to.
2. The world model lets us optimize for empowerment in imagination, reducing the amount of trial and error in the real environment.
3. Mutual informations for deep models are often intractable. We propose two tractable MC estimators for empowerment (action space, latent state space)
4. Learning the world model without task rewards, we demonstrate successful zero-shot and few-show adaptation to a range of challenging control tasks.

Overview

Empowerment

Contributions

Background

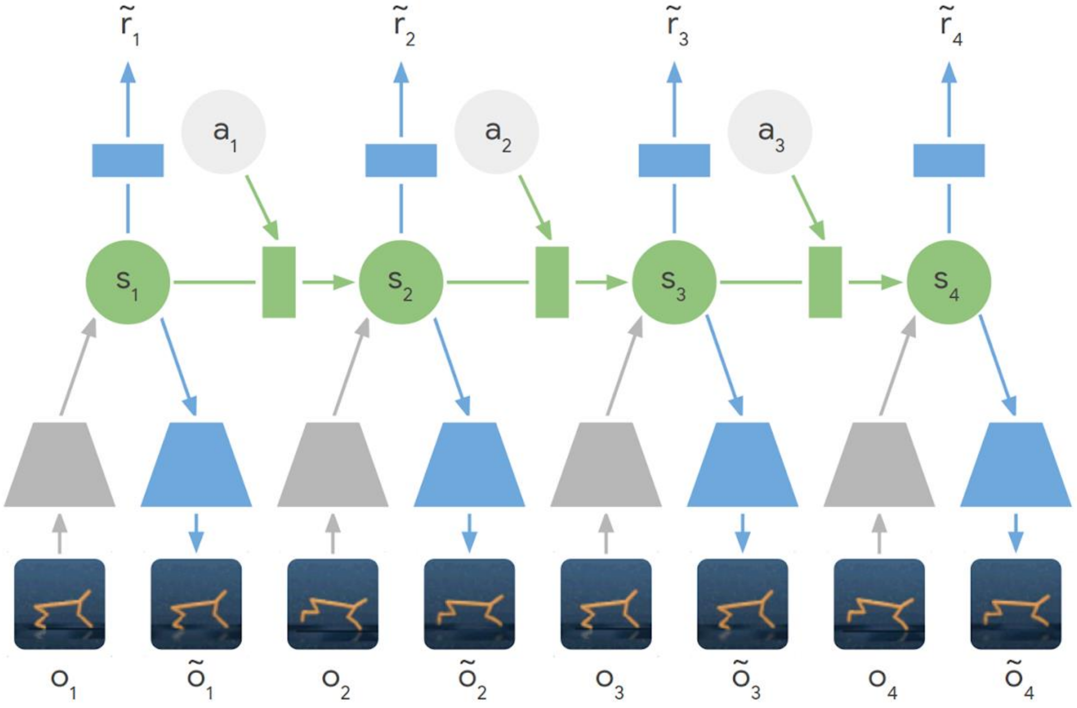
Methods

Experiments

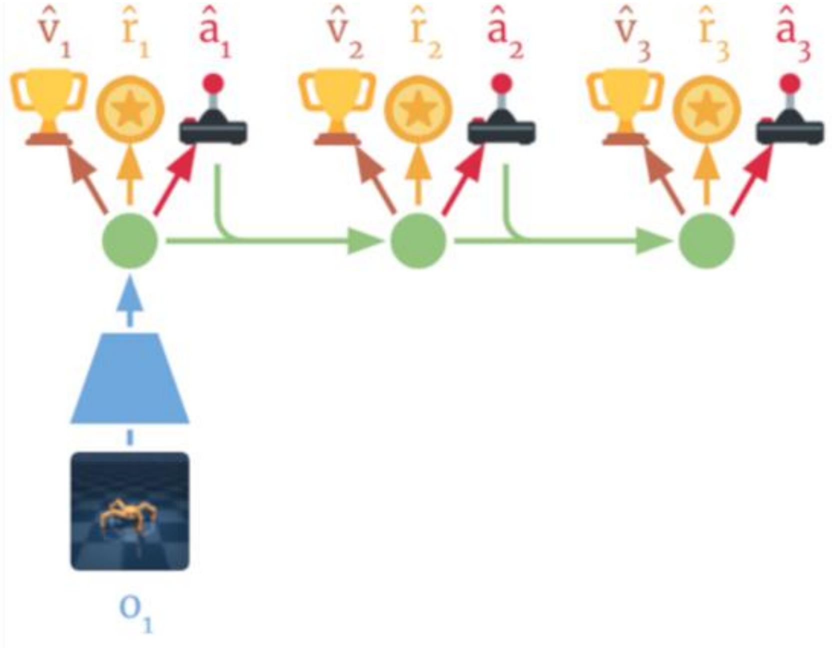
Future work

Background: Learning Latent Dynamics (PlaNet)

$$
\begin{aligned}
&\text{Representation model:} && p(s_t \mid s_{t-1}, a_{t-1}, o_t) \\
&\text{Transition model:} && q(s_t \mid s_{t-1}, a_{t-1}) \\
&\text{Reward model:} && q(r_t \mid s_t).
\end{aligned}
$$



Background: Learning Behaviors (Dreamer)



Action model: $a_\tau \sim q_\phi(a_\tau \mid s_\tau)$

Value model: $v_\psi(s_\tau) \approx \mathrm{E}_{q(\cdot|s_\tau)}\big(\sum_{\tau=t}^{t+H} \gamma^{\tau-t} r_\tau\big).$

$$a_\tau = \tanh\big(\mu_\phi(s_\tau) + \sigma_\phi(s_\tau)\,\epsilon\big), \quad \epsilon \sim \mathrm{Normal}(0, \mathbb{I}).$$

$$\mathrm{V_R}(s_\tau) \doteq \mathrm{E}_{q_\theta, q_\phi}\Big(\sum_{n=\tau}^{t+H} r_n\Big),$$

$$\mathrm{V_N^k}(s_\tau) \doteq \mathrm{E}_{q_\theta, q_\phi}\Big(\sum_{n=\tau}^{h-1} \gamma^{n-\tau} r_n + \gamma^{h-\tau} v_\psi(s_h)\Big)$$

$$\mathrm{V}_\lambda(s_\tau) \doteq (1-\lambda) \sum_{n=1}^{H-1} \lambda^{n-1} \mathrm{V_N^n}(s_\tau) + \lambda^{H-1} \mathrm{V_N^H}(s_\tau),$$

Method: Empowerment Overview

Our general definition of empowerment under policy \pi is the mutual information between sequences of actions and model states:

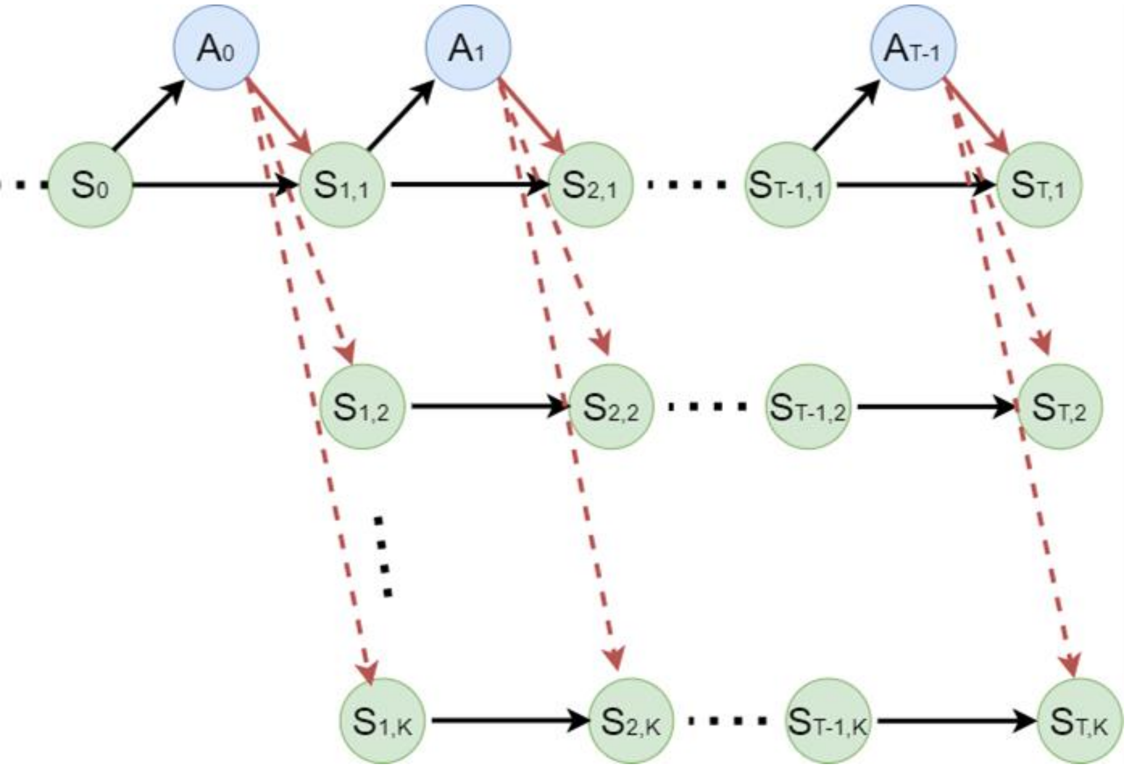
$$\begin{aligned}\mathcal{E}(\pi) = \mathrm{I}(S_{1:T}; A_{1:T} \mid s_0) &= \mathrm{H}(A_{1:T} \mid s_0) - \mathrm{H}(A_{1:T} \mid S_{1:T}, s_0) \\ &= \mathrm{H}(S_{1:T} \mid s_0) - \mathrm{H}(S_{1:T} \mid A_{1:T}, s_0)\end{aligned}$$

Can estimate this objective either in state-space or action-space.

Compute Monte-Carlo estimates of the entropies from multiple imagined rollouts.

Conditional entropy is easy to compute given a set of rollouts. The marginal entropy is estimated as entropy of a mixture distribution across the rollouts.

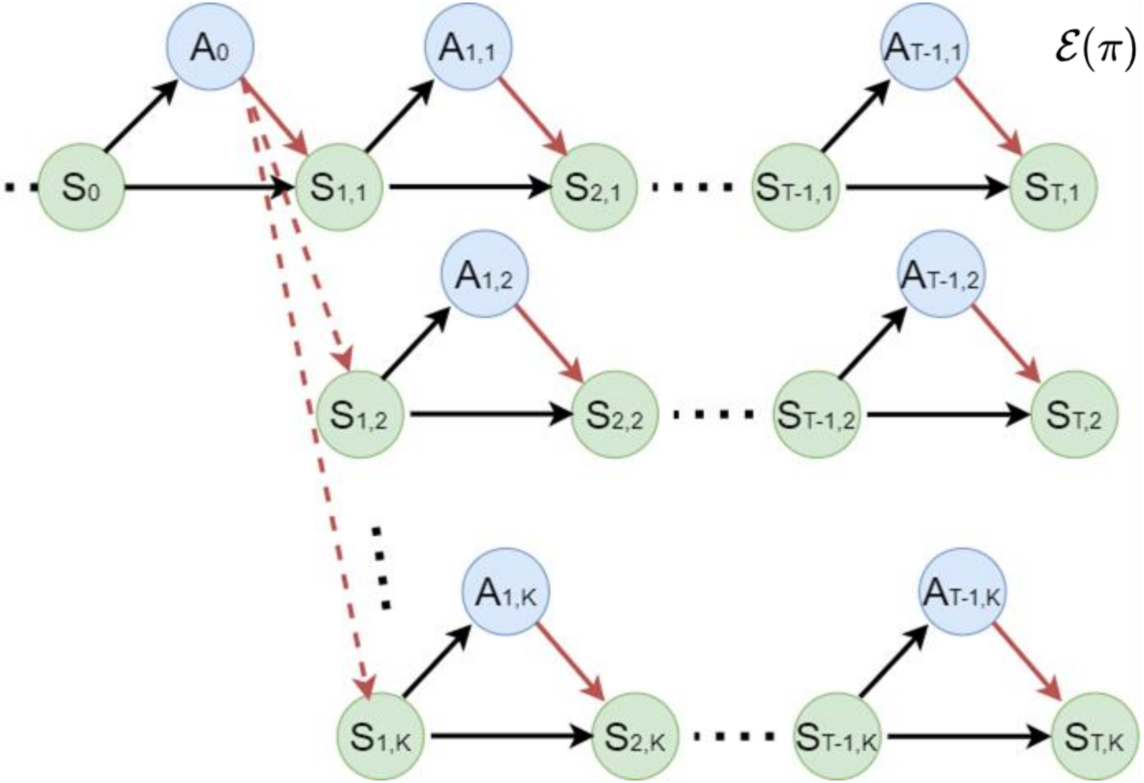
Method: Action Entropy Formulation



$$\mathcal{E}(\pi) = \mathrm{H}(A_{1:T} \mid s_0) - \mathrm{H}(A_{1:T} \mid S_{1:T}, s_0)$$
$$\approx \sum_{t=1}^{T} \Big(\frac{1}{K} \sum_{k=1}^{K} \ln \pi(a_t^k \mid s_t^k)$$
$$- \ln \frac{1}{K} \sum_{k=1}^{K} \pi(\tilde{a}_t \mid s_t^k) \Big)$$

where $\tilde{a}_t$ is resampled from the marginal $\pi(a_t | s_0, a_{1:t-1})$ for estimating the open-loop action entropy.

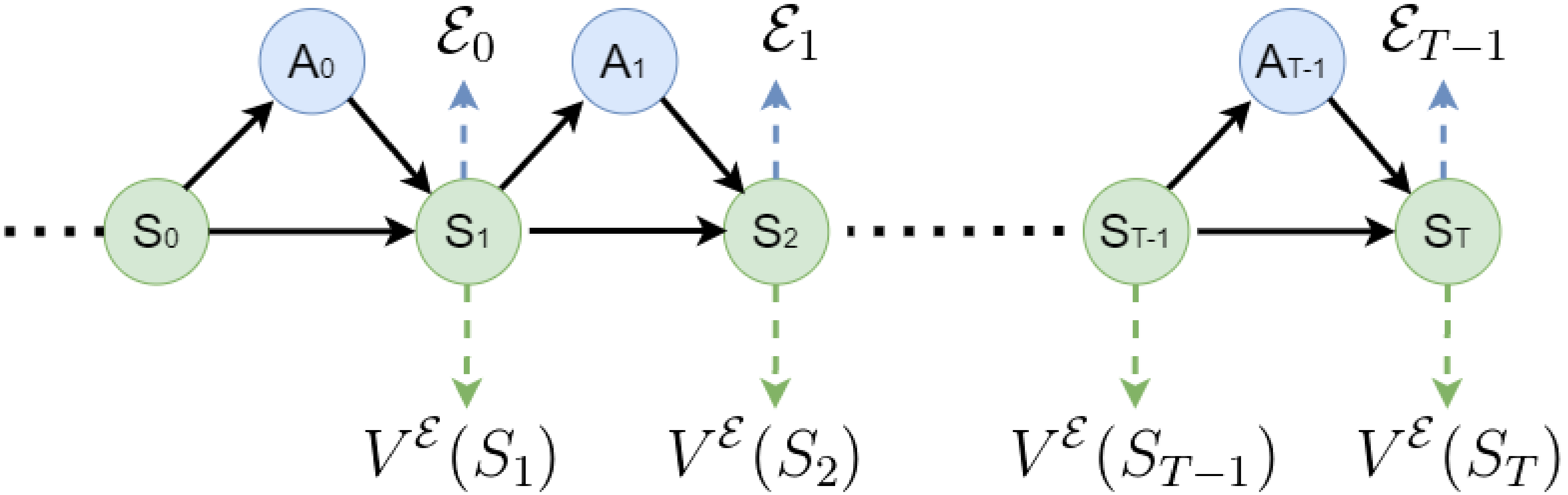
Method: State Entropy Formulation



$$\begin{aligned}\mathcal{E}(\pi) &= \mathrm{H}(S_{1:T}|s_0) - \mathrm{H}(S_{1:T} \mid A_{1:T}, s_0) \\ &\approx \sum_{t=1}^{T}\Big(\frac{1}{K}\sum_{k=1}^{K}\ln p(s_t^k \mid s_{t-1}^k, a_{t-1}^k) \\ &\qquad - \ln\frac{1}{K}\sum_{k=1}^{K} p(\tilde{s}_t \mid s_{t-1}^k, a_{t-1}^k)\Big)\end{aligned}$$

where $\tilde{s}_t$ is resampled from the marginal $p(s_t|s_0, a_{1:t-1})$ for estimating the open-loop state entropy.

Method: Value Learning for Empowerment



Overview

Empowerment

Contributions

Background

Methods

Experiments

Future work

Experiments

Environments:

- Six continuous control tasks of the DeepMind Control Suite.
- Agent is given only raw images as input.
- Challenging tasks: Hopper, Acrobot, Quadruped, etc from pixels.

Evaluation:

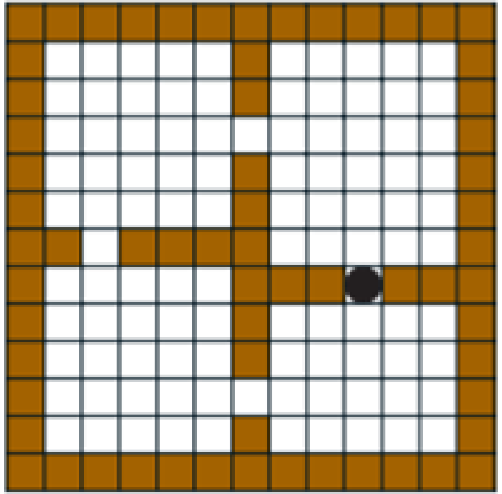
- Agent explores without task rewards and learns the world model.
- Then label experience with rewards to train a task policy in imagination.
- Direct evaluation on the task gives zero-shot performance.
- Additional greedy exploration for the task gives adaptation performance.

Demo: Walker after 5M frames

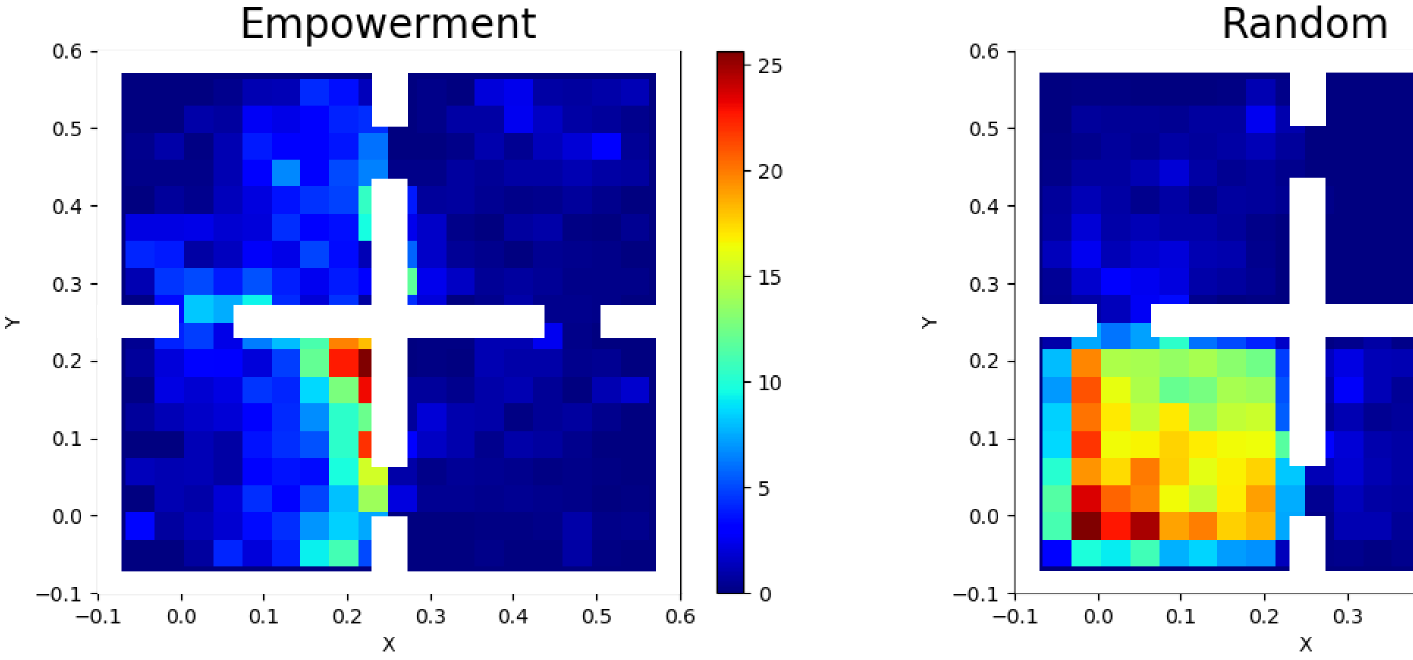
Empowerment

Random

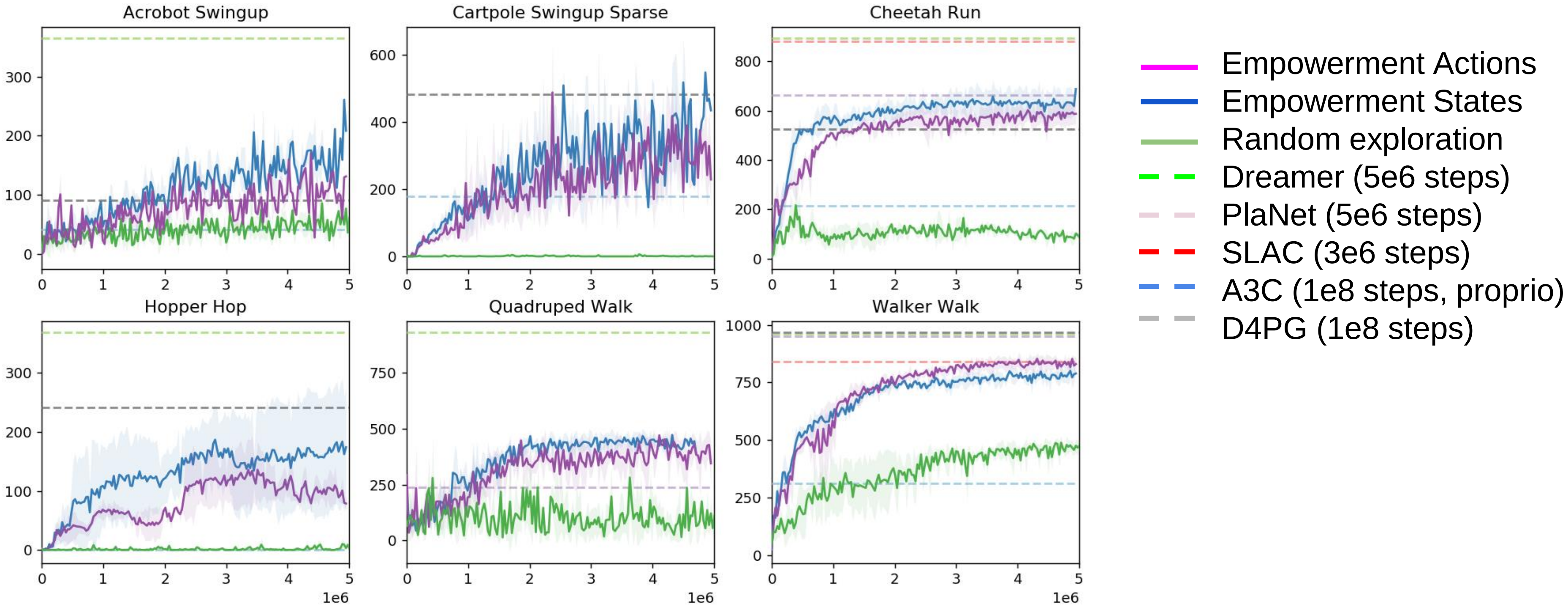
Demo: FourRoom



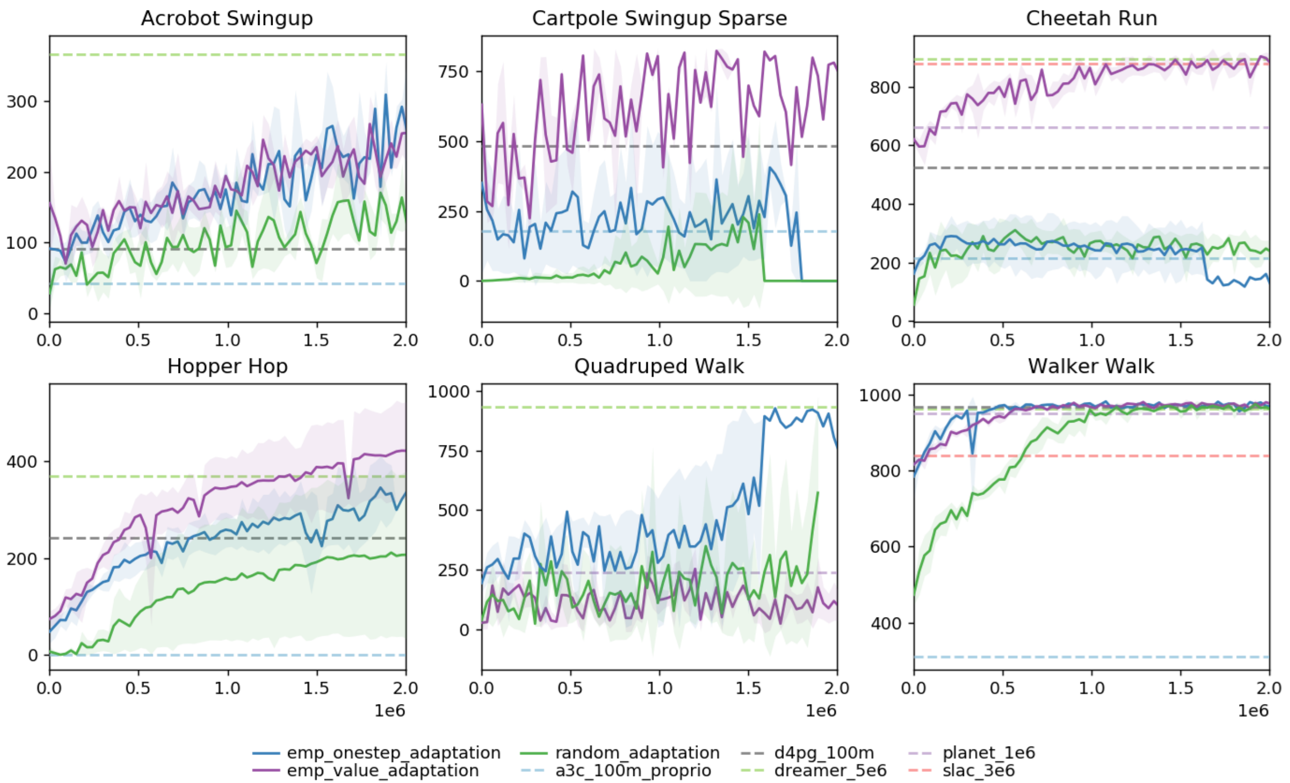
Demo: FourRoom

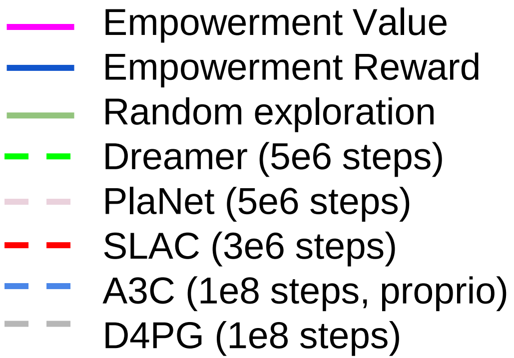


Zero-Shot performance (State vs Action)



Adaption Performance (One-Step vs Value Learning)





Future Work

We see temporal abstraction as critical for further improving exploration.

- For example, Quadruped learns many upside down movements but is less interested in getting on its feet.

Using Kolmogorov Mutual Information as the intrinsic motivation. This requires to track the complexity of a neural network, which is an under-explored area.

References

[1] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Empowerment: A universal agent-centric measure of control. In 2005 IEEE Congress on Evolutionary Computation, volume 1, pages 128–135.IEEE, 2005.

[2] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 16–17, 2017.

[3] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In Advances in neural information processing systems, pages 1471–1479, 2016.

References

[4] B. C. Stadie, S. Levine, and P. Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models.arXiv preprint arXiv:1507.00814, 2015.

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. arXiv preprint arXiv:1810.12894, 2018.

[6] Pathak, D., Gandhi, D., and Gupta, A. Self-Supervised Exploration via Disagreement. In Proceedings of the 36th International Conference on Machine Learning, 2019.

References

[7] T. Jung, D. Polani, and P. Stone. Empowerment for continuous agent—environment systems.Adaptive Behavior, 19(1):16–39, 2011.

[8] K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control.arXiv preprintarXiv:1611.07507, 2016.

[9] M. Karl, M. Soelch, P. Becker-Ehmck, D. Benbouzid, P. van der Smagt, and J. Bayer. Unsuper-vised real-time control through variational empowerment.arXiv preprint arXiv:1710.05101,2017.

[10] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman. Dynamics-aware unsuperviseddiscovery of skills.arXiv preprint arXiv:1907.01657, 2019.