Hypothesis-Driven Exploration for Deep Reinforcement Learning

Caleb Chuck, Supawit Chockchowwat, and Scott Niekum

The University of Texas at Austin

Personal Autonomous Robotics Lab

Research Question

How can we explore efficiently by generating and testing physics-based hypotheses about controllable aspects of the environment?

Other work directs exploration using novelty or extrinsic reward

Count-based approaches [1,2,3]

- Use visitation counts to provide reward for novel states

Reachability-based methods [4,5]

- Use a distance metric to define and reach novel states

Reward-directed exploration [5,6]

- Explore towards where extrinsic reward might be higher

By contrast, exploration based on learning to control objects can greatly improve efficiency.

Core Assumptions

State is factorizable into recognizable objects

- Object Properties: $f^{o_i}(x_{\text{raw}}) \rightarrow x_{o_i}$ (position)
- Object relationships: $\pi^{\Delta x_{o_j}}_{A_{o_i}}(x_{\text{raw}}, a)$

Objects do not change unless acted upon

- Changepoints: $\{x^{(0)}_{o_i}, \ldots, x^{(T)}_{o_i}\} \rightarrow \{c_1, \ldots, c_m\}$
- Segment displacement model: $x^{(t)}_{o_i} + d \approx x^{(t+1)}_{o_i}$

Salient times help explain object changes

- Proximity: Object locations close together
- Attribute change: Changepoints in a different object

Limit search to controllable objects

- Contingency: Directly control by raw actions, or distal control via a different contingent object

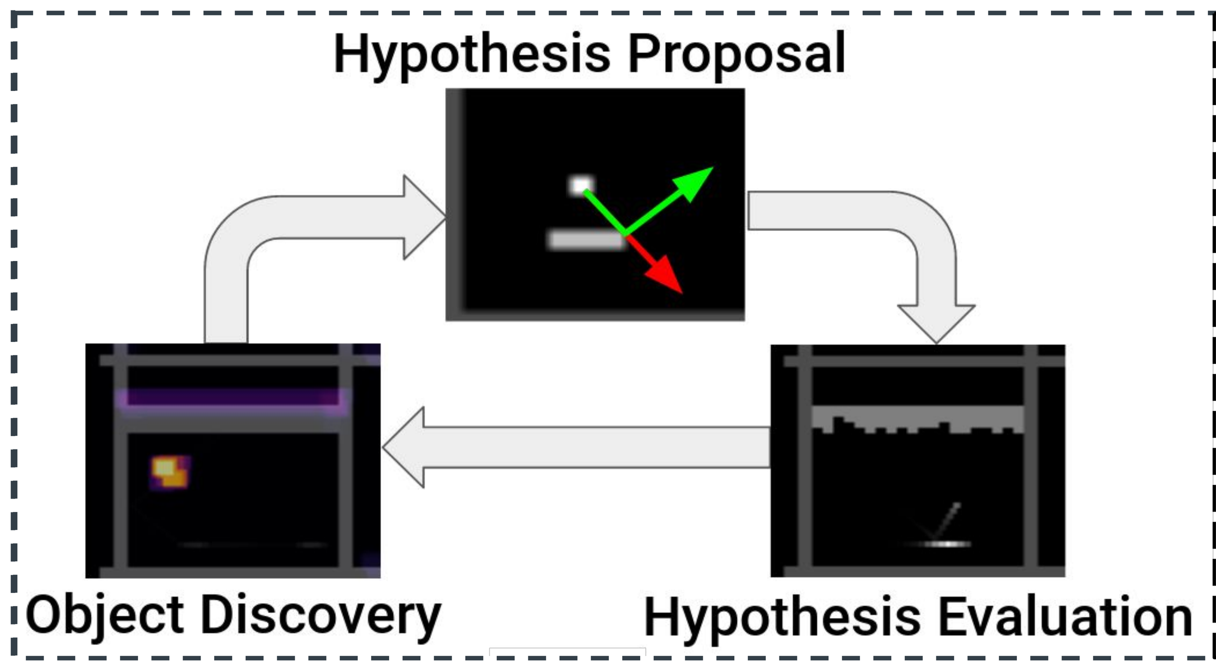
References

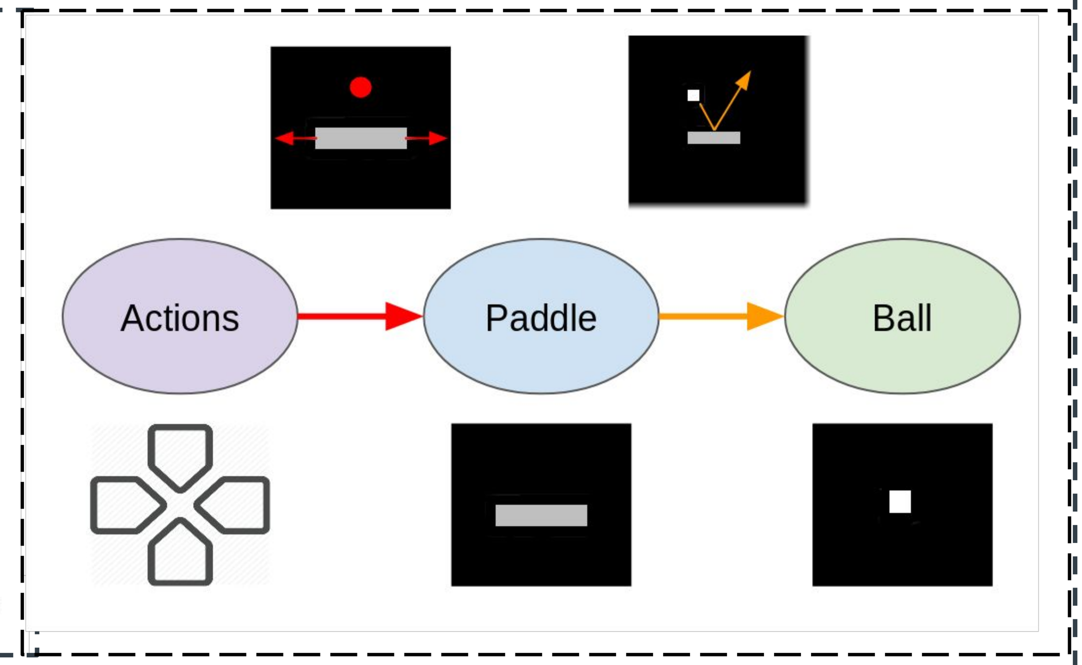
[1] Ostrovski et al. "Count-based exploration with neural density models." *ICLR*, 2017.
[2] Tang et al. "# exploration: A study of count-based exploration for deep reinforcement learning. " *NIPS*. 2017.
[3] Burda et al. "Exploration by random network distillation" *arXiv:*1810.12894,. 2018.
[4] Salinov, Raichuk, Marinier, Vincent et al. " Episodic curiosity through reachability." *ICLR*. 2019.
[5] Ecoffet, et al. "Go-Explore: a New Approach for Hard-Exploration Problems." *arXiv:1901.10995*. 2019.
[6] Hester et al. "Real Time Targeted Exploration in Large Domains." *ICDL*. 2010.
[7] Lowry et al. " Plan online, learn offline: Efficient learning and exploration via model-based control." *ICLR*. 2019.
[8] Hessel, Matteo, et al. "Rainbow: Combining improvements in deep reinforcement learning." *AAAI*, 2018.
[9] Schulman et al. "Proximal Policy Optimization Algorithms." *arXiv:1707.06347*. 2017.
[10] Mnih et al. "Asynchronous Methods for Deep Reinforcement Learning." *ICML*. 2016.
[11] Hansen, et al. "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)." *Evolutionary Computation* 2003.

Key Takeaway

We demonstrate an efficient exploration method which proposes and evaluates hypotheses about controllable object interactions, starting from raw pixels.

Hypothesis Proposal and Evaluation (HyPE) Loop





HyPE Loop

Object Discovery:
Learn a convolutional filter which indicates the location of an object. This filter is learned by searching for features which interact with existing objects.

Hypothesis Proposal:
Generate a set of hypotheses about different ways to control one object using another (or primitive actions). These hypotheses correspond to proposed object relationships.

Hypothesis Evaluation:
Learn to reproduce the hypotheses by rewarding hypothesized control in a reinforcement learning setting which uses states and actions defined by the related objects.

Example 1: Paddle

Paddle Discovery:
Starting from the Actions node, discover some object in the scene which has changepoints that correspond to changes in actions. This is the paddle.

Paddle Control Hypotheses:
Random-action data reveals that certain actions correspond to certain controls. HyPE generates hypotheses which represents right, left and 0 movement in the paddle

Paddle Hypothesis Evaluation:
Learn control policies that move the paddle right, left and 0.

Control-inducing Policy

Example 2: Ball

Ball Discovery:
HyPE searches for a new object from Actions and Paddle, and discovers a filter that exhibits changepoints when interacting with the Paddle, the ball.

Ball Control Hypotheses:
Past data from learning paddle control reveals a specific type of ball changepoint when it is slightly above the paddle. This leads to a near-paddle "bouncing" hypothesis.

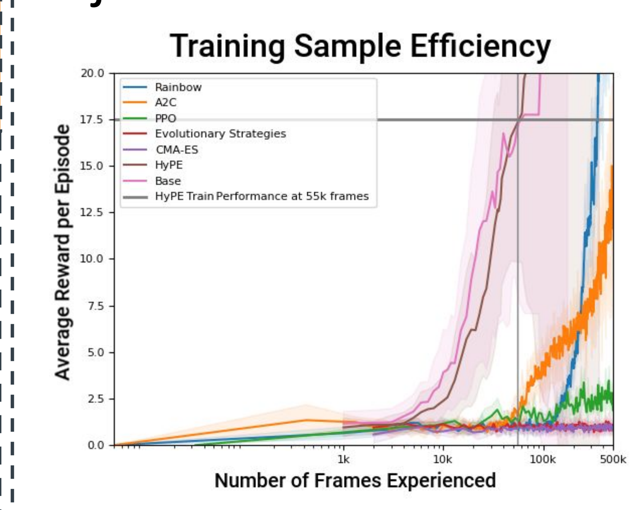
Ball Hypothesis Evaluation:
Learn a control policy to produce ball bounces

Changepoint-inducing Policy

Results

Learning Behaviors

HyPE achieves an order of magnitude improvement in sample efficiency when compared with Rainbow [7], PPO [8], A2C [9] and evolutionary strategies [10]. Most of this sample improvement appears to be from training with relative state between discovered objects.



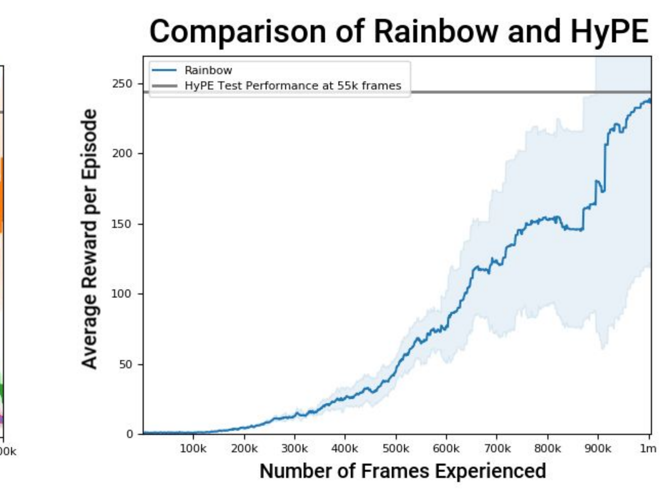
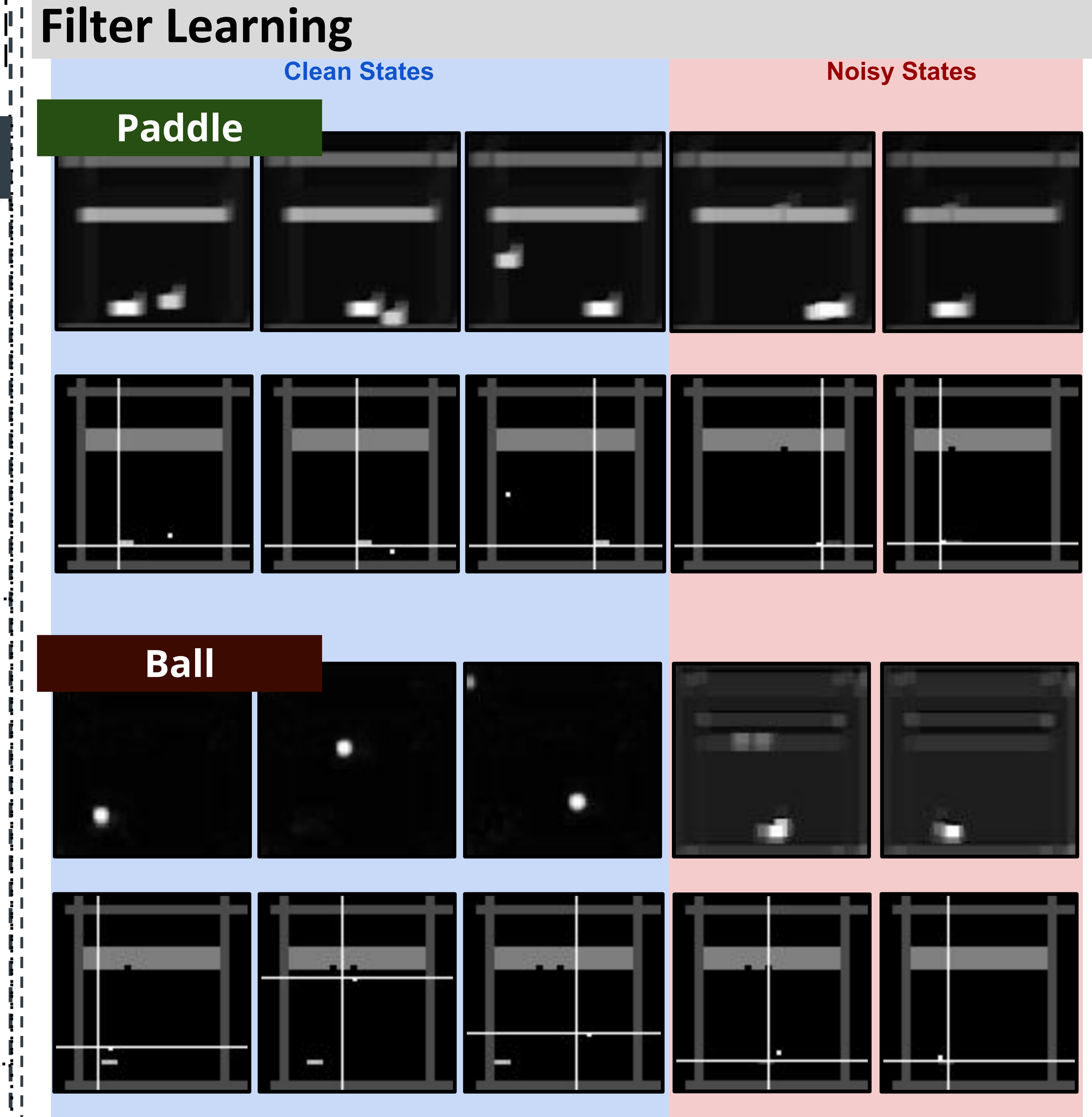


Table 1. Table of training time to find policy with evaluation score of 244 blocks hit, the average test score of HyPE after 55,500 frames of training (standard error 27, 20 trials).

Algorithm	HyPE	Rainbow	A2C & PPO
Timesteps	55,500	$\sim 1,000,000$	$> 1,500,000$

Filter Learning



Proposing object changepoints

Hypothesis	Δx_{o_j}	Δy_{o_j}
$H_{d_0}(x_{o_i}, x_{o_j})$	1.94	0.01
$H_{d_1}(x_{o_i}, x_{o_j})$	0.0	0.0
$H_{d_2}(x_{o_i}, x_{o_j})$	-1.88	0.0

Three hypotheses proposed for control over the paddle, by applying DP-GMMs, indicating 3 mean displacements. values in pixels

Hypothesis	$x_{o_i} - x_{o_j}$	$y_{o_i} - y_{o_j}$
$H(x_{o_i}, x_{o_j})$	-3.94	-2.87

Changepoint hypothesis for ball bouncing, indicating mean relative position between the ball and the paddle, after DP-GMM clustering.