Self-Supervised Exploration via Disagreement

Deepak Pathak*
UC Berkeley

Dhiraj Gandhi*
CMU

Abhinav Gupta
CMU, FAIR

ICML 2019
* equal contribution
Exploration – a major challenge!
Exploration – a major challenge!

- Pathak et.al. “Curiosity-driven Exploration by Self-supervised Exploration”. ICML 2017
Exploration – a major challenge!

- Pathak et.al. “Curiosity-driven Exploration by Self-supervised Exploration”. ICML 2017
Exploration – a major challenge!

Exploration – a major challenge!

- Bellemare et al. “Sample Inefficient [millions of samples].

Sample Inefficient

Simulation
Sample Inefficient

Simulation

Real Robots
Sample Inefficient

“Stuck” in Stochastic Envs

Simulation

Real Robots
Sample Inefficient

Real Robots

Simulation

“Stuck” in Stochastic Envs

Curiosity Exploration w/ Noisy TV & Remote

[Burda*, Edwards*, Pathak* et. al. ICLR’19]

[Juliani et.al., ArXiv’19]
Why inefficient?
current image $x_t$
policy network

\( \pi_\theta(x_t) \)

current image \( x_t \)

[Pathak et al. ICML, 2017]
action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$

[Pathak et al. ICML, 2017]
current image $x_t$

action $a_t$

policy network $\pi_\theta(x_t)$

next image $x_{t+1}$

[Pathak et al. ICML, 2017]
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$

[Pathak et al. ICML, 2017]
current image $x_t$

next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

Prediction Model $f(x_t, a_t)$

[Pathak et al. ICML, 2017]
current image $x_t$

action $a_t$

policy network $\pi_\theta(x_t)$

next image $x_{t+1}$

Prediction Model $f(x_t, a_t)$

[Pathak et al. ICML, 2017]
current image $x_t$

action $a_t$

policy network $\pi_\theta(x_t)$

next image $x_{t+1}$

Prediction Model $f(x_t, a_t)$

predicted next image $\hat{x}_{t+1}$

current image $x_t$

action $a_t$

[Pathak et al. ICML, 2017]
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$

$r_t^i = \|\hat{x}_{t+1} - x_{t+1}\|

Intrinsic Reward $r_t^i$

predicted next image $\hat{x}_{t+1}$

Prediction Model $f(x_t, a_t)$

current image $x_t$

action $a_t$

[Pathak et al. ICML, 2017]
next image $x_{t+1}$

action $a_t$

policy network $\pi_{\theta}(x_t)$

current image $x_t$

Intrinsic Reward $r_t^i = \|\hat{x}_{t+1} - x_{t+1}\|$

predicted next image $\hat{x}_{t+1}$

Prediction Model $f(x_t, a_t)$

current image $x_t$  action $a_t$

[Pathak et al. ICML, 2017]
Environment is “black-box” \( \rightarrow \) hard optimization

\[ r_t^i = \| \hat{x}_{t+1} - x_{t+1} \| \]

Intrinsic Reward \( r_t^i \)

predicted next image \( \hat{x}_{t+1} \)

Prediction Model \( f(x_t, a_t) \)

current image \( x_t \)

action \( a_t \)

policy network \( \pi_\theta(x_t) \)

action \( a_t \)

next image \( x_{t+1} \)
current image $x_t$

action $a_t$

next image $x_{t+1}$

policy network $\pi_\theta(x_t)$

REINFORCE $\rightarrow \max_\theta \mathbb{E} \left( \sum_{t=1}^{T} r^i_t \right)$

$\sum_{t=1}^{T} r^i_t$ = $\|\hat{x}_{t+1} - x_{t+1}\|$

Intrinsic Reward $r^i_t$

predicted next image $\hat{x}_{t+1}$

predicted next image $x_{t+1}$

current image $x_t$

action $a_t$

Prediction Model $f(x_t, a_t)$

[Pathak et al. ICML, 2017]
REINFORCE $\rightarrow \max_\theta \mathbb{E} \left( \sum_{t=1}^{T} r_t^i \right)$

$r_t^i = \|\hat{x}_{t+1} - x_{t+1}\|

\text{Intrinsic Reward} r_t^i$
current image $x_t$

next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

Intrinsic Reward $r^i_t$

Predicted next image $\hat{x}_{t+1}$

Prediction Model $f(x_t, a_t)$

REINFORCE $\max_\theta \mathbb{E} \left( \sum_{t=1}^{T} r^i_t \right)$

$r^i_t = \| \hat{x}_{t+1} - x_{t+1} \|$
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$

$$\min \|x_{t+1} - \hat{x}^1_{t+1}\| \quad \|x_{t+1} - \hat{x}^2_{t+1}\| \quad \|x_{t+1} - \hat{x}^n_{t+1}\|$$

$$f_1 \quad f_2 \quad \cdots \quad f_n$$

$x_t \ a_t$
next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$
current image $x_t$

action $a_t$

policy network $\pi_\theta(x_t)$

next image $x_{t+1}$

$r_t^i = \sigma \left\{ \frac{x_{t+1} - \hat{x}_{t+1}^1}{\min ||x_{t+1} - \hat{x}_{t+1}^1||, ||x_{t+1} - \hat{x}_{t+1}^2||, ||x_{t+1} - \hat{x}_{t+1}^n||} \right\}$

$f_1(x_t, a_t) \quad f_2(x_t, a_t) \quad \cdots \quad f_n(x_t, a_t)$
Intrinsic Reward

\[ r^i_t = \sigma \left\{ \begin{array}{c}
\hat{x}_{t+1}^1 \\
\hat{x}_{t+1}^2 \\
\vdots \\
\hat{x}_{t+1}^n
\end{array} \right\} \]

\[ \min \left\| x_{t+1} - \hat{x}_{t+1}^1 \right\|, \left\| x_{t+1} - \hat{x}_{t+1}^2 \right\|, \ldots, \left\| x_{t+1} - \hat{x}_{t+1}^n \right\| \]

action \( a_t \)

policy network \( \pi_\theta(x_t) \)

next image \( x_{t+1} \)

current image \( x_t \)
Intrinsic Reward

Disagreement

next image $x_{t+1}$

action $a_t$

policy network $\pi_\theta(x_t)$

current image $x_t$

$$r^i_t = \sigma \left\{ \begin{array}{c}
\hat{x}^1_{t+1} \\
\hat{x}^2_{t+1} \\
\vdots \\
\hat{x}^n_{t+1}
\end{array} \right\}$$

$$\min \|x_{t+1} - \hat{x}^1_{t+1}\|, \|x_{t+1} - \hat{x}^2_{t+1}\|, \|x_{t+1} - \hat{x}^n_{t+1}\|$$
Deterministic Environments

performs as well as state-of-the-art methods
Deterministic Environments

performs as well as state-of-the-art methods
Deterministic Environments

performs as well as state-of-the-art methods
Deterministic Environments

performs as well as state-of-the-art methods
Stochastic Environments
Every model’s goes to mean $\rightarrow$ variance drops $\rightarrow$ unstuck
Stochastic Environments: 3D Navigation

Every model’s goes to mean $\rightarrow$ variance drops $\rightarrow$ unstuck

- w/o TV
- Noisy TV w/ Remote

Reward (not for training)

Number of Frames (in millions)
Stochastic Environments: 3D Navigation

Every model’s goes to mean $\rightarrow$ variance drops $\rightarrow$ unstuck
current state $x_t$

next state $x_{t+1}$

action $a_t$

policy network $\pi_{\theta}(x_t)$

$r_t^i = \sigma \left( \min \left\{ \|x_{t+1} - \hat{x}_{t+1}^1\|, \|x_{t+1} - \hat{x}_{t+1}^2\|, \|x_{t+1} - \hat{x}_{t+1}^n\| \right\} \right)$

Disagreement

Curiosity Reward

$f_1$ $f_2$ $\cdots$ $f_n$
Disagreement

next state \( x_{t+1} \)

action \( a_t \)

policy network \( \pi_\theta(x_t) \)

current state \( x_t \)

Curiosity Reward

\[ r^i_t = \sigma \left\{ \begin{array}{c} f_1 \\hat{x}^1_{t+1} \\ f_2 \\hat{x}^2_{t+1} \\ \vdots \\ f_n \\hat{x}^n_{t+1} \end{array} \right\} \]

\[ \min \| x_{t+1} - \hat{x}^1_{t+1} \|, \| x_{t+1} - \hat{x}^2_{t+1} \|, \ldots, \| x_{t+1} - \hat{x}^n_{t+1} \| \]
\[ r_t^i \triangleq \mathbb{E}_\theta \left[ \| f(x_t, a_t; \theta) - \mathbb{E}_\theta[f(x_t, a_t; \theta)] \|_2 \right] \]
Disagreement

\[ r_t^i \triangleq \mathbb{E}_\theta \left[ \| f(x_t, a_t; \theta) - \mathbb{E}_\theta[f(x_t, a_t; \theta)] \|^2_2 \right] \]

No dependency on the environment!
Differentiable Exploration

\[ r_t^i \triangleq \mathbb{E}_\theta \left[ \| f(x_t, a_t; \theta) - \mathbb{E}_\theta [f(x_t, a_t; \theta)] \|_2^2 \right] \]

No dependency on the environment!
Differentiable Exploration

Differentiable Exploration

\[
\min_{\theta_1, \ldots, \theta_k} \sum_{i=1}^{k} \left\| f_{\theta_i}(x_t, \pi(x_t; \theta_P)) - x_{t+1} \right\|_2
\]

Model Optimization

---

Differentiable Exploration

Model Optimization

\[
\min_{\theta_1, \ldots, \theta_k} \sum_{i=1}^{k} \left\| f_{\theta_i}(x_t, \pi(x_t; \theta_P)) - x_{t+1} \right\|_2
\]

Policy Optimization

\[
\max_{\theta_p} \sum_{i=1}^{k} \left\| f_{\theta_i}(x_t, \pi(x_t; \theta_P)) - \left(\frac{1}{k}\right) \sum_{j=1}^{k} f_{\theta_j}(x_t, \pi(x_t; \theta_P)) \right\|_2
\]

Differentiable Exploration

Differentiable Exploration

Differentiable Exploration

Differentiable Exploration

Position Control:

1. Position
2. Direction
3. Gripper Angle
4. Gripper Distance

Differentiable Exploration

Efficiency over REINFORCE

Object Interaction Rate

Training Samples

Differentiable Exploration

Efficiency over REINFORCE

Differentiable Exploration

Efficiency over REINFORCE


Pushing skill
Differentiable Exploration

Efficiency over REINFORCE

Pushing skill

Picking skill

Summary: Exploration via Disagreement
Summary: Exploration via Disagreement

- Similar to state-of-the-art in deterministic envs
  (Atari games)
Summary: Exploration via Disagreement

- Similar to state-of-the-art in deterministic envs
  (Atari games)

- Does not get stuck in stochastic scenarios
  (Stochastic Atari; Unity-TV)
Summary: Exploration via Disagreement

- Similar to state-of-the-art in deterministic envs
  (Atari games)

- Does not get stuck in stochastic scenarios
  (Stochastic Atari; Unity-TV)

- Differentiable reformulation for real robots
  (Sawyer Robot)
Code Available

https://pathak22.github.io/exploration-by-disagreement/
Thank you!