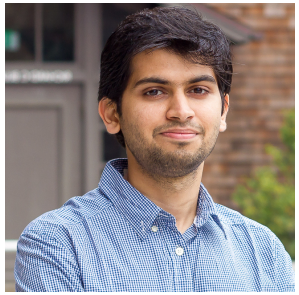
Self-Supervised Exploration via Disagreement



Deepak Pathak*
UC Berkeley

Dhiraj Gandhi*
CMU

Abhinav Gupta
CMU, FAIR

ICML 2019

* equal contribution

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Carnegie Mellon University

Exploration – a major challenge!

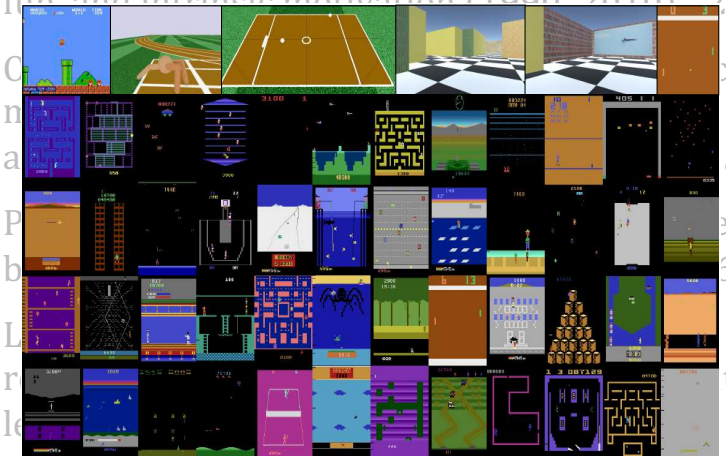
Exploration – a major challenge!

- Schmidhuber, Jurgen. “A possibility for implementing curiosity and boredom in model building neural controllers”, 1991.
- Schmidhuber, Jurgen. “Formal theory of creativity, fun, and intrinsic motivation (1990–2010)”, 2010.
- Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. Frontiers in neurorobotics, 2009.
- Poupart *et.al.* “An analytic solution to discrete bayesian reinforcement learning”. ICML, 2006.
- Lopes *et.al.* “Exploration in model-based reinforcement learning by empirically estimating learning progress”. NIPS, 2012.
- Bellemare *et.al.* “Unifying count-based exploration and intrinsic motivation”. NIPS, 2016.
- Mohamed *et.al.* “Variational information maximisation for intrinsically motivated reinforcement learning”. NIPS, 2015.
- Houthooft *et.al.* “VIME: Variational information maximizing exploration”. NIPS, 2016.
- Gregor *et.al.* “Variational intrinsic control”. ICLR Workshop, 2017.
- Pathak *et.al.* “Curiosity-driven Exploration by Self-supervised Exploration”. ICML 2017
- Ostrovski *et.al.* “Count-based exploration with neural density models”. ICML, 2017.
- Burda*, Edwards*, Pathak* *et.al.* “Large-Scale Study of Curiosity-driven Learning”. ICLR 2019
- Eysenbach et al. “Diversity is all you need: Learn skills without a reward function”. ICLR 2019.
- Savinov et al. "Episodic curiosity through reachability". ICLR 2019.

Exploration – a major challenge!

- Schmidhuber, Jurgen. “A possibility for implementing curiosity and boredom in model building neural controllers”, 1991.
- Schmidhuber, Jurgen. “Formal theory of creativity, fun, and intrinsic motivation (1990–2010)”, 2010.
- Oudeyer, P.-Y. and Kaplan, F. What is intrinsic motivation? a typology of computational approaches. Frontiers in neurorobotics, 2009.
- Poupart *et.al.* “An analytic solution to discrete bayesian reinforcement learning”. ICML, 2006.
- Lopes *et.al.* “Exploration in model-based reinforcement learning by empirically estimating learning progress”. NIPS, 2012.
- Bellemare *et.al.* “Unifying count-based exploration and intrinsic motivation”. NIPS, 2016.
- Mohamed *et.al.* “Variational information maximisation for intrinsically motivated reinforcement learning”. NIPS, 2015.
- Houthooft *et.al.* “VIME: Variational information maximizing exploration”. NIPS, 2016.
- Gregor *et.al.* “Variational intrinsic control”. ICLR Workshop, 2017.
- Pathak *et.al.* “Curiosity-driven Exploration by Self-supervised Exploration”. ICML 2017
- Ostrovski *et.al.* “Count-based exploration with neural density models”. ICML, 2017.
- Burda*, Edwards*, Pathak* *et.al.* “Large-Scale Study of Curiosity-driven Learning”. ICLR 2019
- Eysenbach et al. “Diversity is all you need: Learn skills without a reward function”. ICLR 2019.
- Savinov et al. "Episodic curiosity through reachability". ICLR 2019.

Exploration – a major challenge!



- Bellemare *et.al.* “Unifying count-based exploration and intrinsic motivation”. NIPS, 2016.
- Mohamed *et.al.* “Variational information maximisation for intrinsically motivated reinforcement learning”. NIPS, 2015.
- Houthooft *et.al.* “VIME: Variational information maximizing exploration”. NIPS, 2016.
- Gregor *et.al.* “Variational intrinsic control”. ICLR Workshop, 2017.
- Pathak *et.al.* “Curiosity-driven Exploration by Self-supervised Exploration”. ICML 2017
- Ostrovski *et.al.* “Count-based exploration with neural density models”. ICML, 2017.
- Burda*, Edwards*, Pathak* *et.al.* “Large-Scale Study of Curiosity-driven Learning”. ICLR 2019
- Eysenbach et al. “Diversity is all you need: Learn skills without a reward function”. ICLR 2019.
- Savinov et al. "Episodic curiosity through reachability". ICLR 2019.

Exploration – a major challenge!

- Schmidhuber, Jurgen. “A possibility for implementing curiosity and boredom in model building neural controllers”, 1991.
- Schmidhuber, Jurgen. “Formal theory of creativity, fun, and intrinsic motivation (1990–2010)”, 201
- Oudeyer, P.-Y. and Kaplan, F. What is intrin motivation? a typology of computation approaches. Frontiers in neurorobo
- Poupart *et.al.* “An analytic sol bayesian reinforcement l
- Lopes *et.al.* “Explor reinforcement l imating learning prog
- Bellemare *et.al.* -based exploration and intrinsic moti IPS, 2016.
- Mohamed *et* formation maximis otivated reinf , 2015.
- ariational information ion”. NIPS, 2016.
- ariational intrinsic control”. ICLR , 2017.
- ak *et.al.* “Curiosity-driven Exploration by Self-upervised Exploration”. ICML 2017
- Ostrovski *et.al.* “Count-based exploration with neural density models”. ICML, 2017.
- Burda*, Edwards*, Pathak* *et.al.* “Large-Scale Study of Curiosity-driven Learning”. ICLR 2019
- Eysenbach et al. “Diversity is all you need: Learn skills without a reward function”. ICLR 2019.
- Savinov et al. "Episodic curiosity through reachability". ICLR 2019.

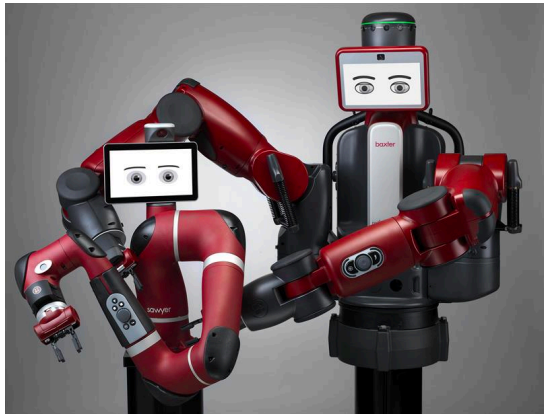
Sample Inefficient [millions of samples]

Sample Inefficient

Simulation

Sample Inefficient

Simulation

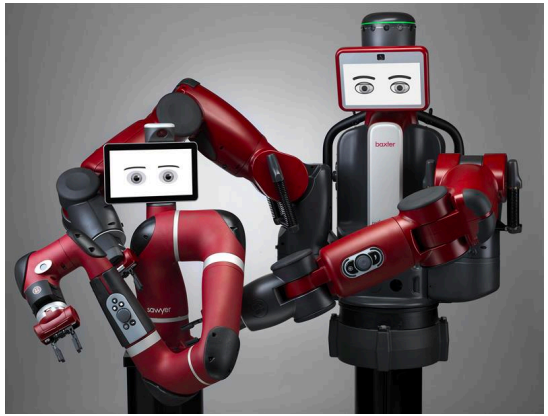


Real Robots

Sample Inefficient

✓

Simulation



✗

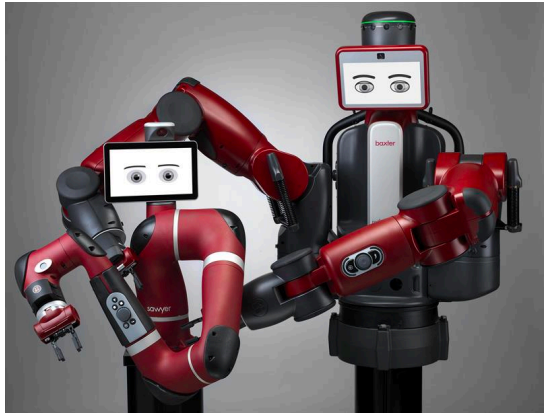
Real Robots

“Stuck” in Stochastic Envs

Sample Inefficient

✓

Simulation



✗

Real Robots

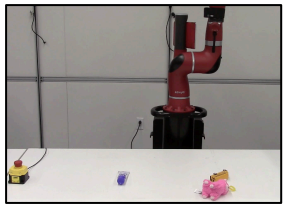
“Stuck” in Stochastic Envs

Curiosity Exploration
w/ Noisy TV & Remote

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

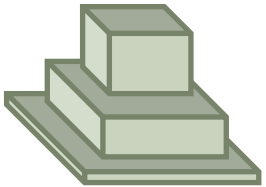
Why inefficient?

[Pathak et al. ICML, 2017]

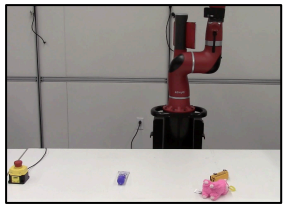


current image $\boldsymbol{x_t}$

[Pathak et al. ICML, 2017]

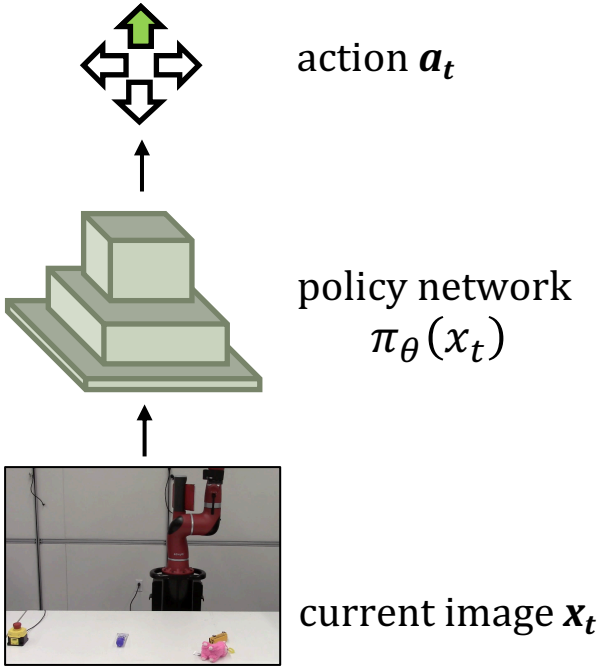


policy network
$\pi_\theta(x_t)$

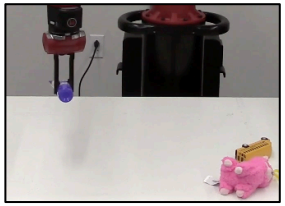


current image $\boldsymbol{x_t}$

[Pathak et al. ICML, 2017]

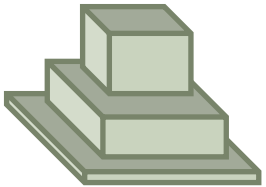


[Pathak et al. ICML, 2017]

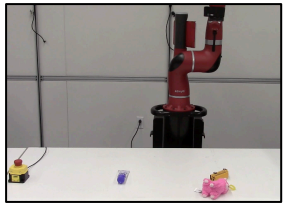


next image x_{t+1}

action a_t

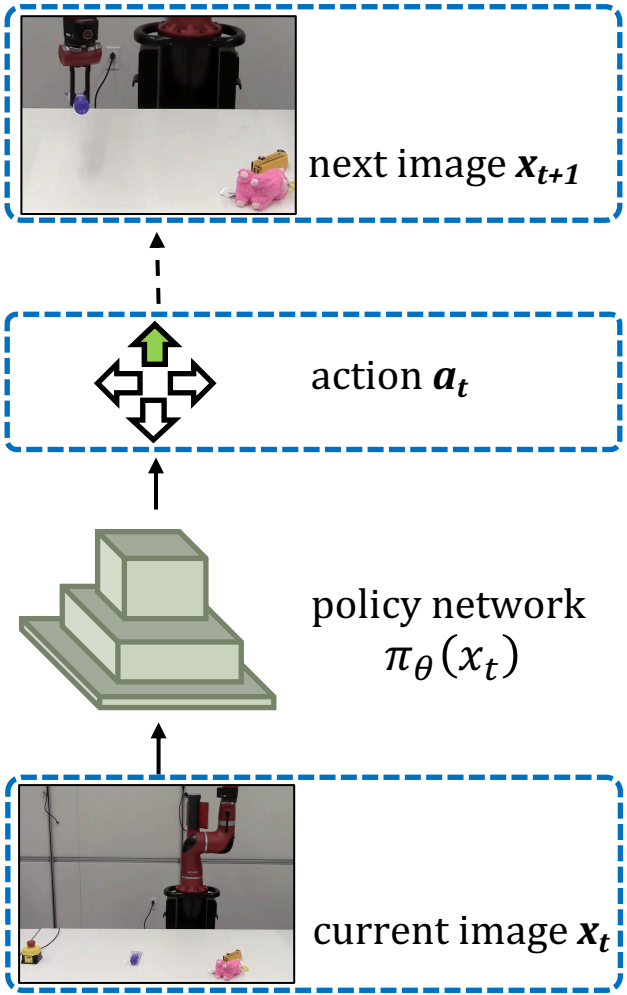


policy network
$\pi_\theta(x_t)$

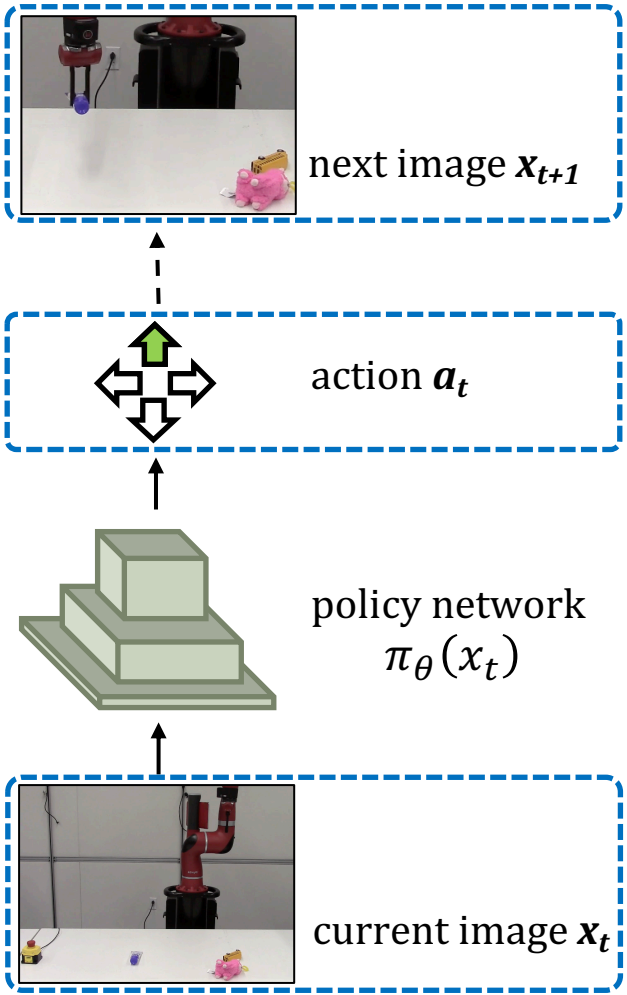


current image x_t

[Pathak et al. ICML, 2017]

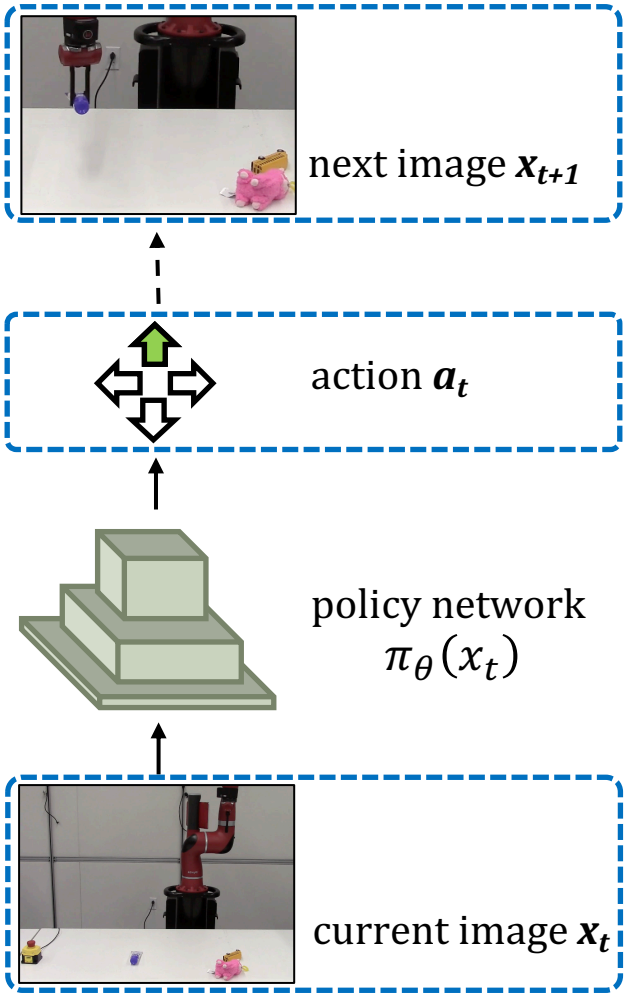


[Pathak et al. ICML, 2017]

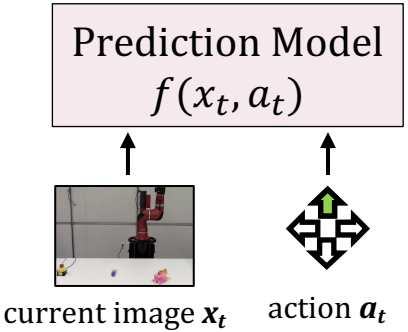


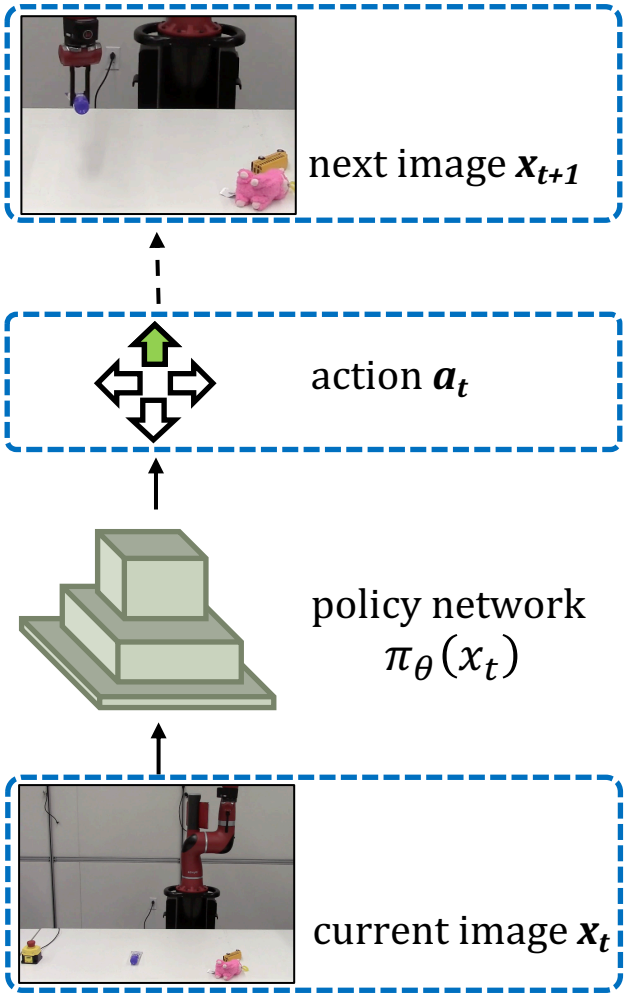
Prediction Model
$f(x_t, a_t)$

[Pathak et al. ICML, 2017]

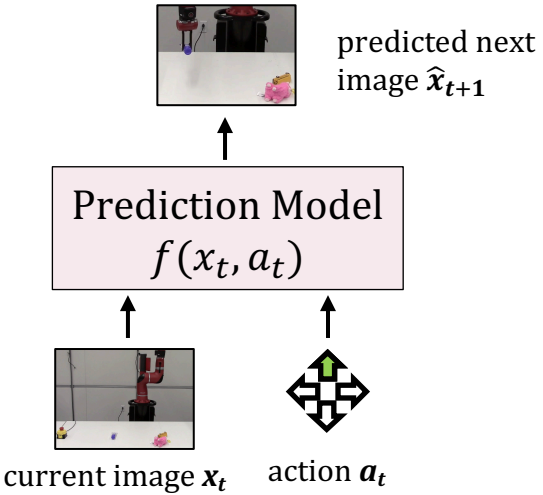


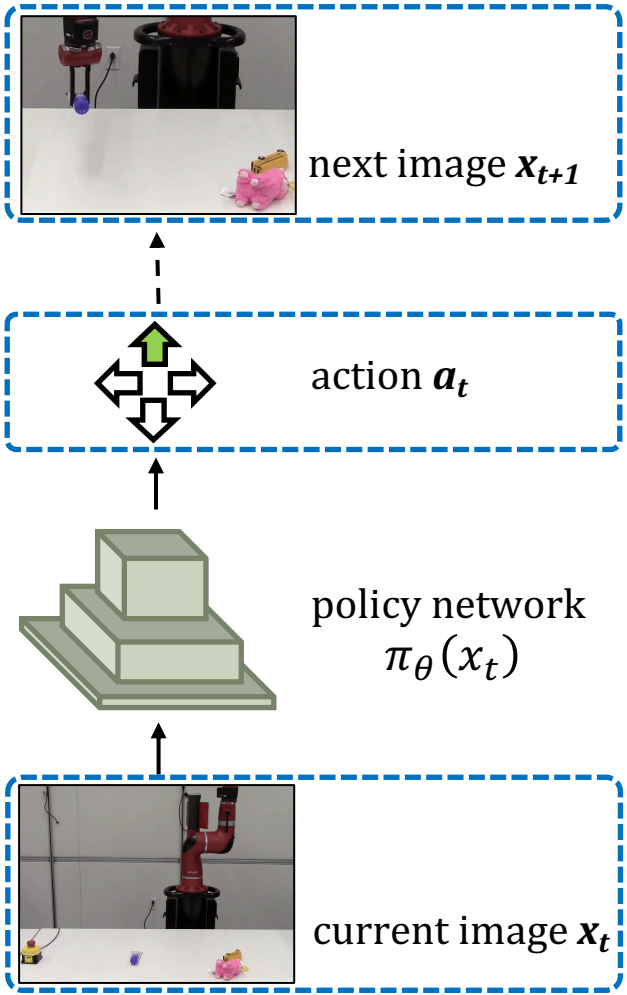
[Pathak et al. ICML, 2017]





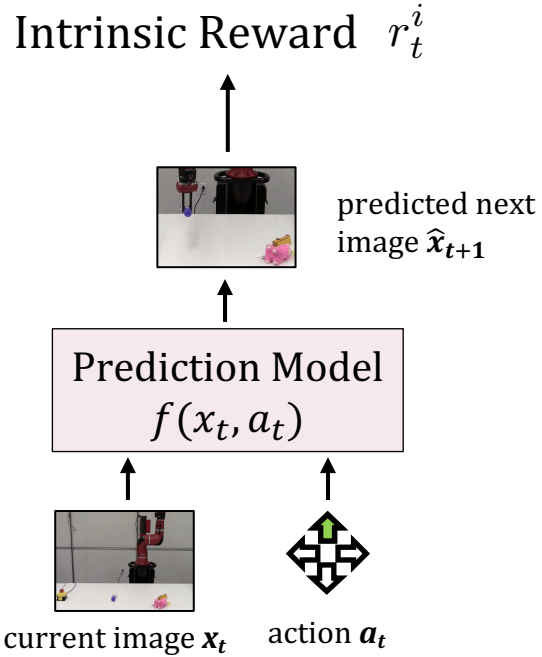
[Pathak et al. ICML, 2017]

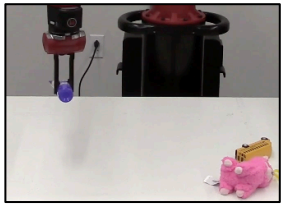




[Pathak et al. ICML, 2017]

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



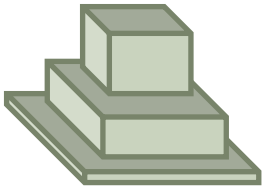


next image x_{t+1}

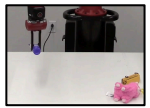
$$r_t^i = \|\hat{x}_{t+1} - x_{t+1}\|$$

action a_t

Intrinsic Reward r_t^i

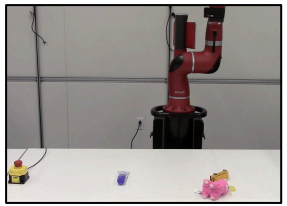


policy network
$\pi_\theta(x_t)$

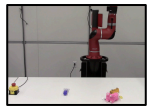


predicted next image $\hat{x}_{t+1}$

Prediction Model
$f(x_t, a_t)$



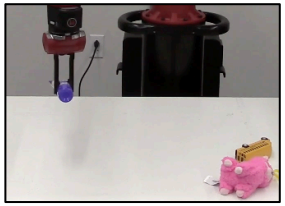
current image x_t



current image x_t

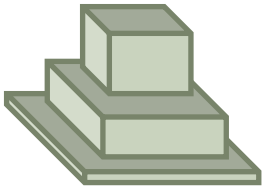
action a_t

[Pathak et al. ICML, 2017]

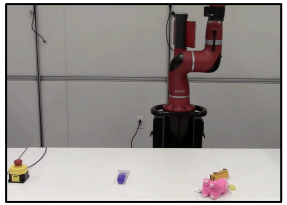


next image x_{t+1}

action a_t



policy network
$\pi_\theta(x_t)$

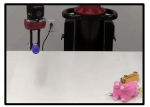


current image x_t

Environment is "black-box"
→ hard optimization

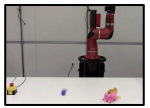
$$r_t^i = \|\hat{x}_{t+1} - x_{t+1}\|$$

Intrinsic Reward r_t^i



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

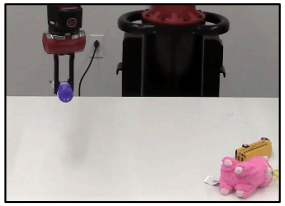
Prediction Model
$f(x_t, a_t)$



current image x_t

action a_t

[Pathak et al. ICML, 2017]

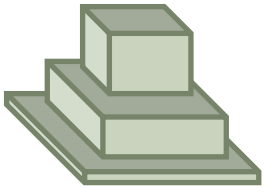


next image x_{t+1}

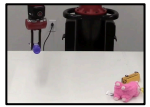
REINFORCE $\longrightarrow$ $\max_{\theta} \mathbb{E}\left(\sum_{t=1}^{T} r_t^i\right)$

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

action a_t $\longleftrightarrow$ Intrinsic Reward r_t^i

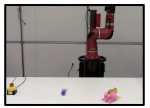


policy network
$\pi_\theta(x_t)$



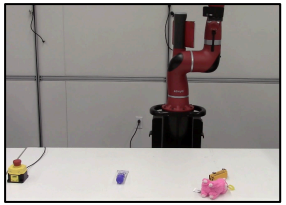
predicted next image $\hat{x}_{t+1}$

Prediction Model
$f(x_t, a_t)$



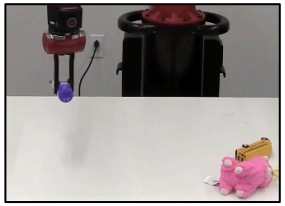
current image x_t

action a_t



current image x_t

[Pathak et al. ICML, 2017]

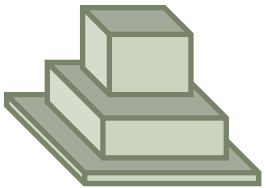


next image x_{t+1}

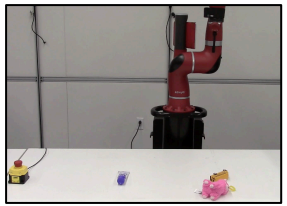
REINFORCE $\longrightarrow$ $\max_{\theta} \mathbb{E}\left(\sum_{t=1}^{T} r_t^i\right)$

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

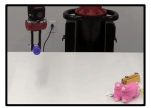
action a_t $\longleftrightarrow$ Intrinsic Reward r_t^i



policy network
$\pi_\theta(x_t)$

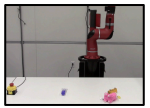


current image x_t



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

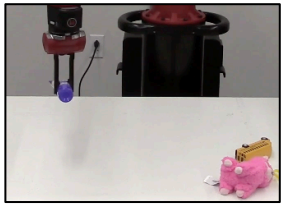
Prediction Model
$f(x_t, a_t)$



current image x_t

action a_t

[Pathak et al. ICML, 2017]



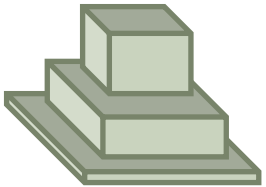
next image x_{t+1}

REINFORCE $\longrightarrow$ $\max_{\theta} \mathbb{E}\left(\sum_{t=1}^{T} r_t^i\right)$

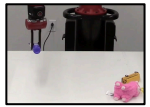
$r_t^i = \|\hat{x}_{t+1} - x_{t+1}\|$

action a_t

Intrinsic Reward r_t^i

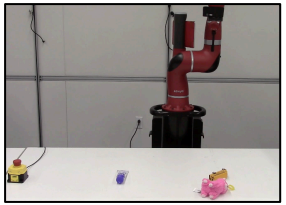


policy network
$\pi_\theta(x_t)$

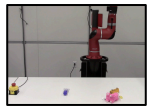


predicted next image $\hat{x}_{t+1}$

Prediction Model
$f(x_t, a_t)$



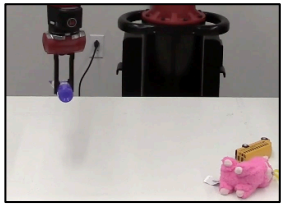
current image x_t



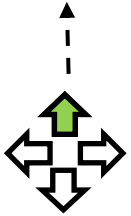
current image x_t

action a_t

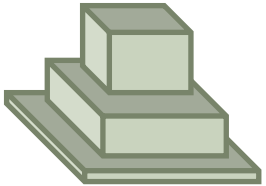
[Pathak et al. ICML, 2017]



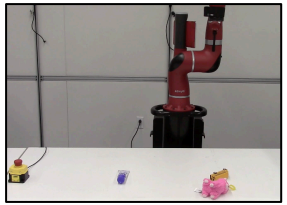
next image $\boldsymbol{x_{t+1}}$



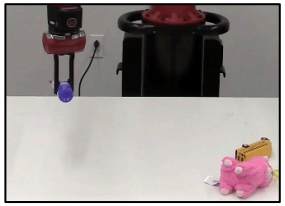
action $\boldsymbol{a_t}$



policy network
$\pi_\theta(x_t)$

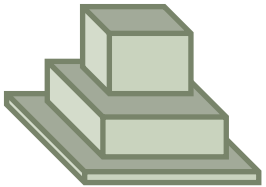


current image $\boldsymbol{x_t}$

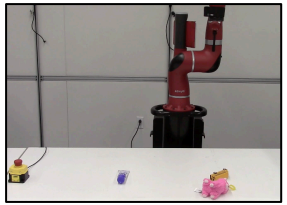


next image $\boldsymbol{x_{t+1}}$

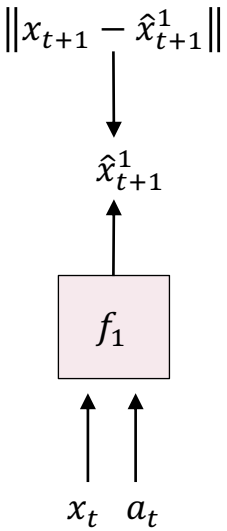
action $\boldsymbol{a_t}$

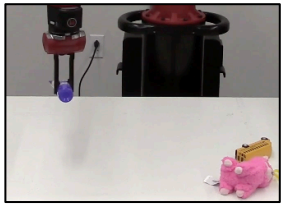


policy network
$\pi_\theta(x_t)$



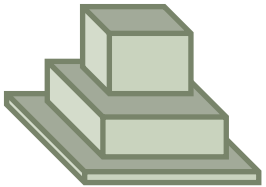
current image $\boldsymbol{x_t}$



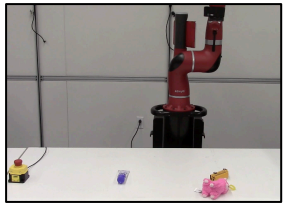


next image $\boldsymbol{x_{t+1}}$

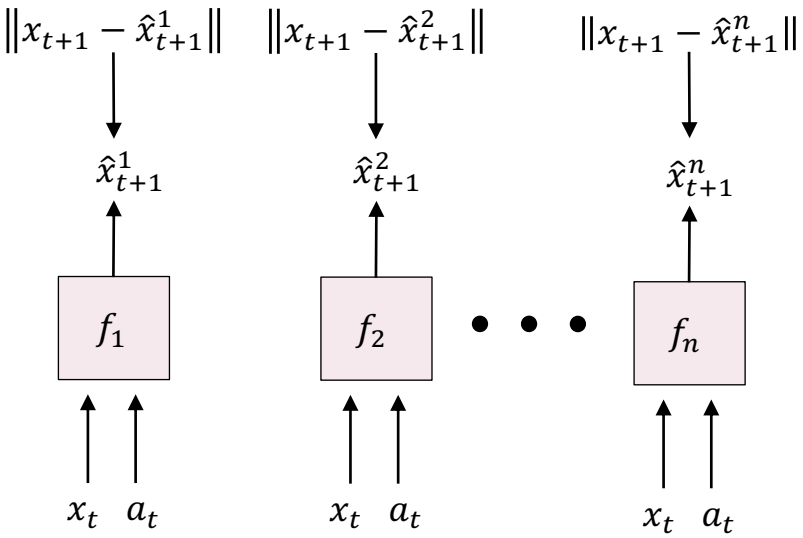
action $\boldsymbol{a_t}$

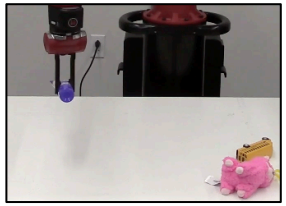


policy network
$\pi_\theta(x_t)$



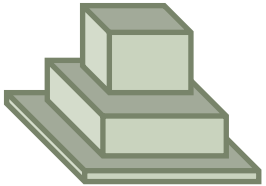
current image $\boldsymbol{x_t}$



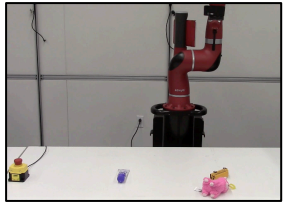


next image $\boldsymbol{x_{t+1}}$

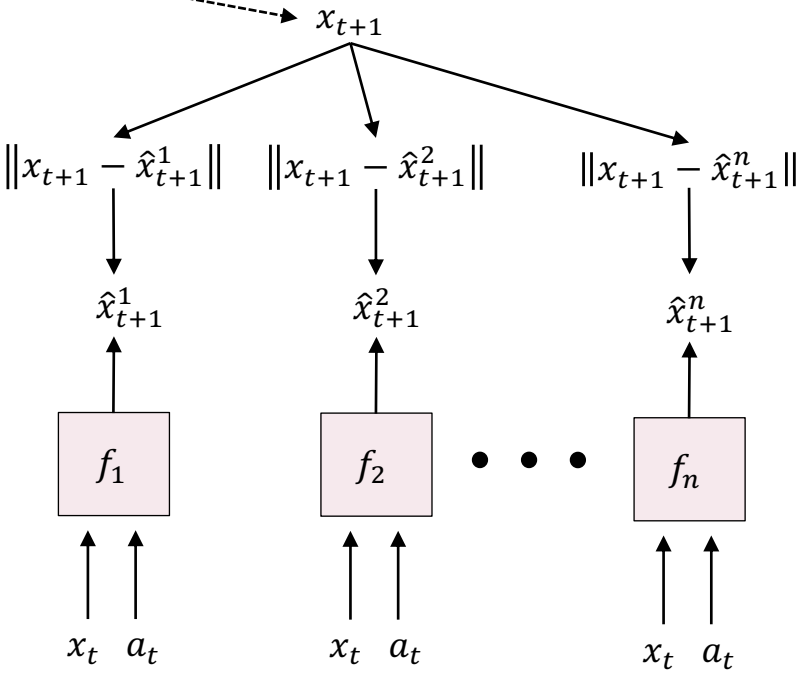
action $\boldsymbol{a_t}$

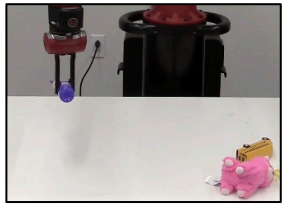


policy network
$\pi_\theta(x_t)$



current image $\boldsymbol{x_t}$

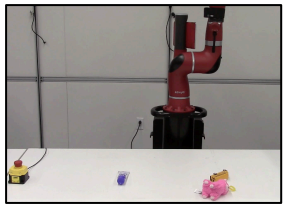




next image x_{t+1}

action a_t

policy network
$\pi_\theta(x_t)$



current image x_t

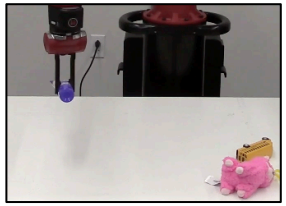
x_{t+1}

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

$\hat{x}^1_{t+1}$ $\hat{x}^2_{t+1}$ $\hat{x}^n_{t+1}$

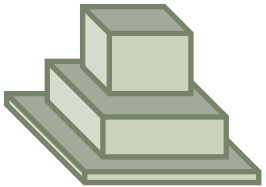
f_1 f_2 f_n

x_t a_t x_t a_t x_t a_t

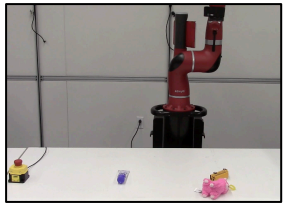


next image $\boldsymbol{x_{t+1}}$

action $\boldsymbol{a_t}$



policy network
$\pi_\theta(x_t)$



current image $\boldsymbol{x_t}$

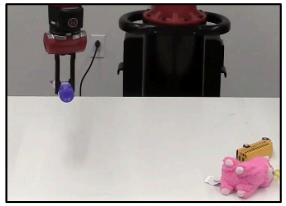
x_{t+1}

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

$\hat{x}^1_{t+1}$ $\hat{x}^2_{t+1}$ $\hat{x}^n_{t+1}$

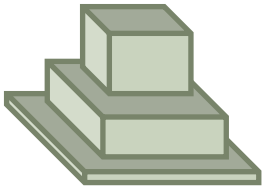
f_1 f_2 $\bullet\ \bullet\ \bullet$ f_n

x_t a_t x_t a_t x_t a_t

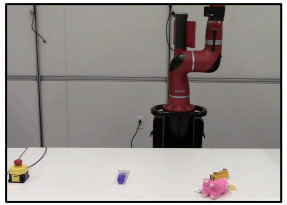


next image $\boldsymbol{x_{t+1}}$

action $\boldsymbol{a_t}$



policy network
$\pi_\theta(x_t)$



current image $\boldsymbol{x_t}$

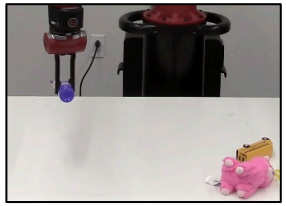
x_{t+1}

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

$r_t^i = \sigma\{\hat{x}^1_{t+1}\ \hat{x}^2_{t+1}\ \hat{x}^n_{t+1}\}$

f_1 f_2 $\bullet\ \bullet\ \bullet$ f_n

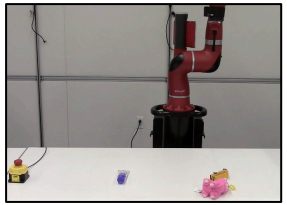
x_t a_t | x_t a_t | x_t a_t



next image x_{t+1}

action a_t

policy network
$\pi_\theta(x_t)$



current image x_t

$r_t^i = \sigma\{\}$

Intrinsic Reward

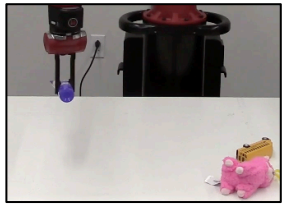
x_{t+1}

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

$\hat{x}_{t+1}^1$ $\hat{x}_{t+1}^2$ $\hat{x}_{t+1}^n$

f_1 f_2 $\bullet\bullet\bullet$ f_n

x_t a_t x_t a_t x_t a_t



next image x_{t+1}

action a_t

policy network
$\pi_\theta(x_t)$

current image x_t

Intrinsic Reward

$r_t^i = \sigma\{\}$

Disagreement

x_{t+1}

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

$\hat{x}_{t+1}^1$ $\hat{x}_{t+1}^2$ $\hat{x}_{t+1}^n$

f_1 f_2 $\bullet\bullet\bullet$ f_n

x_t a_t x_t a_t x_t a_t

Deterministic Environments

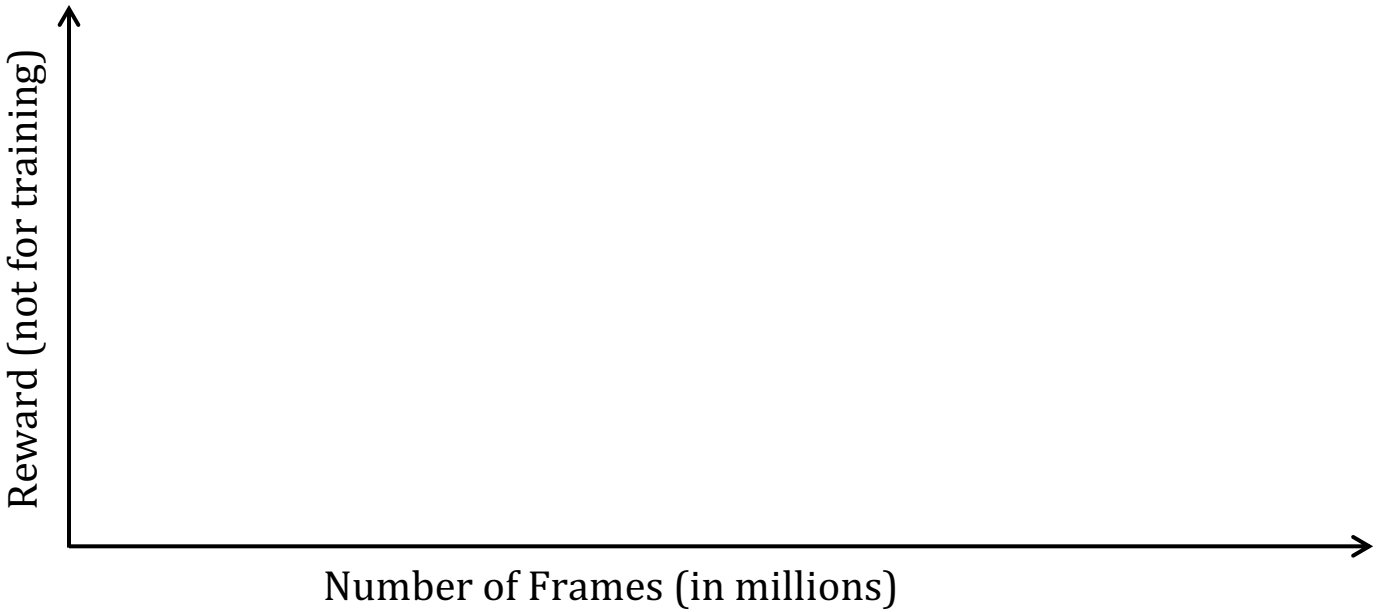
performs as well as state-of-the-art methods

Deterministic Environments

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

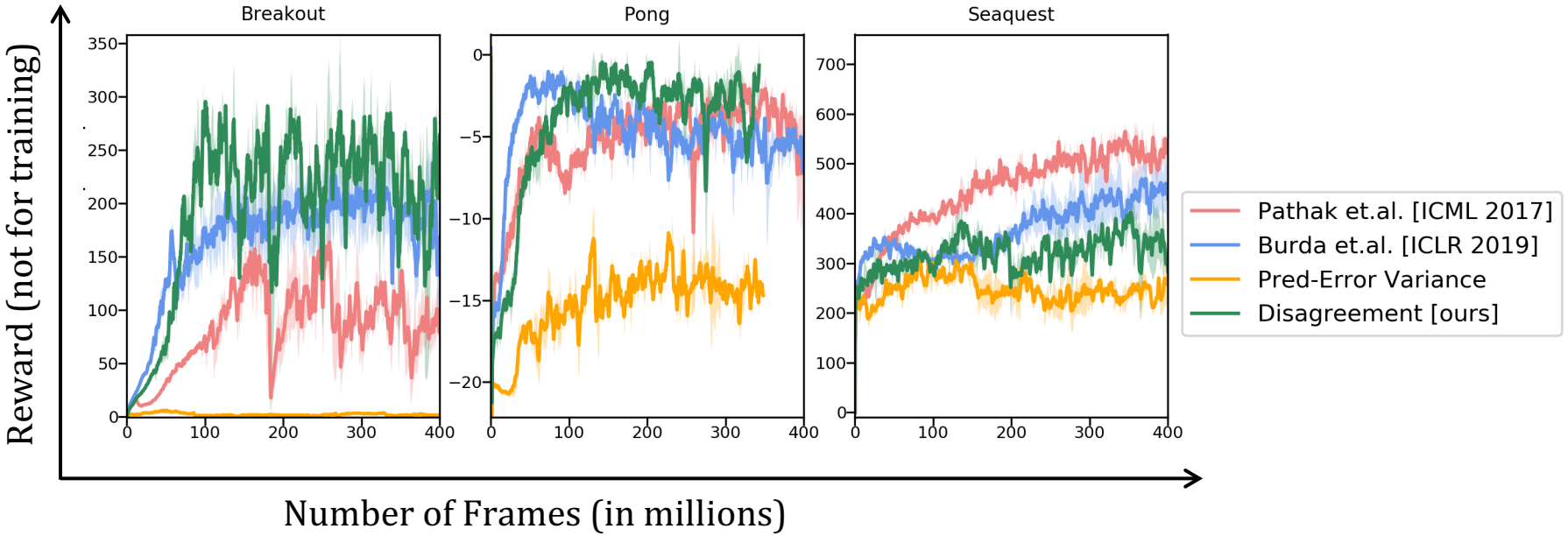
Deterministic Environments

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



Deterministic Environments

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



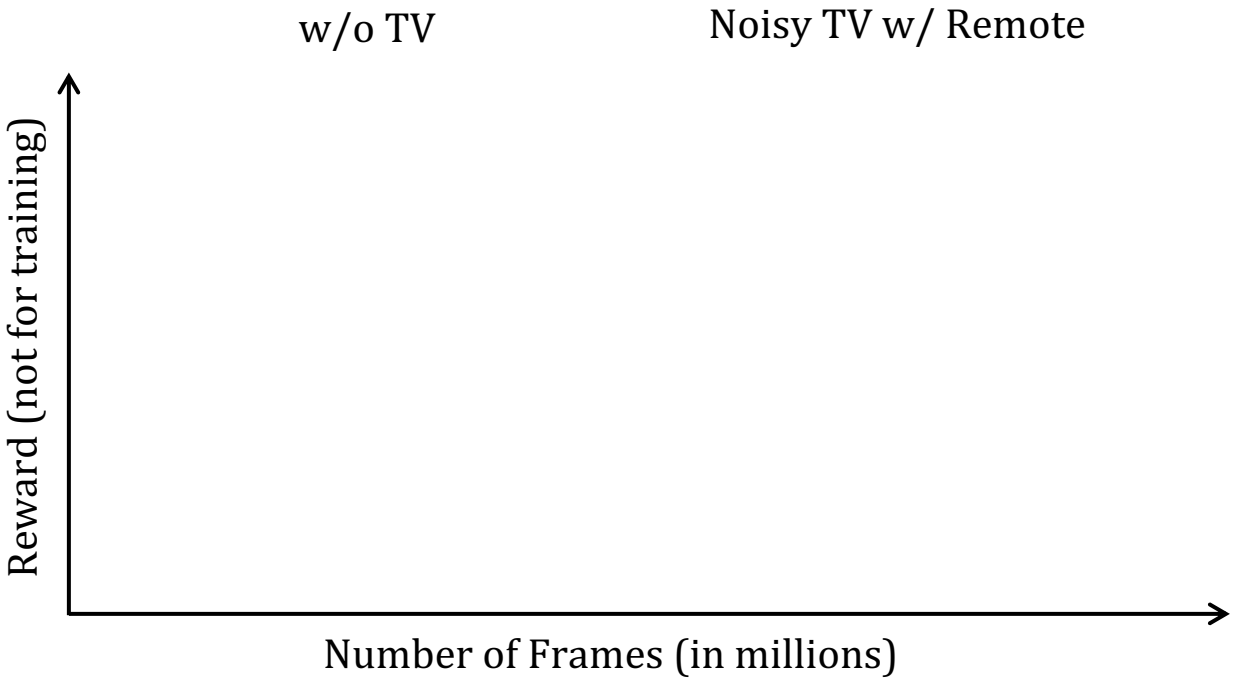
Stochastic Environments

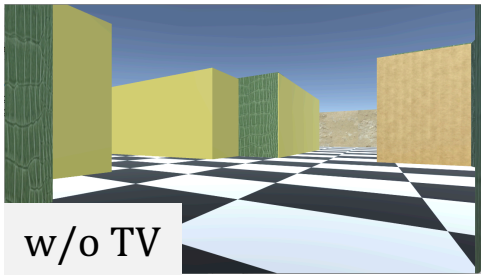
Stochastic Environments

Every model's goes to mean → variance drops → unstuck

Stochastic Environments: 3D Navigation

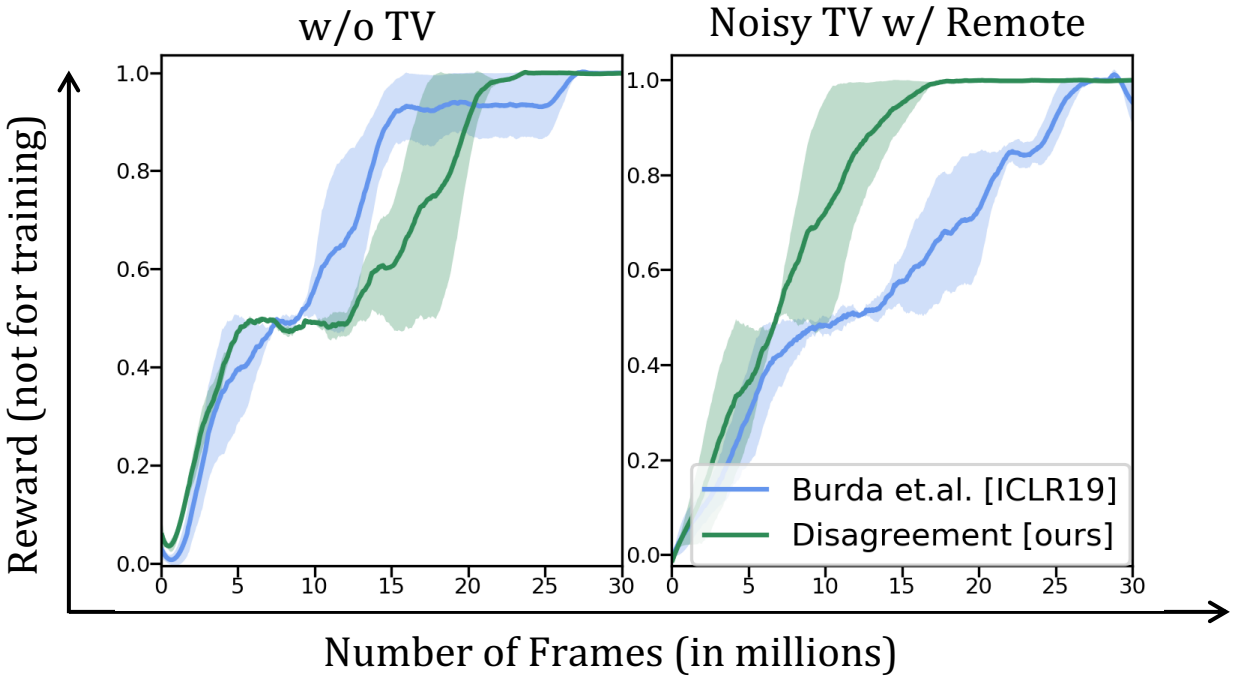
Every model's goes to mean → variance drops → unstuck

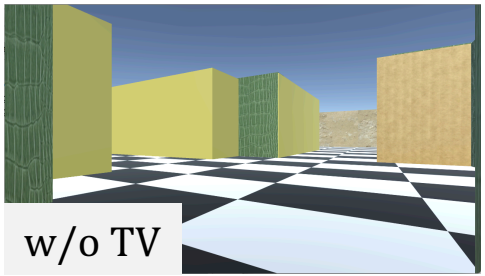


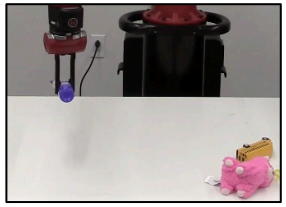


Stochastic Environments: 3D Navigation

Every model's goes to mean → variance drops → unstuck

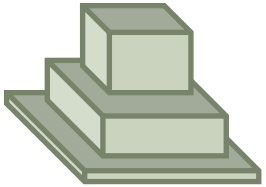




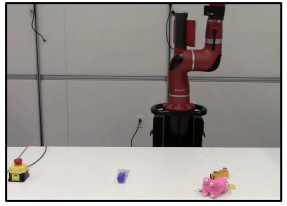


next state x_{t+1}

action a_t



policy network
$\pi_\theta(x_t)$



current state x_t

Curiosity
Reward

$r_t^i = \sigma\{\}$

Disagreement

x_{t+1}

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

$\hat{x}_{t+1}^1$ $\hat{x}_{t+1}^2$ $\hat{x}_{t+1}^n$

f_1 f_2 $\bullet\ \bullet\ \bullet$ f_n

x_t a_t x_t a_t x_t a_t

next state x_{t+1}

action a_t

policy network
$\pi_\theta(x_t)$

current state x_t

Curiosity
Reward

$r_t^i = \sigma\left\{ \right.$

Disagreement

next state x_{t+1}

Disagreement

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

x_{t+1}

$\| x_{t+1} - \hat{x}_{t+1}^n \|$

$r_t^i = \sigma\{ \hat{x}_{t+1}^1 \quad \hat{x}_{t+1}^2 \quad \hat{x}_{t+1}^n \}$

policy network
$\pi_\theta(x_t)$

Curiosity
Reward

f_1 f_2 $\bullet \bullet \bullet$ f_n

x_t a_t x_t a_t x_t a_t

current state x_t

next state x_{t+1}

Disagreement

x_{t+1}

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

$x_{t+1} - \hat{x}_{t+1}^n\|$

policy network

$r_t^i = \sigma\{$ $\hat{x}_{t+1}^1$ $\hat{x}_{t+1}^2$ $\hat{x}_{t+1}^n$ $\}$

Curiosity

No dependency on the environment!

J_1 J_2 f_n

x_t a_t x_t a_t x_t a_t

current state x_t

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

Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration

Model Optimization

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

Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration

Model Optimization

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

Policy Optimization

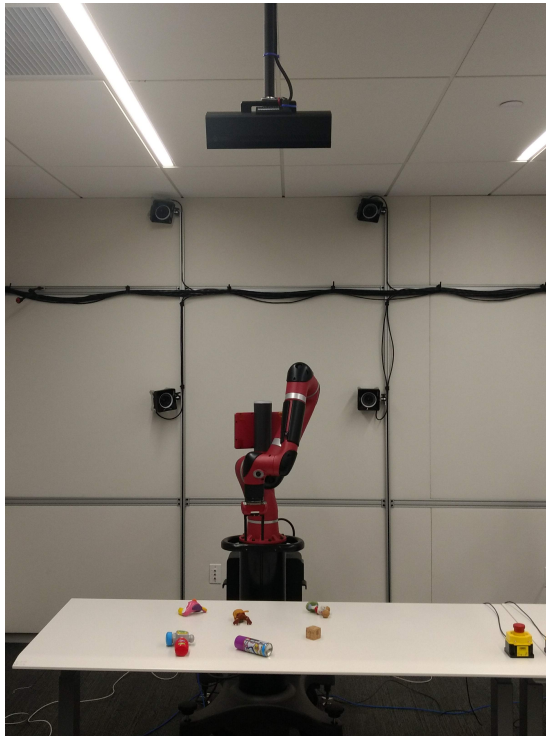
$$\max_{\theta_P} \sum_{i=1}^{k} \left\| f_{\theta_i}\left(x_t, \pi(x_t;\theta_P)\right) - \left(\frac{1}{k}\right) \sum_{j=1}^{k} f_{\theta_j}\left(x_t, \pi(x_t;\theta_P)\right) \right\|_2$$

Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration

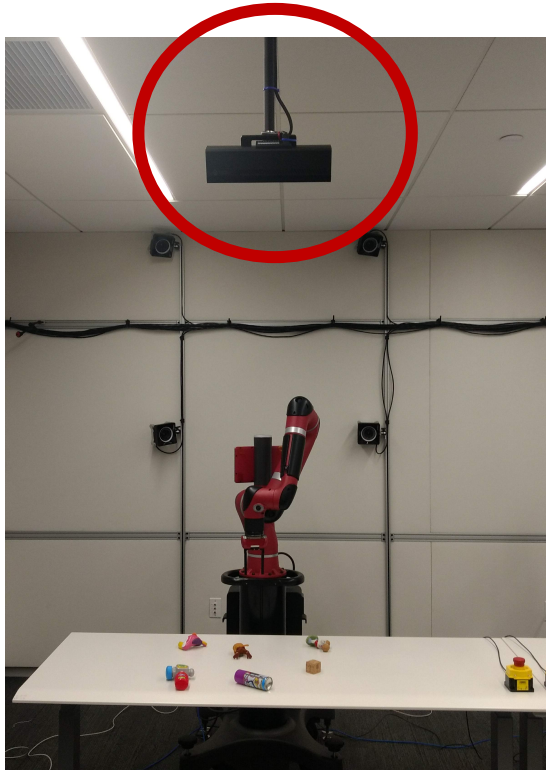
Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration



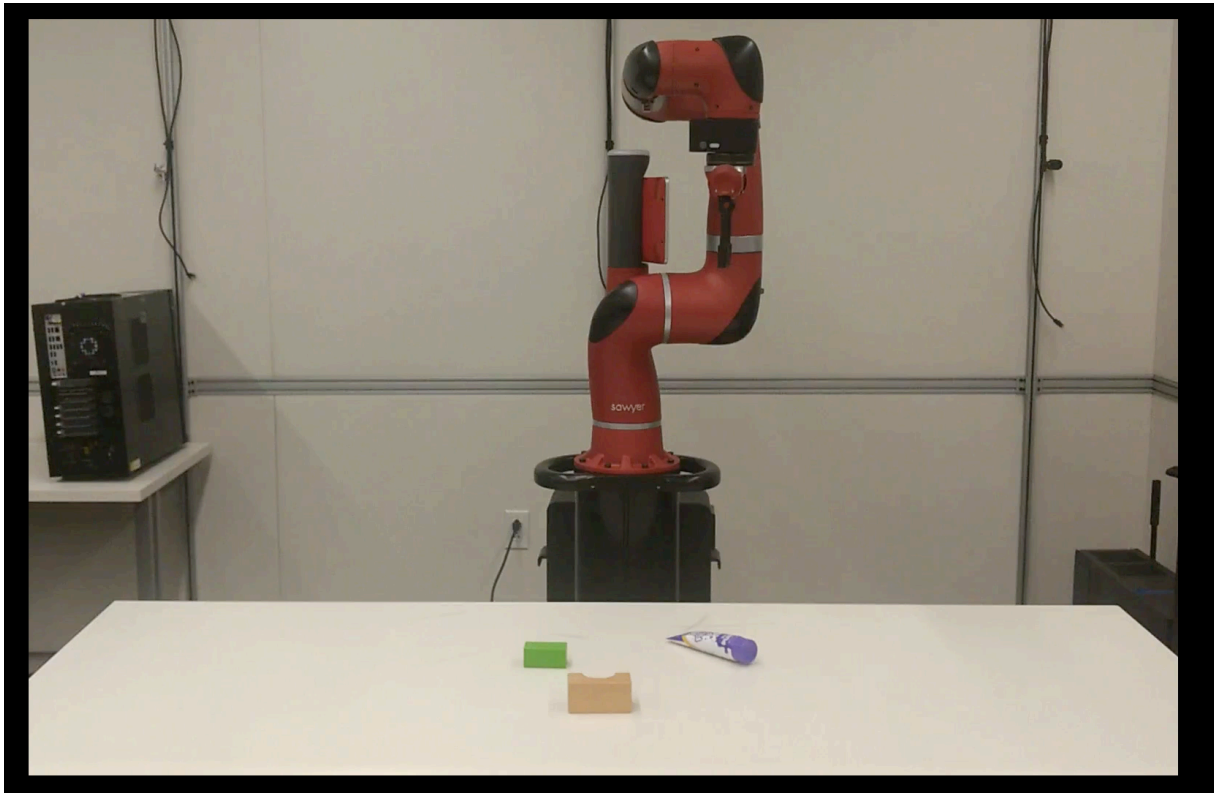
Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration



Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration

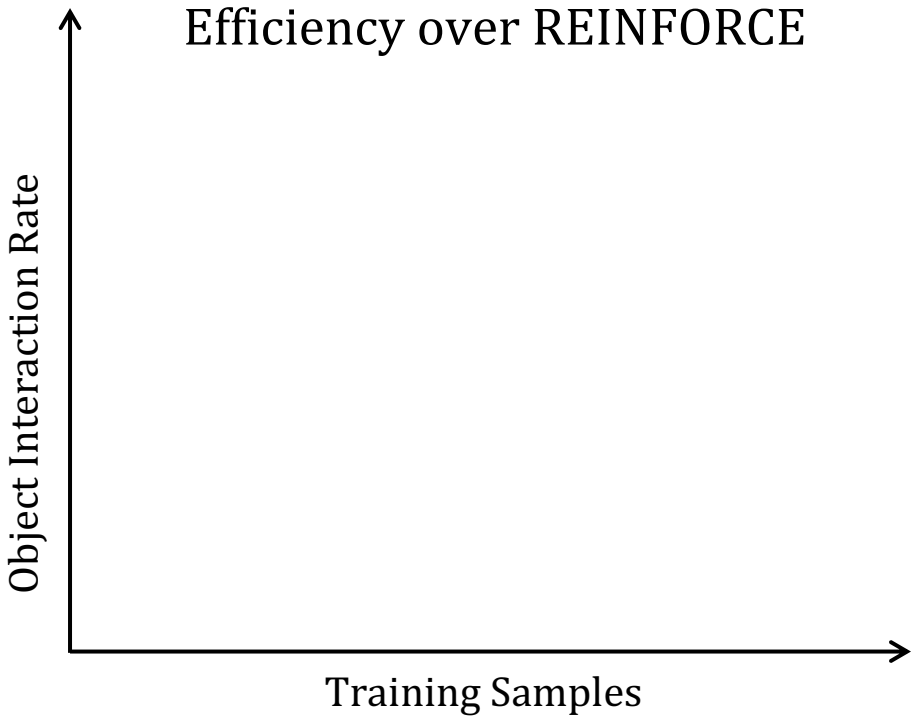


Position Control:

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

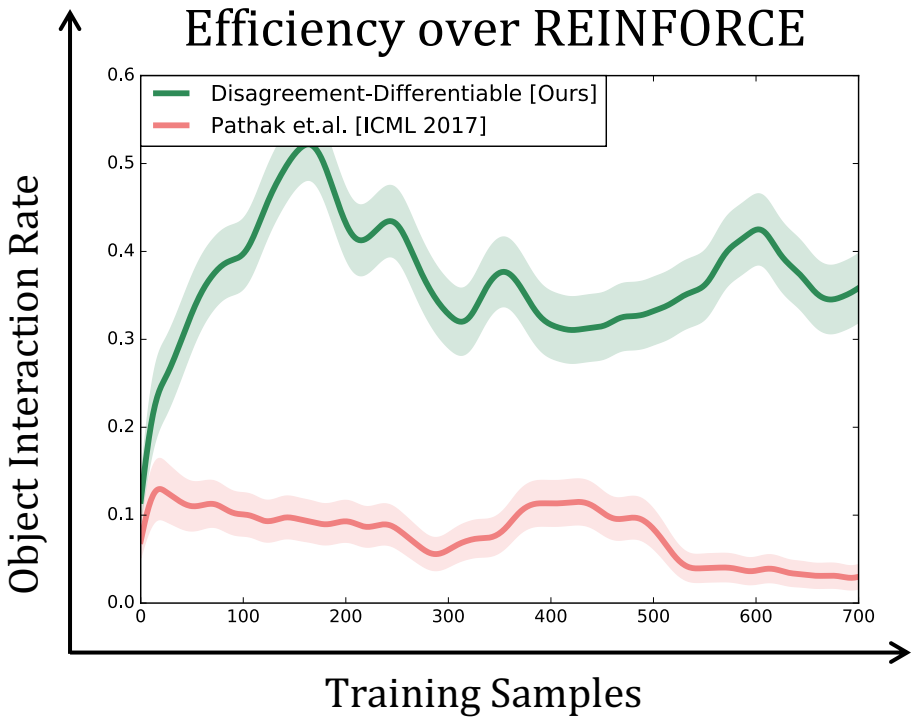
Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration



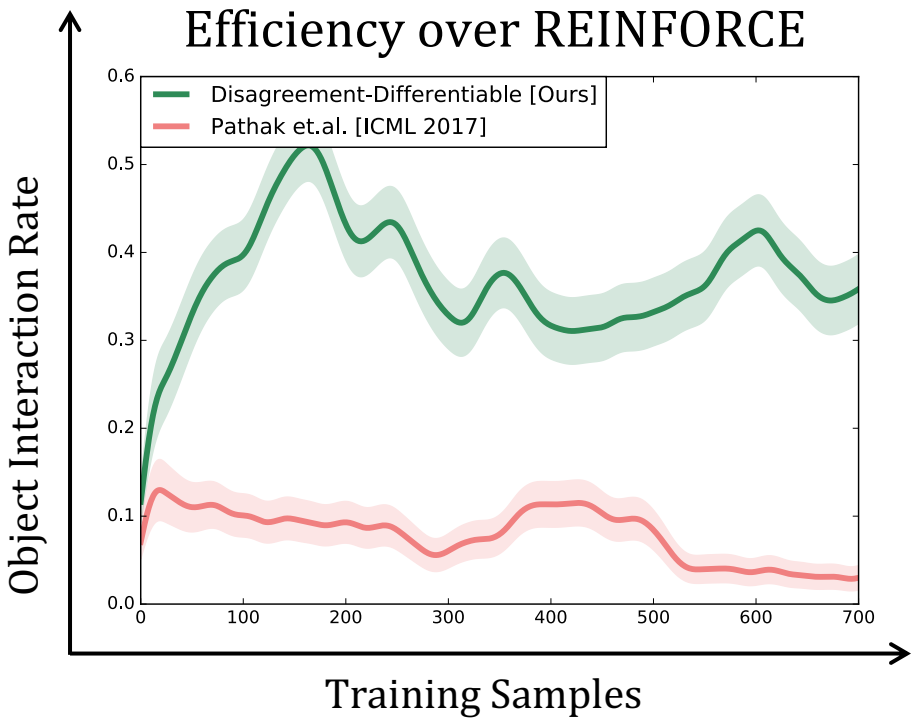
Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration

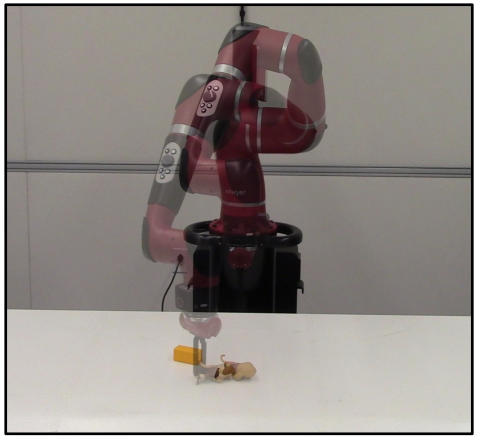


Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

Differentiable Exploration

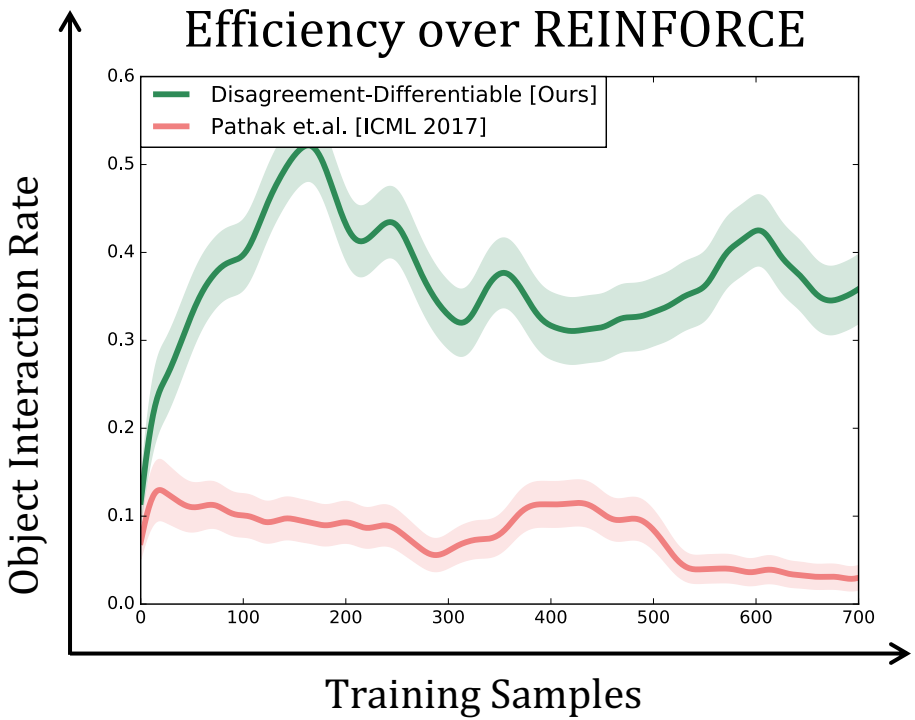


Pushing skill

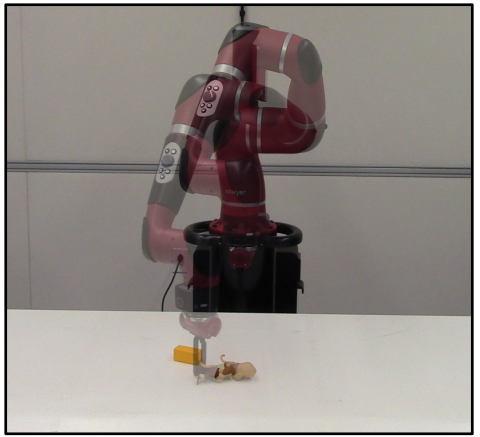


Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

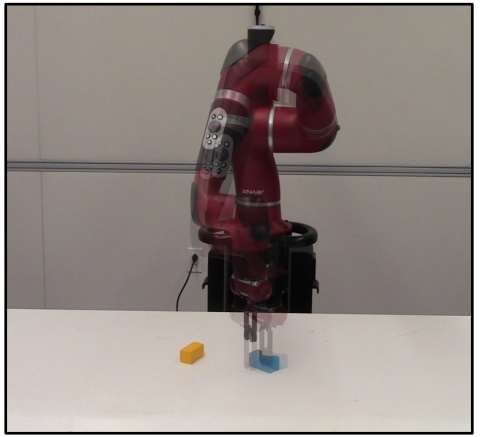
Differentiable Exploration



Pushing skill



Picking skill



Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.

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

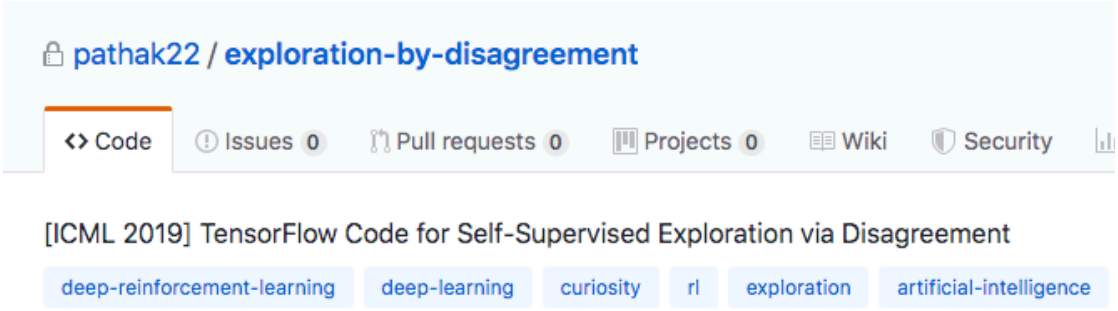
  (Stochastic Atari; Unity-TV)

- Differentiable reformulation for real robots

  (Sawyer Robot)

Code Available

https://pathak22.github.io/exploration-by-disagreement/



Poster # 39
(today)

Thank you!

Pathak*, Gandhi*, Gupta. “Self-Supervised Exploration via Disagreement“. ICML, 2019.