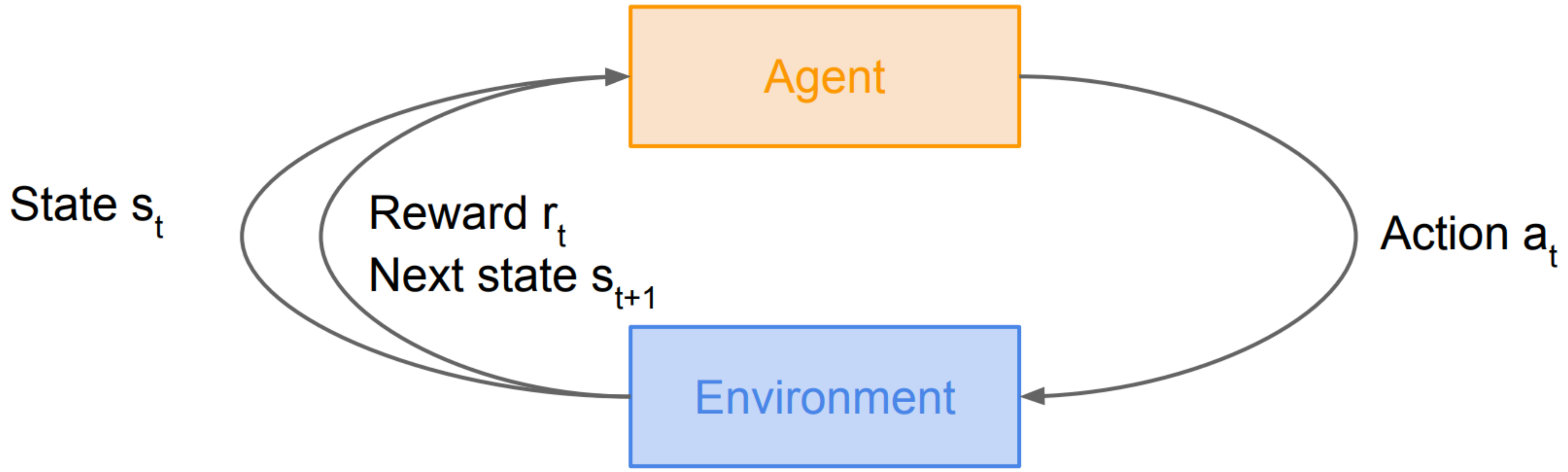
Curiosity-driven Exploration by Self-supervised Prediction

Author: Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, Trevor Darrell

ICML 2017

PRESENTER: CHIA-CHEN HSU

Reinforcement Learning



Credit: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture14.pdf

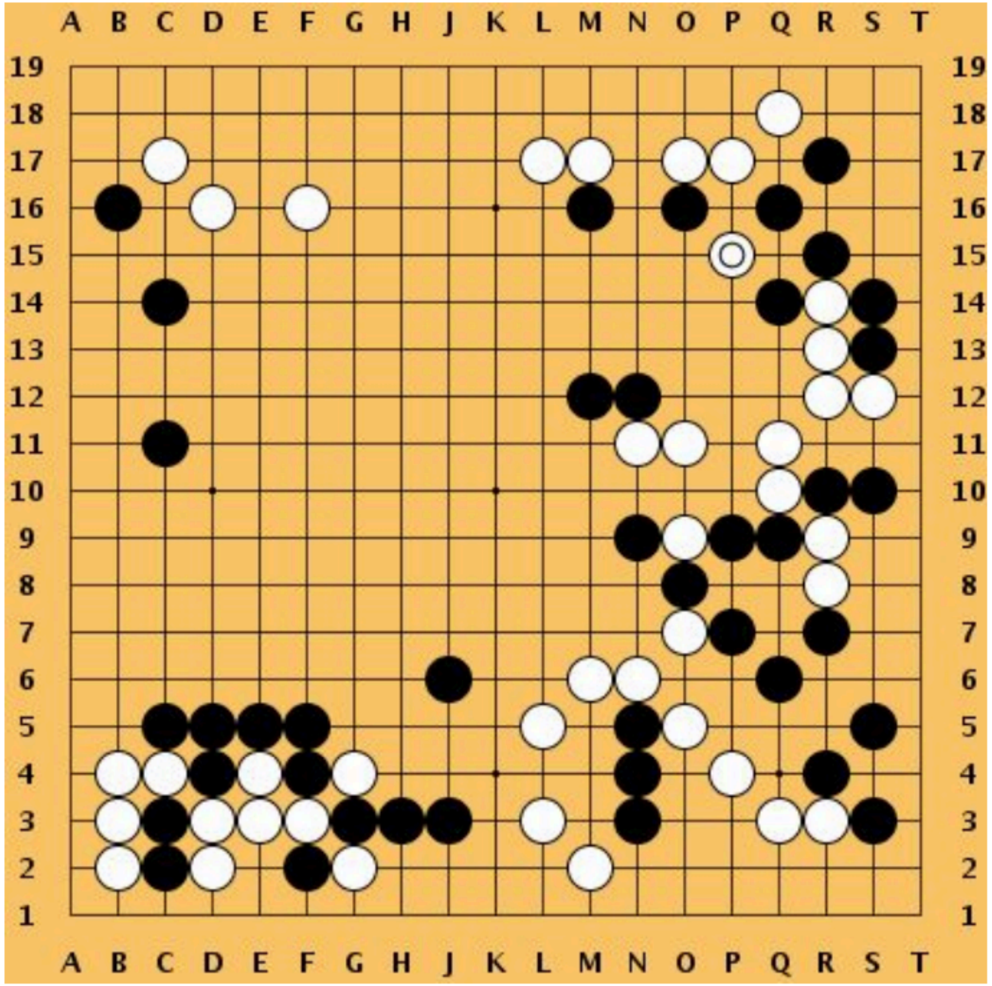
Example – Alpha Go

Objective: Win the game!

State: Position of all pieces

Action: Where to put the next piece down

Reward: 1 if win at the end of the game, 0 otherwise



Credit: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture14.pdf

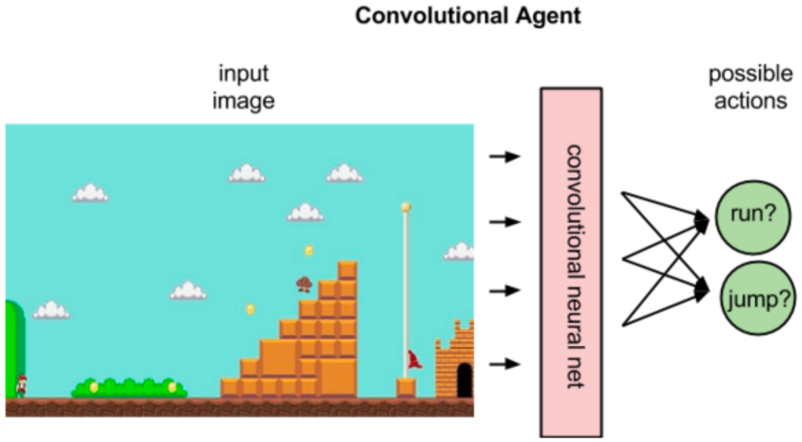
Example -- Games

Objective: Complete the game with the highest score

State: Raw pixel inputs of the game state

Action: Game controls e.g. Left, Right, Up, Down

Reward: Score increase/decrease at each time step



Credit: http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture14.pdf

Reward--Motivation

“Forces” that energize an organism to act and that direct its activity.

Extrinsic Motivation: being moved to do something because of some external reward ($$, a prize, etc.).

Intrinsic Motivation: being moved to do something because it is inherently enjoyable.

- Curiosity, Exploration, Manipulation, Play, Learning itself . . .
- Encourage the agent to explore “novel” states
- Encourage the agent to perform actions that reduce the error/uncertainty in the agent’s ability to predict the consequence of its own actions

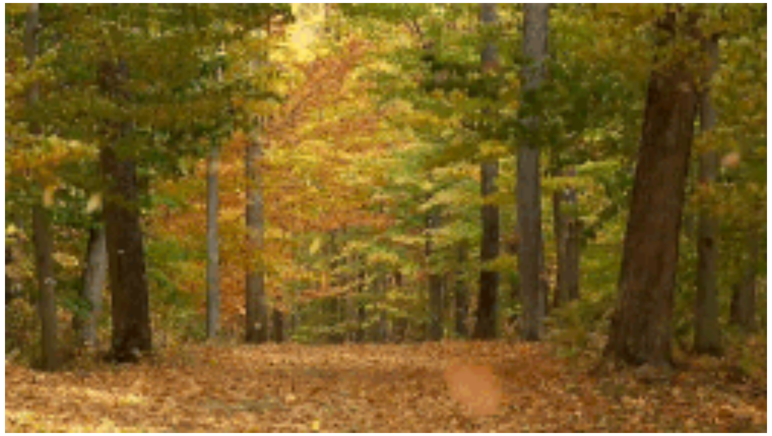
Challenge of Intrinsic Motivated

Imagine: movement of tree leaves in a breeze

- Pixel prediction would be high

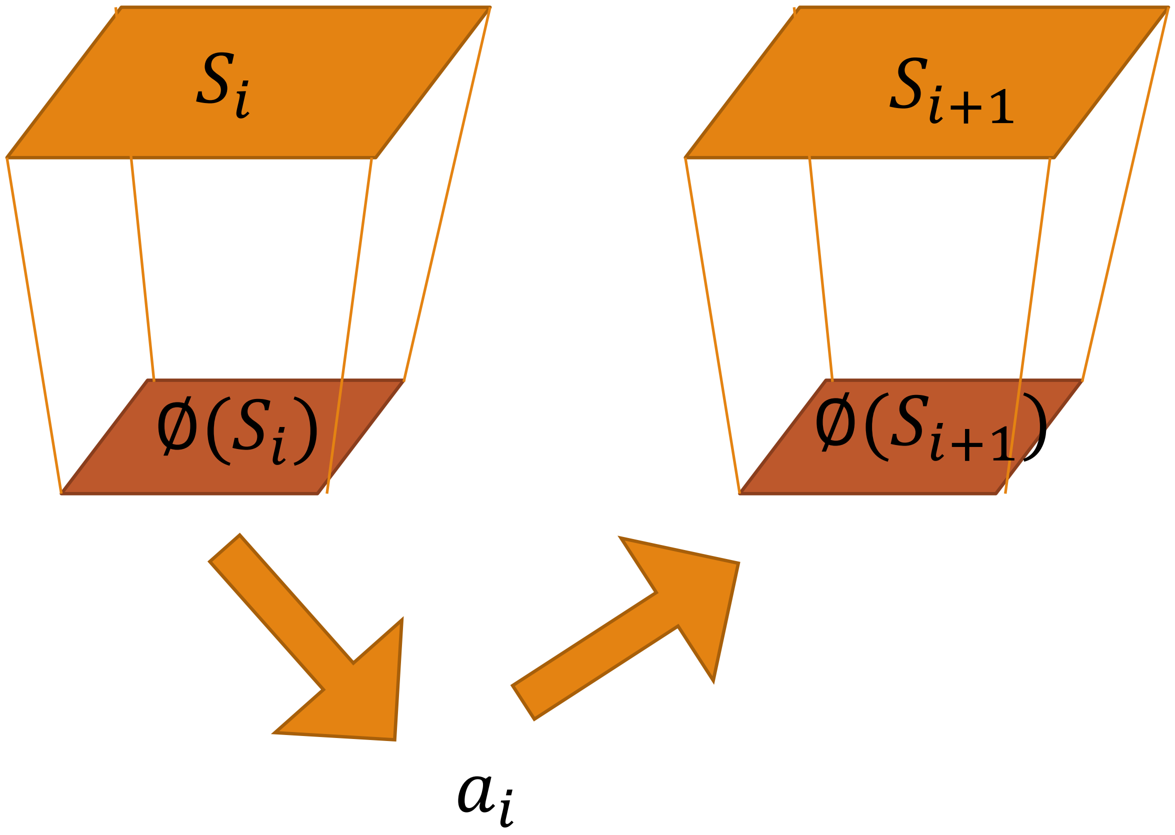
Observation

- (1) things that can be controlled by the agent;
- (2) things that the agent cannot control but that can affect the agent (e.g. a vehicle driven by another agent),
- (3) things out of the agent's control and not affecting the agent (e.g. moving leaves).



Goal : predict what change of states are caused by agent or will affect the agent

Self-supervised prediction



Inverse

$$g(\emptyset(S_i)\,,\emptyset(S_{i+1})) \rightarrow \widehat{a}_i$$

$$\min_{\theta_I} L_I(\hat{a}_t, a_t) \qquad (3)$$

Forward

$$\mathrm{f}(\emptyset(S_i), a_i) \rightarrow \widehat{\emptyset(S_i)}$$

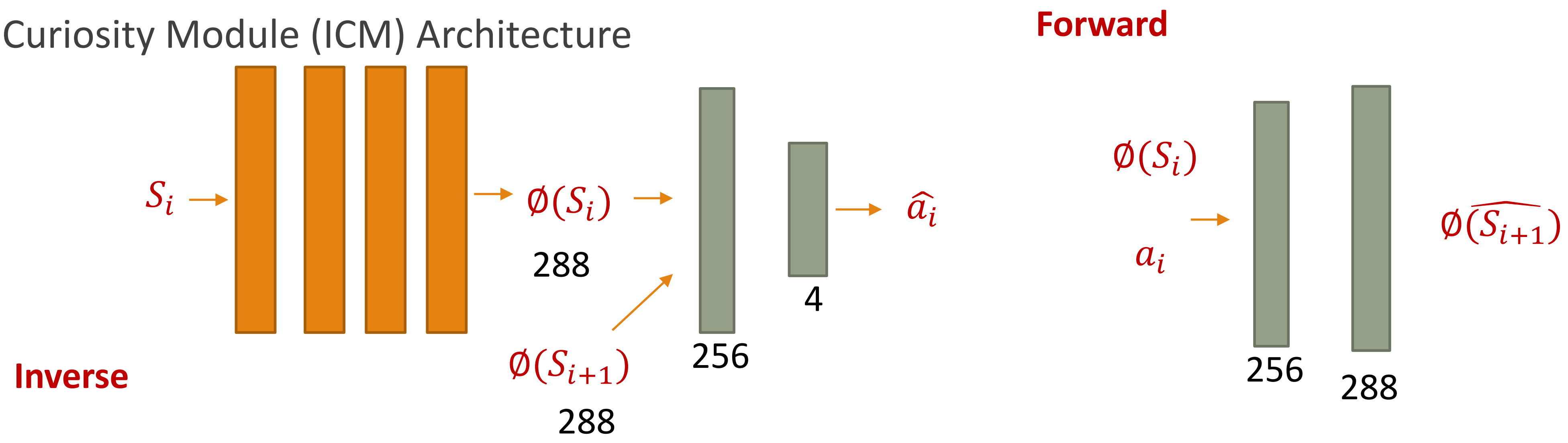
$$L_F\Big(\phi(s_t), \hat{\phi}(s_{t+1})\Big) = \frac{1}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$$

Reward

$$r_t^i = \frac{\eta}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$$

Architecture

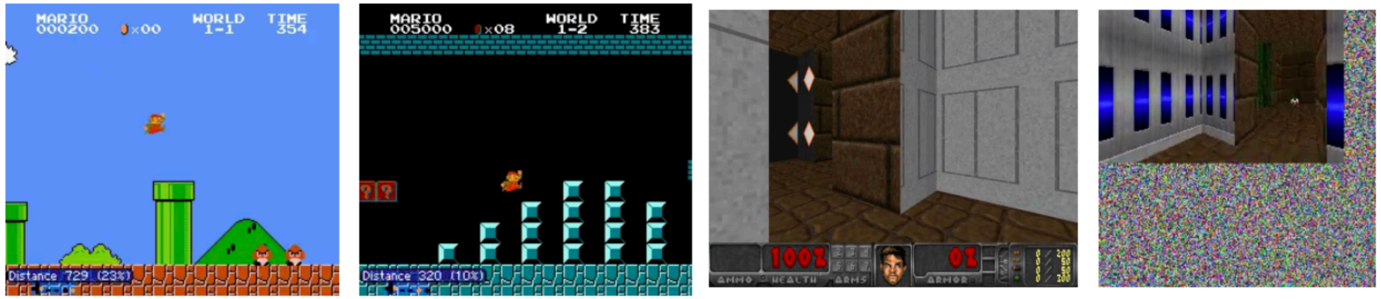
- A3C
  - Proposed by Google DeepMind. State-of-the-art RL architecture
  - 4 **convolution + LSTM** with 256 units + 2 fully connected
  - Two **separate fully connected layers** are used to predict
    - The value function
    - The action from the LSTM feature representation
- Intrinsic Curiosity Module (ICM) Architecture



Experiment

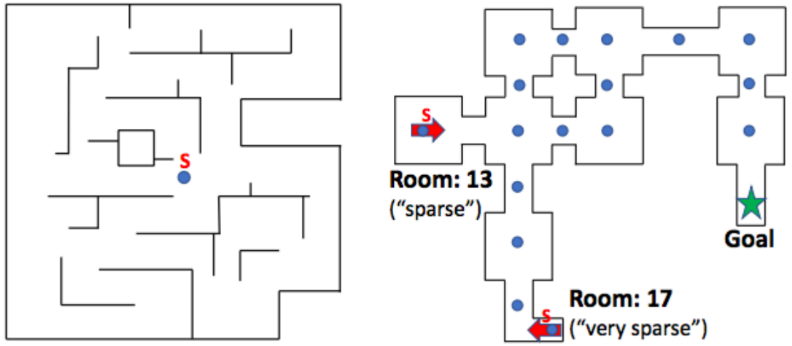
Environment

1. Super Mario Bros
2. VisDoom



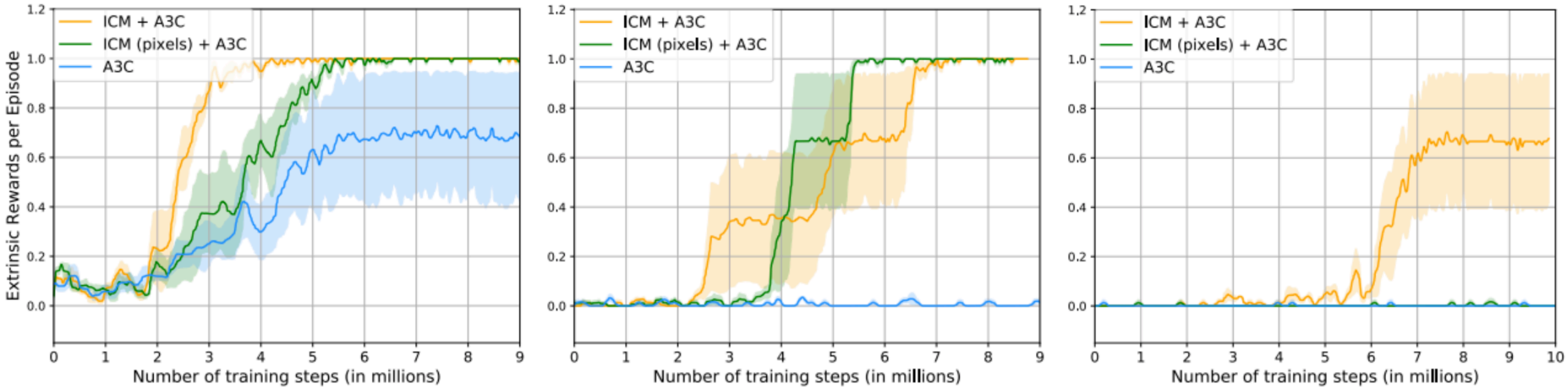
Setting

1. Sparse extrinsic reward on reaching a goal
2. Exploration without extrinsic reward



(a) Train Map Scenario (b) Test Map Scenario

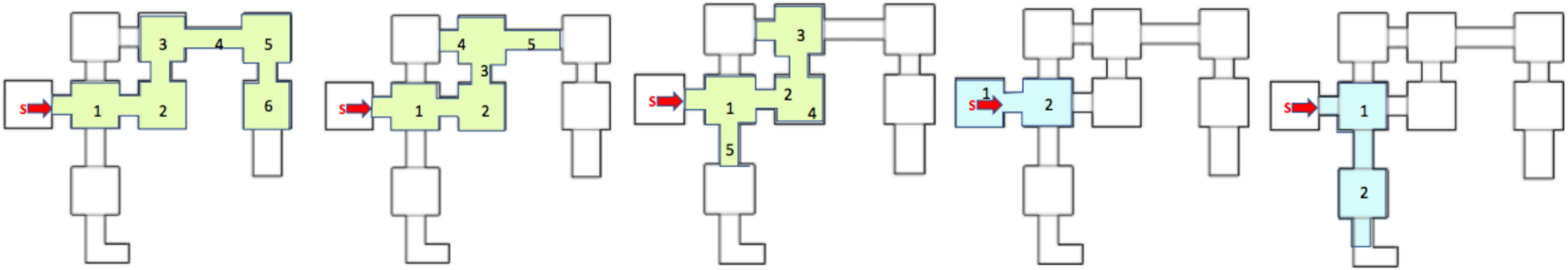
Sparse extrinsic reward on reaching a goal



(a) “dense reward” setting (b) “sparse reward” setting (c) “very sparse reward” setting

Exploration

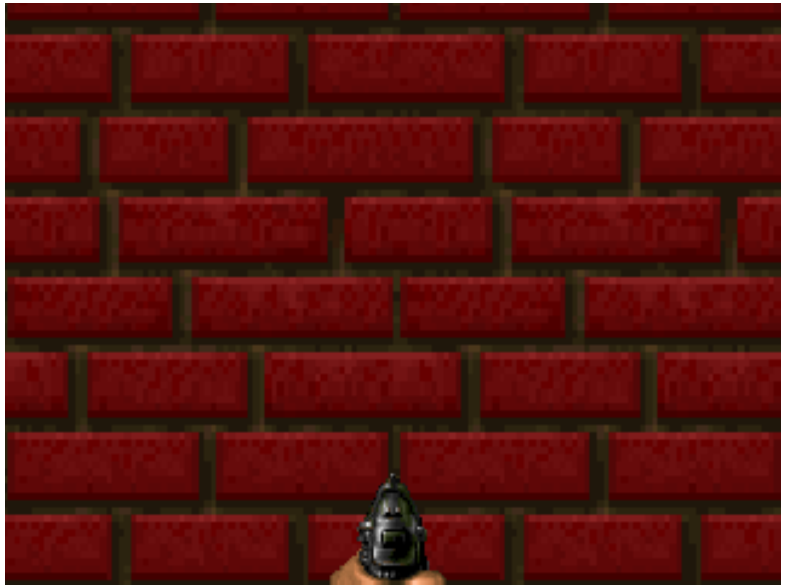
VisDoom



Mario

30% of level 1

Demo



NIPS2016[1]

ICML 2017
(This paper)

ICLR2017[2]
Winner, Visual Doom AI Competition2016

Backup

Self-supervised prediction--Reward

Two subsystems

- A reward generator that outputs a curiosity-driven intrinsic reward signal
  - Rewards $\mathbf{r_t = r^i_t + r^e_t}$

- A policy that outputs a sequence of actions to maximize that reward signal. In addition to intrinsic

$$\max_{\theta_P} \mathbb{E}_{\pi(s_t;\theta_P)}[\Sigma_t r_t]$$

Intrinsic Curiosity Module (ICM) Architecture

The inverse model

- first maps the input state (st) into a feature vector ϕ(st) using a series of four convolution layers, each with 32 filters, kernel size 3x3, stride of 2 and padding of 1. ELU non-linearity
- The dimensionality of ϕ(st) is 288.
- For the inverse model, ϕ(st) and ϕ(st+1) are concatenated into a single feature vector and passed as inputs into a fully connected layer of 256
- Fully connected layer with 4 units to predict one of the four possible actions.

The forward model

- Concatenating ϕ(st) with at and passing it into a sequence of two fully connected layers with 256 and 288 units respectively.

Self-supervised prediction

Forward

$$\hat{a}_t = g\Big(s_t, s_{t+1}; \theta_I\Big) \qquad (2)$$

$$\min_{\theta_I} L_I(\hat{a}_t, a_t) \qquad (3)$$

Inverse

$$\hat{\phi}(s_{t+1}) = f\Big(\phi(s_t), a_t; \theta_F\Big)$$

$$L_F\Big(\phi(s_t), \hat{\phi}(s_{t+1})\Big) = \frac{1}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2 \qquad (5)$$

Reward

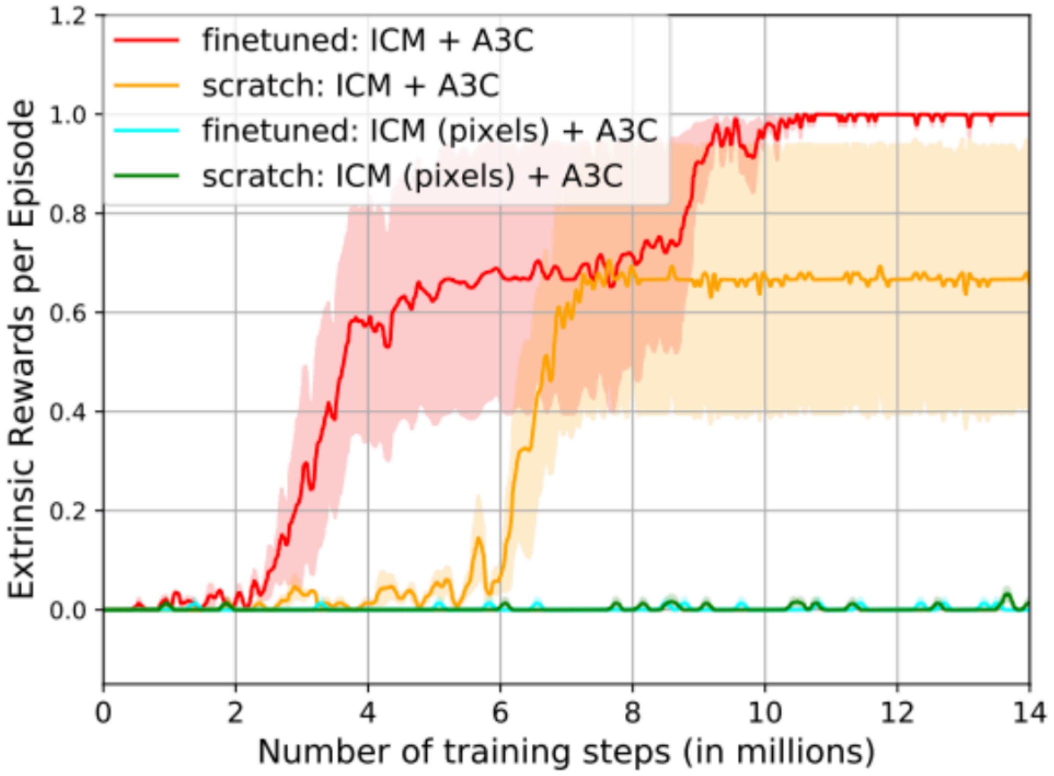
$$r_t^i = \frac{\eta}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$$

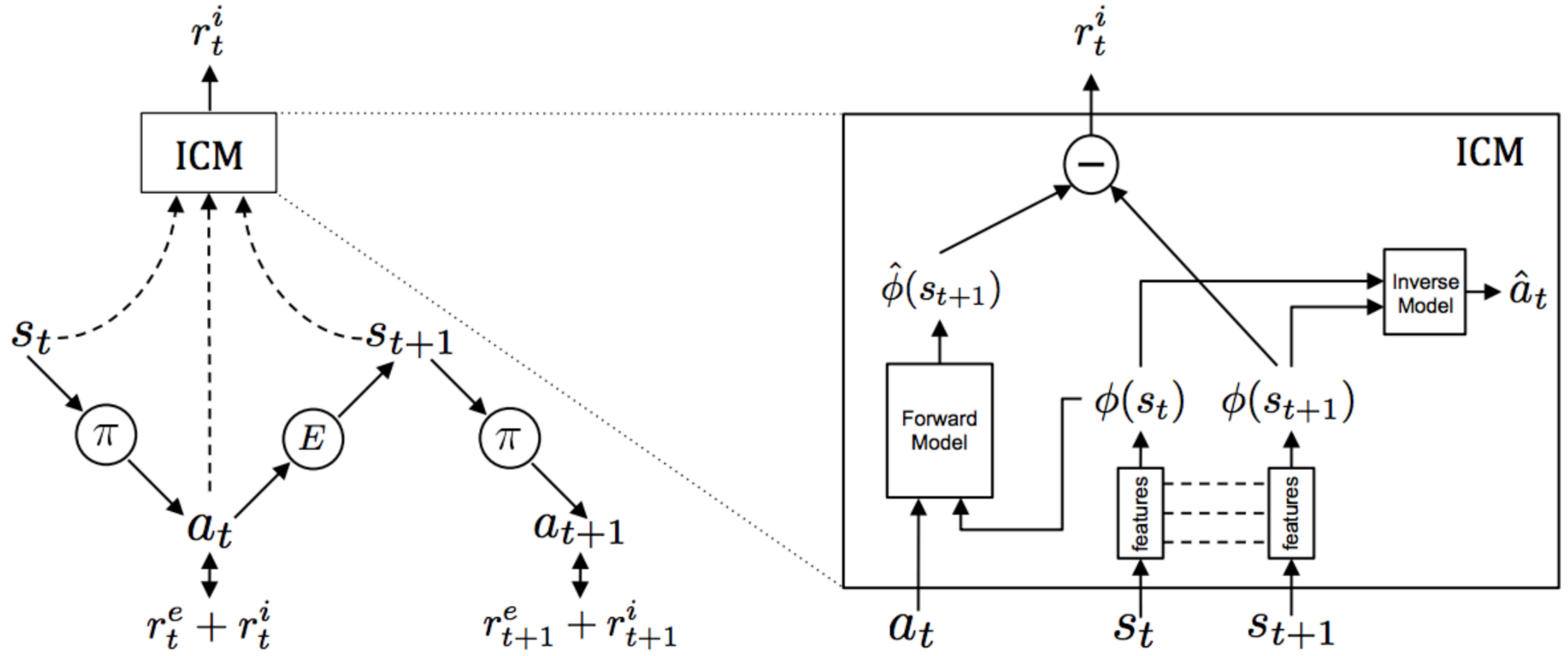
Intrinsic Reward in RL

1. Explore “Novel” state

2. Reduce error/uncertainty

Fine tuned with curiosity vs external





r_t^i
ICM
s_t
π
a_t
E
s_{t+1}
π
a_{t+1}
$r_t^e + r_t^i$
$r_{t+1}^e + r_{t+1}^i$
r_t^i
ICM
$\hat{\phi}(s_{t+1})$
Forward Model
$\phi(s_t)$
$\phi(s_{t+1})$
Inverse Model
$\hat{a}_t$
features
features
a_t
s_t
s_{t+1}

http://realai.org/intrinsic-motivation/

http://swarma.blog.caixin.com/archives/164137

https://data-sci.info/2017/05/16/%E4%B8%8D%E9%9C%80%E8%A6%81%E5%A4%96%E9%83%A8reward%E7%9A%84%E5%A2%9E%E5%BC%B7%E5%BC%8F%E5%AD%B8%E7%BF%92-curiosity-driven-exploration-self-supervised-prediction/

https://weiwenku.net/d/100573787 **