

CURIOSITY - DRIVEN EXPLORATION BY SELF-SUPERVISED PREDICTION (ICM)

Kim. J. H

1, November, 2021

What is Curiosity ?

Curiosity = Intrinsic Motivation = Intrinsic reward signal



as **Human agents**,
we are accustomed to operating with **rewards that are so sparse** that we only experience them once or twice in a lifetime, if at all.

- Motivation/curiosity have been used to explain the **need to explore the environment and discover novel states.**
- Curiosity is **a way of learning new skills** which **might come handy for pursuing rewards in the future.**

Why Curiosity ?

In many **Real-world scenarios**,
rewards **extrinsic** to the agent are **extremely sparse**, or **absent altogether**.

➡ Human agents can solve tasks **with curiosity** under those situations.

➡ How about other agents trained with Reinforcement Learning ?

➡ If **with Curiosity**, It might be alright !



Experiments in sparse reward Game environments
e.g. *VizDoom*, *Super Mario Bros*

Related Works so far

Intrinsic motivation/rewards formulations (two classes)

- Encourage the agent to **explore “novel” states** through **recording the counts of visited states**.
(Bellemare et al., 2016; Lopes et al., 2012; Poupart et al., 2006)
- Encourage the agent to **perform actions** that **reduce the error** in the agent’s ability to **predict the consequence of its own actions**.
(Houthoofd et al., 2016; Mohamed & Rezende, 2015; ...)





Both methods fail

- Hard to build **in high-dimensional continuous state** spaces
(e.g. images)
- Hard to deal with the **stochasticity of the agent-environment** system
(i.e. **noise** in the agent’s actuation, **inherent stochasticity** in the environment)
- **Challenge of generalization** across physically distinct but functionally similar parts of an environment

Curiosity – driven Exploration

Key Insight

Only reward the agent when it encounters states that are **hard to predict** but are **“learnable”**

1. States that are **hard to predict**  **Prediction error** as **intrinsic reward**
2. States that are **learnable**  **Transform state to **feature space** that related to the action performed by agent**

Inverse dynamics

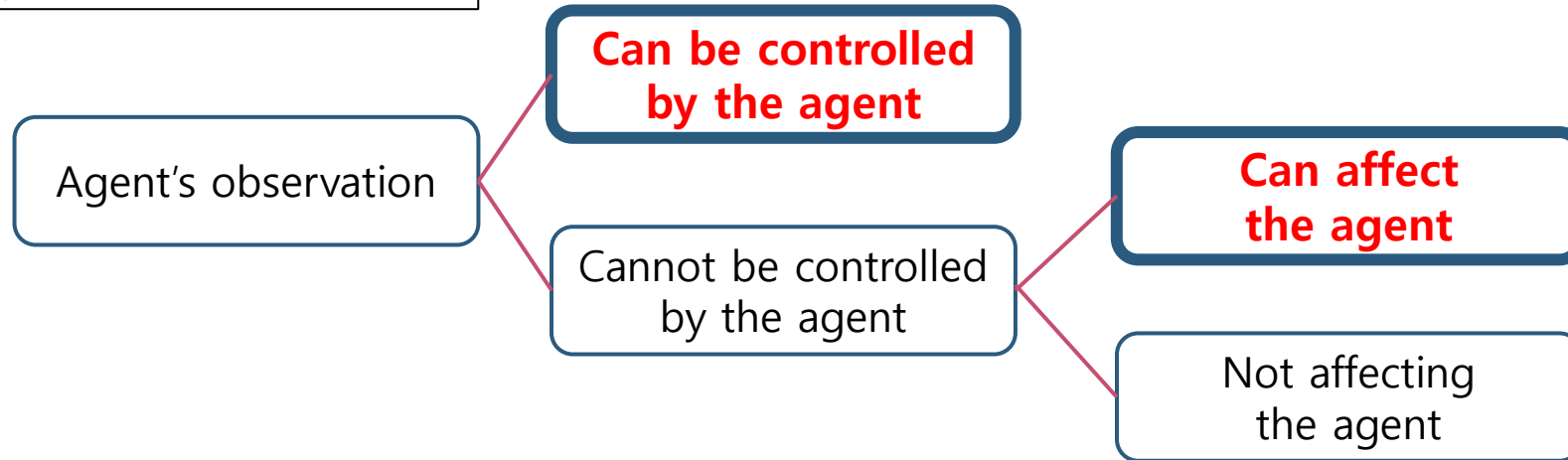
Feature extraction (only features related to the agent)

Forward dynamics

Predicts the feature representation of the next state

Curiosity – driven Exploration

Inverse dynamics model



Good feature space for curiosity



Can be controlled
by the agent

+

Can affect
the agent

Encoder
(feature extractor)

$$s_t \rightarrow \varphi(s_t)$$

Predicts the action

$$\varphi(s_t), \varphi(s_{t+1}) \rightarrow \hat{a}_t$$



$$\hat{a}_t = g(s_t, s_{t+1}; \theta_I)$$
$$\min_{\theta_I} L_I(\hat{a}_t, a_t)$$

L_I is the discrepancy between \hat{a}_t and a_t

Curiosity – driven Exploration

Forward dynamics model

Takes as inputs a_t and $\varphi(s_t)$ and predicts $\varphi(s_{t+1})$

$$\begin{aligned}\hat{\varphi}(s_{t+1}) &= f(\varphi(s_t), a_t; \theta_F) \\ L_F(\varphi(s_t), \hat{\varphi}(s_{t+1})) &= \frac{1}{2} \|\hat{\varphi}(s_{t+1}) - \varphi(s_{t+1})\|_2^2\end{aligned}$$

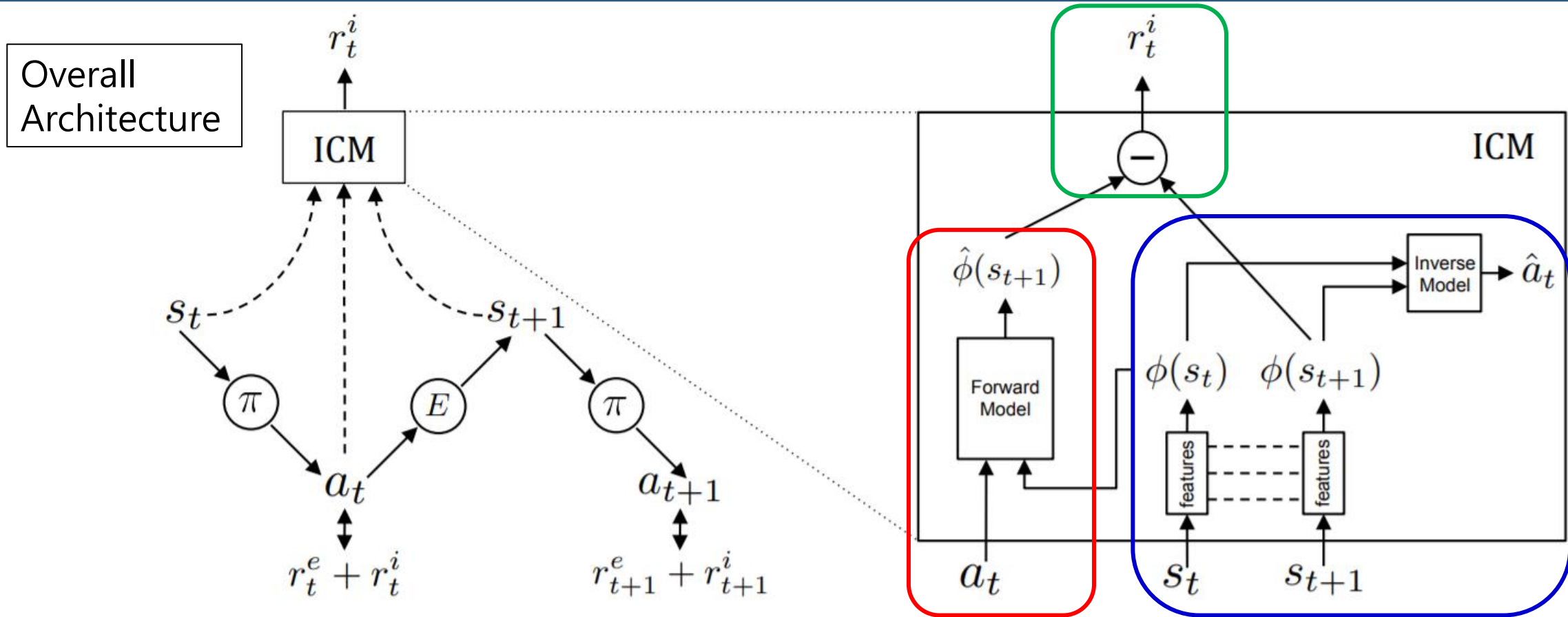


The intrinsic reward signal

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

(total reward : $r_t = r_t^i + r_t^e$)

Curiosity – driven Exploration



Forward Model : next state feature prediction $\hat{\phi}(s_{t+1}) = f(\phi(s_t), a_t; \theta_F)$

Inverse Model : feature extraction + self-supervised learning $\hat{a}_t = g(s_t, s_{t+1}; \theta_I)$

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

Total reward : $r_t = r_t^i + r_t^e$

Curiosity – driven Exploration

overall
Optimization Problem

$$\min_{\theta_P, \theta_I, \theta_F} \left[\underbrace{-\lambda \mathbb{E}_{\pi(s_t; \theta_P)} [\Sigma_t r_t]}_{\text{Policy Gradient}} + \underbrace{(1 - \beta) L_I}_{\text{Inverse dynamics}} + \underbrace{\beta L_F}_{\text{Forward dynamics}} \right]$$

Policy Gradient
(maximize the expected return)

Inverse dynamics

Forward dynamics

$\lambda > 0$ (**weights the importance of the policy gradient loss** against the importance of learning the intrinsic reward signal)

$0 \leq \beta \leq 1$ (**weights the inverse model loss** against the forward model loss)

Experiment

Three roles of ICM

1. Solving tasks **with sparse rewards**
2. Helping agent **explore** its environment in the quest **for new knowledge**
3. **Generalization** to unseen scenarios

Experiment settings

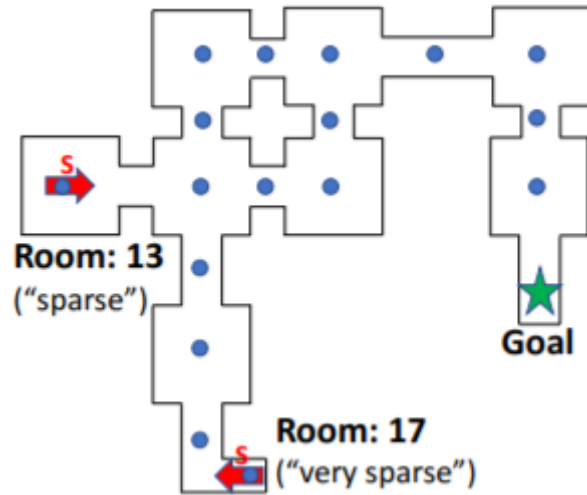
1. **Sparse** extrinsic reward
2. Exploration with **no extrinsic reward**
3. Generalization to **unseen scenarios**

Environment & Agent

in *VizDoom*, *Super Mario Bros*
A3C w/ ϵ -greedy exploration
A3C w/ ICM
A3C w/ ICM-pixels(w/o Inverse model)

Experiment - Sparse Extrinsic Reward

Sparse Extrinsic Reward Setting



VizDoom environment

Dense rewards : starting from **blue dots**

Sparse rewards : starting from **room 13**
(avg 270 steps away from the Goal under an optimal policy)

Very-Sparse rewards : starting from **room 17**
(avg 350 steps away from the Goal under an optimal policy)

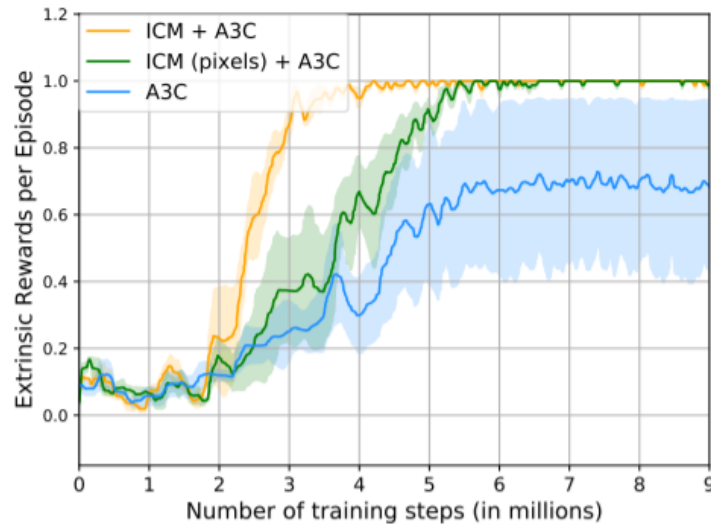
Extrinsic rewards are **only provided** when **reaching to Goal** position.

Termination conditions : **finding Goal** or **max 2100 steps**

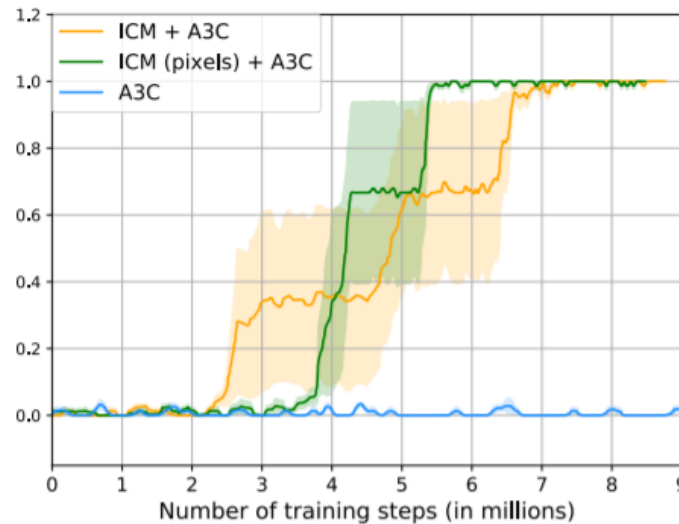
Experiment - Sparse Extrinsic Reward

Sparse Extrinsic Reward

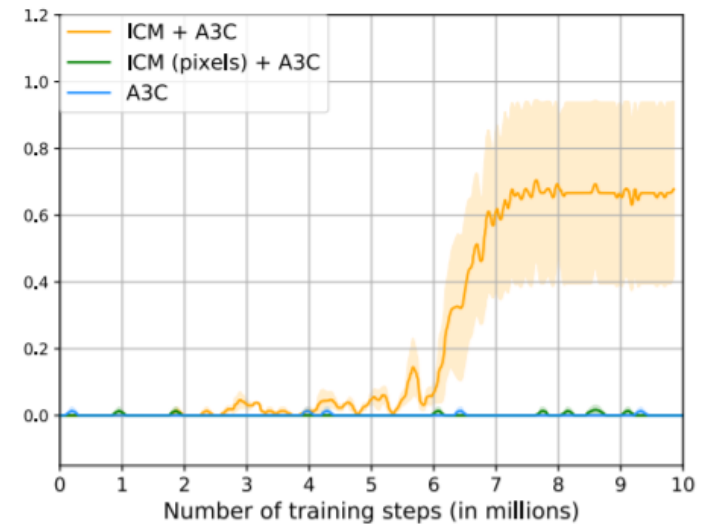
A3C : A3C w/ ϵ -greedy exploration



(a) “dense reward” setting



(b) “sparse reward” setting



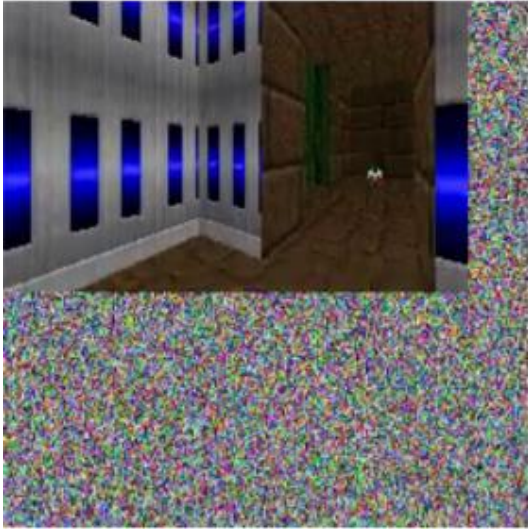
(c) “very sparse reward” setting

mean \pm std error, three independent runs of each algorithm, no tuning of random seeds

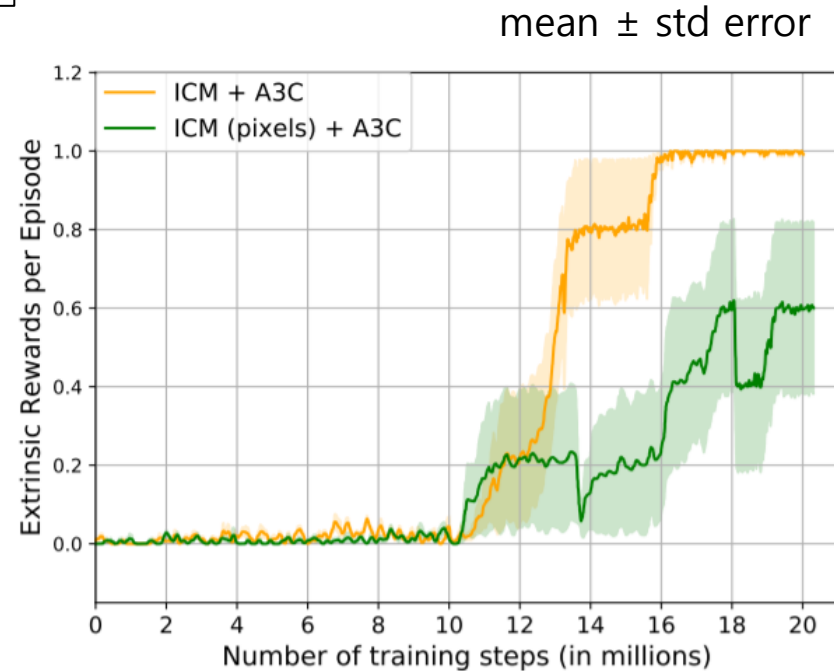
- A3C baseline(w/o curiosity) performance degrades with sparser rewards.
- A3C + ICM-pixel never succeed in very-sparse reward case.
- A3C + ICM agents are superior in all cases.

Experiment - Sparse Extrinsic Reward

Robustness to uncontrollable dynamics



40% of white noise
(not affecting the agent)



- Experiment on the sparse reward setup.
- A3C + ICM achieves a perfect score.
- A3C + ICM-pixels suffers significantly despite having succeeded at the sparse reward task w/o white noise.

Experiment - Sparse Extrinsic Reward

Comparison to SOTA (TRPO + VIME)

TRPO : general more sample efficient than A3C

VIME : Variational Information Maximizing Exploration
(state-of-the-art exploration method at that time, (Houthoof et al.,2016))

Method ("sparse" reward setup)	Mean (Median) Score (at convergence)
TRPO	26.0 % (0.0 %)
A3C	0.0 % (0.0 %)
VIME + TRPO	46.1 % (27.1 %)
ICM + A3C	100.0 % (100.0 %)

- Experiment on the sparse reward setup.
- A3C + ICM work significantly better than others

Experiment - No Extrinsic Reward

No Extrinsic Reward Setting

A good exploration policy :
allowing the agent to **visit as many states as possible** even w/o any goals.

VizDoom setting

- No Extrinsic Reward
- Terminates in 2100 steps
- The farthest rooms are over 250 steps away (for an optimally-moving agent)

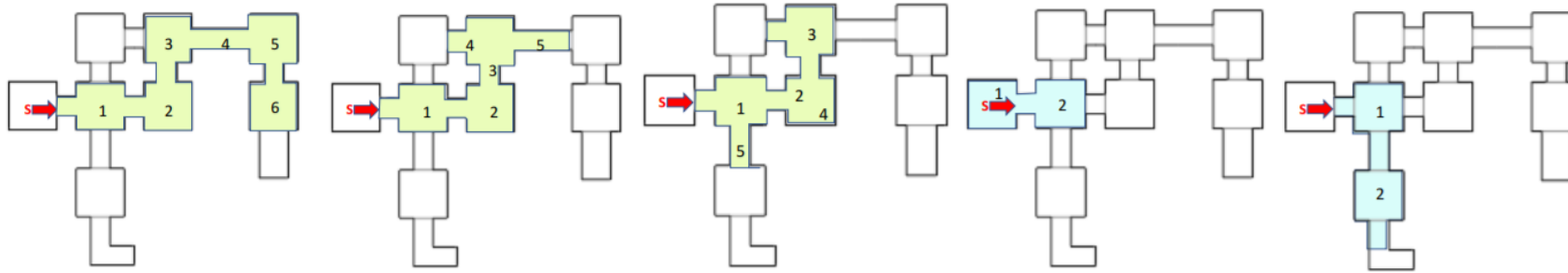
Super Mario Bros setting

- No Extrinsic Reward (for killing or dodging enemies or avoiding fatal events)

Experiment - No Extrinsic Reward

No Extrinsic Reward

VizDoom



ICM

Random Exploration

- **ICM** agents **explore a much larger state space** w/o any extrinsic reward
- **Random exploration** agents have hard time **getting around local minima**

Mario

- **ICM** agents **can learn to cross over 30% of Level-1**
- **Automatically discovered killing or dodging enemies or avoiding fatal events** behaviors for remaining curious to reach new states.

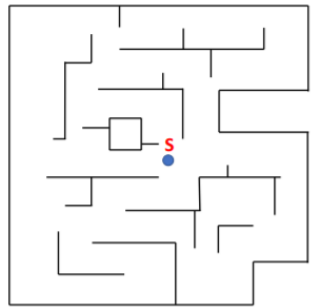
Experiment - Generalization

Generalization to Novel Scenarios

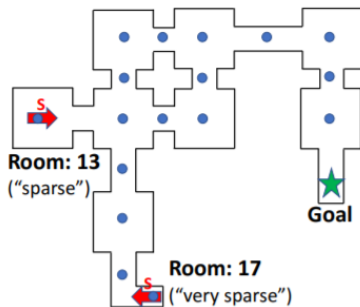
Generalization? or **Simply Memorizing** the training set?

- **No reward exploratory behavior**
- Evaluate the resulting exploration policy in three ways
 - a) apply the learned policy **"as is"** to a new scenario (Mario)
 - b) adapt the policy by **fine-tuning** with curiosity reward only (Mario)
 - c) adapt the policy to maximize **some extrinsic reward** (VizDoom)

VizDoom



(a) Train Map Scenario



(b) Test Map Scenario

**Pre-trained
vs.
From scratch**

on very sparse setting

Mario

Pre-trained by **ICM on Level-1**
Tested on **Level-2, 3 "as is"** and **"fine-tuned"**

measure : the distance covered by the agent

Experiment - Generalization

Generalization - Mario

Level Ids	Level-1	Level-2				Level-3			
Accuracy Iterations	Scratch 1.5M	Run as is 0	Fine-tuned 1.5M	Scratch 1.5M	Scratch 3.5M	Run as is 0	Fine-tuned 1.5M	Scratch 1.5M	Scratch 5.0M
Mean \pm stderr	711 \pm 59.3	31.9 \pm 4.2	466 \pm 37.9	399.7 \pm 22.5	455.5 \pm 33.4	319.3 \pm 9.7	97.5 \pm 17.4	11.8 \pm 3.3	42.2 \pm 6.4
% distance > 200	50.0 \pm 0.0	0	64.2 \pm 5.6	88.2 \pm 3.3	69.6 \pm 5.7	50.0 \pm 0.0	1.5 \pm 1.4	0	0
% distance > 400	35.0 \pm 4.1	0	63.6 \pm 6.6	33.2 \pm 7.1	51.9 \pm 5.7	8.4 \pm 2.8	0	0	0
% distance > 600	35.8 \pm 4.5	0	42.6 \pm 6.1	14.9 \pm 4.4	28.1 \pm 5.4	0	0	0	0

As is :

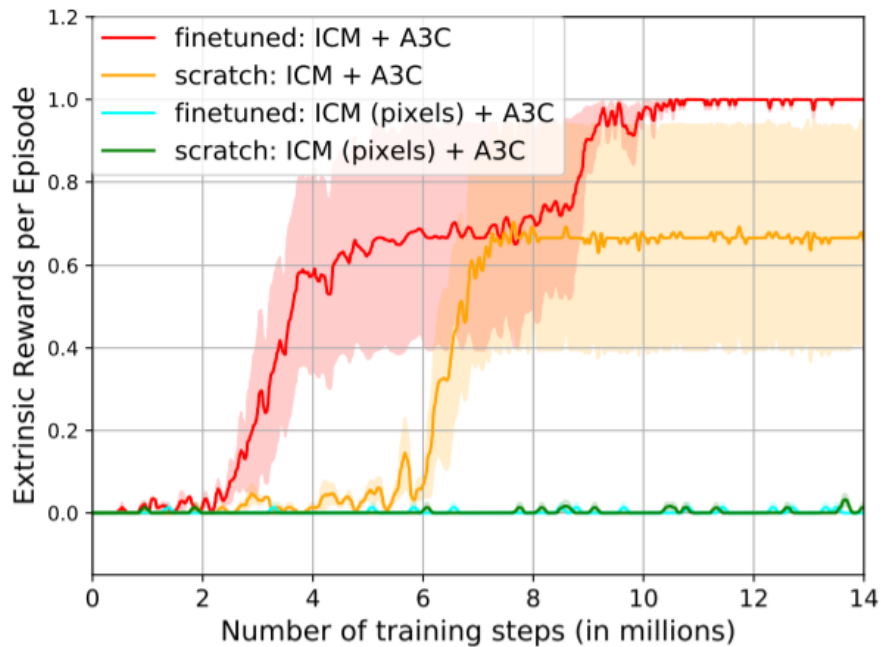
- on **Level-2 not well** (significantly different visual appearance, night world)
- on **Level-3 surprisingly well** (similar visual appearance, day world)

Fine-tuning :

- on **Level-2 better than from scratch** (overcoming different visual appearance)
even **better than more iterations of scratch**,
(the agent is able to use the knowledge acquired by playing Level-1 to better explore the subsequent levels)
- on **Level-3 performance deteriorated** (too hard to overcome only with curiosity, boredom)

Experiment - Generalization

Generalization - VizDoom



Pre-trained and fine-tuned A3C + ICM agent learns **faster** and achieves **higher** reward than **ICM agent trained from scratch** even when external rewards are provided (**red**, **yellow**)

ICM-pixels does not generalize

Summary

1. Proposing **ICM** (Intrinsic-Curiosity Module)
 - **scales to high dimensional** visual inputs (bypassing the difficult of predicting pixels)
 - ensures the **exploration strategy** of the agent is **unaffected by nuisance factors in the environment**
2. Find **good exploration policy** on sparse, very-sparse, and even no reward setting
3. Evaluating **Generalization** to unseen scenarios(separate testing set)
 - by applying the learned policy to a new scenario **"as is"**
 - by **fine-tuning** the learned policy on a new scenario

QnA

Reference

<https://arxiv.org/abs/1705.05363>

D. Pathak et al. “Curiosity-driven Exploration by Self-supervised Prediction.”

<https://github.com/utilForever/rl-paper-study/blob/main/1st/200608%20-%20Curiosity-driven%20Exploration%20by%20Self-supervised%20Prediction%2C%20Pathak%20et%20al%2C%202017.pdf>

최하늘(Haneul Choi), Curiosity-driven Exploration 리뷰.

<https://bluediary8.tistory.com/30>

미스터탁, Curiosity-driven Exploration 리뷰.