

---

# Supplementary Material

## Curiosity-driven Exploration by Self-supervised Prediction

---

Deepak Pathak<sup>1</sup> Pulkit Agrawal<sup>1</sup> Alexei A. Efros<sup>1</sup> Trevor Darrell<sup>1</sup>

### 1. Experimental Setup

To evaluate our curiosity module on its ability to improve exploration and provide generalization to novel scenarios, we will use two simulated environments. This section describes the details of the experimental setup. The code, environment setups, trained models and video of results are publicly available at <http://pathak22.github.io/noreward-rl/>.

**Training details** All agents in this work are trained using *visual inputs* that are pre-processed in manner similar to (Mnih et al., 2016). The input RGB images are converted into gray-scale and re-sized to  $42 \times 42$ . In order to model temporal dependencies, the state representation ( $s_t$ ) of the environment is constructed by concatenating the current frame with the three previous frames. Closely following (Mnih et al., 2015; 2016), we use action repeat of four during training time in VizDoom and action repeat of six in Mario. However, we sample the policy without any action repeat during inference. Following the asynchronous training protocol in A3C, all the agents were trained asynchronously with twenty workers using adaptive stochastic gradient descent, ADAM (Kingma & Ba, 2015).

**A3C architecture** The input state  $s_t$  is passed through a sequence of four convolution layers with 32 filters each, kernel size of  $3 \times 3$ , stride of 2 and padding of 1. An exponential linear unit (ELU; (Clevert et al., 2015)) is used after each convolution layer. The output of the last convolution layer is fed into a LSTM with 256 units. Two separate fully connected layers are used to predict the value function and the action from the LSTM feature representation.

**Intrinsic Curiosity Module (ICM) architecture** The intrinsic curiosity module consists of the forward and the inverse model. The encoder model first maps the input state ( $s_t$ ) into a feature vector  $\phi(s_t)$  using a series of four con-

---

<sup>1</sup>University of California, Berkeley. Correspondence to: Deepak Pathak <pathak@berkeley.edu>.

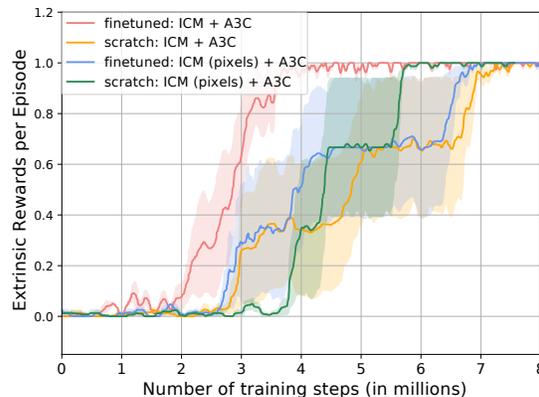


Figure 1. Curiosity pre-trained ICM + A3C when finetuned on the test map with environmental rewards outperforms ICM + A3C trained from scratch using both environmental and curiosity reward on the “sparse” reward setting of *VizDoom*. The pixel prediction based ICM agent does not generalize, while our ICM agent generalizes very well to the new map and novel textures. Results in “very sparse” scenario are in shown in main paper in Figure 8. Discussion is in Section 4.3.

volution layers, each with 32 filters, kernel size  $3 \times 3$ , stride of 2 and padding of 1. ELU non-linearity is used after each convolution layer. The dimensionality of  $\phi(s_t)$  (i.e. the output of the fourth convolution layer) is 288. For the inverse model,  $\phi(s_t)$  and  $\phi(s_{t+1})$  are concatenated into a single feature vector and passed as inputs into a fully connected layer of 256 units followed by an output fully connected layer with 4 units to predict one of the four possible actions. The forward model is constructed by concatenating  $\phi(s_t)$  with  $a_t$  and passing it into a sequence of two fully connected layers with 256 and 288 units respectively. Note that there are no shared parameters between ICM architecture and policy architecture because these two are optimizing the objective in approximately opposite directions.

The value of  $\beta$  is 0.2,  $\eta$  is 0.01, and  $\lambda$  is 0.1. The equation (7) is minimized with learning rate of  $1e-4$ . We used ADAM (Kingma & Ba, 2015) optimizer with its parameters not shared across the workers.

**Baselines** Our first baseline, ‘ICM-pixels’, contains only forward model to predict next observation in pixel space.

Architecture follows the ICM feature encoder architecture to encode input observation into features  $\phi(s_t)$ , followed by 3 deconvolutional layers of kernel size  $3 \times 3$ , stride of 2 and 32 filters. A final deconvolutional layer generates the next observation in pixel space. The value of  $\eta$  is 0.5 for ‘ICM-pixels’.

Our second baseline, ‘ICM-aenc’, follows the same architecture of ‘ICM-pixels’. The difference is in the way curiosity bonus is calculated in the feature space  $\phi(s_t)$ . The value of  $\eta$  is 0.01 for ‘ICM-aenc’.

### References

- Clevert, Djordj-Arné, Unterthiner, Thomas, and Hochreiter, Sepp. Fast and accurate deep network learning by exponential linear units (elus). *arXiv:1511.07289*, 2015.
- Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *ICLR*, 2015.
- Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.
- Mnih, Volodymyr, Badia, Adria Puigdomenech, Mirza, Mehdi, Graves, Alex, Lillicrap, Timothy P, Harley, Tim, Silver, David, and Kavukcuoglu, Koray. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.