[Supplementary Material] Curiosity-Bottleneck: Exploration by Distilling Task-Specific Novelty

Youngjin Kim [1 2] **Wontae Nam*** [3] **Hyunwoo Kim*** [2] **Ji-Hoon Kim** [4] **Gunhee Kim** [2]

1. Algorithm Details

We explain the hyperparameters of PPO (Schulman et al., 2017) algorithm and our model architecture. We implement most of the experiments based on code[1] released by Burda et al. (2019). Thus, on *Treasure Hunt* and Atari environments, we use their default setting for preprocessing and hyperparameters of PPO .

1.1. Preprocessing Details

Table 1 describes the details of the preprocessing methods that we use for *Treasure Hunt* and Atari environments. We divide gray-scale observations by 255 and stack four frames for the policy input. For exploration models, we normalize gray-scale observations by subtracting the running mean and then dividing by the running standard deviation. We then clip the normalized observations to be within a range of $[-5, 5]$.

In static image classification experiments, we simply divide the input image by 255 to keep pixel values within $[0, 1]$.

Table 1. Preprocessing details for the experiments on Treasure Hunt and Atari environments.

HYPERPARAMETERS	VALUE
GRAY-SCALE	TRUE
OBSERVATION SIZE	84×84
EXTRINSIC REWARD CLIPPING	$[-1, 1]$
INTRINSIC REWARD CLIPPING	FALSE
MAX FRAMES PER EPISODE	18K
TERMINAL ON LOSS OF LIFE	FALSE
MAX AND SKIP FRAMES	4
RANDOM STARTS	FALSE

*Equal contribution [1]NALBI Inc. [2]Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea [3]Machine Learning Lab, KC Co. Ltd., South Korea [4]Clova AI Research, NAVER Corp., South Korea. Correspondence to: Gunhee Kim <gunhee.kim@snu.ac.kr>, Ji-Hoon Kim <genesis.kim@navercorp.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]https://github.com/openai/random-network-distillation

1.2. PPO Hyperparameters

We choose PPO with generalized advantage estimation (Schulman et al., 2015) as our base policy optimization algorithm. Table 2 summarizes the hyperparameters of the PPO algorithm. The complete implementation of PPO and all other exploration methods can be found in the codes accompanying this paper.

1.3. CNN Architectures

For CNN architectures, we use a standard convolutional encoder followed by a single fully connected prediction layer for all environments. Table 3–4 summarizes our model architecture for the static image classification task and the *Treasure Hunt* and Atari experiments, respectively. K is the kernel size, C is the number of channels, S is the stride size, P in the max-pool operation is the size of pooling regions, and the number within the fully-connected operation is the output dimension.

2. Experimental Details

2.1. Static Image Classification

2.1.1. HYPERPARAMETERS FOR DISTRACTION

Table 5 summarizes the hyperparameters for three distraction types in the static image classification task.

2.1.2. THE SDR SCORE

Unfortunately, there is no off-the-shelf metric to quantify how much a novelty detection algorithm is resistant to distraction. Therefore, we introduce a new evaluation metric, the *Signal-to-Distraction Ratio* (SDR) score, which can quantify the robustness of exploration methods to task-irrelevant information. The *SDR* score is defined as

$$\text{SDR} = \frac{\sum_{p_d \in \mathbf{s}_d} \sigma(\mathbf{M}_{\cdot,p_d}) \text{Corr}(\mathbf{M}_{\cdot,p_d}, \mathbf{s}_t)}{\sum_{p_t \in \mathbf{s}_t} \sigma(\mathbf{M}_{p_t,\cdot}) \text{Corr}(\mathbf{M}_{p_t,\cdot}, \mathbf{s}_d)},$$
$$\text{where } M_{p_t,p_d} = \frac{1}{N} \sum_{i=1}^{N} m(x_i, p_t, p_d) \quad (1)$$

where $\mathbf{s}_d = \mathbf{s}_t = [0.1, 0.2, \ldots, 0.9]$ is respectively a set of the retention ratio p_t and the distraction probability p_d

Table 2. Hyperparameters for PPO algorithms.

HYPERPARAMETERS	VALUE
ROLLOUT LENGTH	128
COEFFICIENT OF EXTRINSIC REWARD	2
COEFFICIENT OF INTRINSIC REWARD	1
NUMBER OF PARAMETER UPDATES ON *Treasure Hunt*	10K
NUMBER OF PARAMETER UPDATES ON ATARI	40K
NUMBER OF PARALLEL ENVIRONMENTS ON *Treasure Hunt*	16
NUMBER OF PARALLEL ENVIRONMENTS ON ATARI	64
LEARNING RATE	0.0001
OPTIMIZATION ALGORITHM	ADAM (KINGMA & BA, 2015)
ENTROPY COEFFICIENT	0.001
DECAY PARAMETER γ_e FOR EXTRINSIC REWARD	0.999
DECAY PARAMETER γ_i FOR INTRINSIC REWARD	0.99

Table 3. The architecture configure for the static image classification task.

INPUT IMAGE $x \in R^{28\times28}$
CONV($K=5$, $C=32$, $S=1$)
RELU
MAX-POOL($P=2$, $S=2$)
CONV($K=5$, $C=64$, $S=1$)
RELU
MAX-POOL($P=2$, $S=2$)
FLATTEN
FULLY-CONNECTED (60)
$z \sim N(\mu_z, \sigma_z)$, $z \in R^{30}$
FULLY-CONNECTED (10)
SOFTMAX

Table 4. The architecture configure for the *Treasure Hunt* and Atari experiments.

STACKED INPUT IMAGES $x \in R^{84\times84\times4}$
CONV($K=8$, $C=32$, $S=4$)
RELU
CONV($K=4$, $C=64$, $S=2$)
RELU
CONV($K=3$, $C=64$, $S=1$)
RELU
FLATTEN
FULLY-CONNECTED (128)
$z \sim N(\mu_z, \sigma_z)$, $z \in R^{64}$
FC (1)

of our consideration in a vector form. $\sigma(\cdot)$ is the standard deviation for a given vector.

$m(x_i, p_t, p_d)$ is the quantified novelty measure for a test observation x_i at specific p_t and p_d, and thus $\mathbf{M}_{p_t,p_d}$ denotes the average of the novelty measure for N number of all test observations. Therefore, $\mathbf{M}$ is the matrix of the heat map in Fig. 1, and the rows and columns of $\mathbf{M}$ are vectors each corresponding to $\mathbf{M}_{p_t,\cdot}$ and $\mathbf{M}_{\cdot,p_d}$ at specific p_t or p_d. Corr$(\mathbf{M}_{\cdot,p_d}, \mathbf{s}_t)$ is the correlation between the novelty values and the retention ratio $\mathbf{s}_t$ at p_d . Since $\mathbf{s}_t$ is linearly increasing from 0.1 to 0.9, it can measure how similarly the novelty values changes according to $\mathbf{s}_t$ values. Likewise, Corr$(\mathbf{M}_{p_t,\cdot}, \mathbf{s}_d)$ is the correlation between the novelty values and the distraction probability $\mathbf{s}_d$ at p_t.

The numerator of SDR is the average of σ-weighted correlation between between the novelty values and the retention ratios $\mathbf{s}_t$ for all distraction probabilities $p_d \in \mathbf{s}_d$. The denominator has the similar form with the numerator only expect that $\mathbf{s}_t$ and $\mathbf{s}_d$ are swapped. In sum, the SDR score is the highest when an algorithm generates the same $\mathbf{M}$ with the ideal color map (*i.e.* the novelty values linearly decrease as p_t increases at every fixed p_d, and at the same time the values are constant as p_d increases at every fixed p_t). Therefore, a higher *SDR* score indicates a novelty measure that is more tolerant to distractive information.

Table 5. Hyperparameters for three distraction types that we use in the static image classification task.

TYPE	HYPERPARAMETER	VALUE
RANDOM BOX	BOX SIZE	7×7
	PIXEL NOISE	$\eta \sim N(0, 0.3)$
	NUMBER OF BOXES	$n \sim$ UNIFORM$(2, 8)$
OBJECT	IMAGE PATCH SIZE	12×12
PIXEL NOISE	PIXEL NOISE	$\eta \sim N(0, 0.3)$

2.2. Treasure Hunt

Grid-world. The map size is 11×11, and 5 actions are possible by the agent: 4 directional moves and standing still.

Target items. A target item is hidden somewhere in the map.

It becomes visible to agent only if the distance between them is less than 4. However, if the agent steps out of that region, the item is invisible again.

Rewards. When the agent lands on the exact position of the item, it gains 3 points.

Sequences of items. Once the agent earns an item, another hidden item is randomly created somewhere else in the map.

Traces of items. Up to 3 traces of recently obtained items remain visible in light grey pentagons.

Movement onset condition. A distraction box occurs at a random location when the agent remains stationary in a specific location for more than 7 steps out of 8.

Location onset condition. A distraction box occurs at a random location when the agent is inside a 3×3 sized region at each corner on the map.

2.2.1. Hyperparameters for Distraction

Table 6. Hyperparameters for *Random Box* distraction in *Treasure hunt*.

Hyperparameter	Value
Box size	1×1 Block in the map
Pixel Noise	$\eta \sim N(0, 0.3)$
Number of boxes	$n \sim$ Uniform$(5, 50)$

2.3. Atari

2.3.1. Hyperparameters For Distraction

Table 7 summarizes the hyperparameters for *Random Box* distraction that we use in Atari environments.

Table 7. Hyperparameters for *Random Box* distraction in Atari.

Hyperparameter	Value
Box size	12×12
Pixel Noise	$\eta \sim N(0, 255. \times 0.3)$
Number of boxes	$n \sim$ Uniform$(10, 40)$
Distraction Probability	0.1

3. More Experimental Results

3.1. Heat maps on MNIST Dataset

We present more heat map examples similar to Fig.4 for the MNIST dataset in Fig.1. Our method can produce more similar heat maps to the ideal cases that baselines. Note that the heat maps for our method in Fig.1 (c) show relatively slanted gradation compared to the ones for the Fashion-MNIST dataset. It is due to the lower complexity of the MNIST dataset that makes distractive information more stand out.

3.2. Grad-CAM Activation Maps

Fig.2 showcases examples of *Grad-CAM* (Selvaraju et al., 2017) visualizations of distraction-free observations at test time using the agent that is trained with *location* and *movement* onset conditions of distraction. The PPO agent equipped with *Curiosity-Bottleneck* (See Fig.2 (c)), like all other models, focuses on task-relevant information.

References

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In *ICLR*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. URL http://arxiv.org/abs/1506.02438.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL http://arxiv.org/abs/1707.06347.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

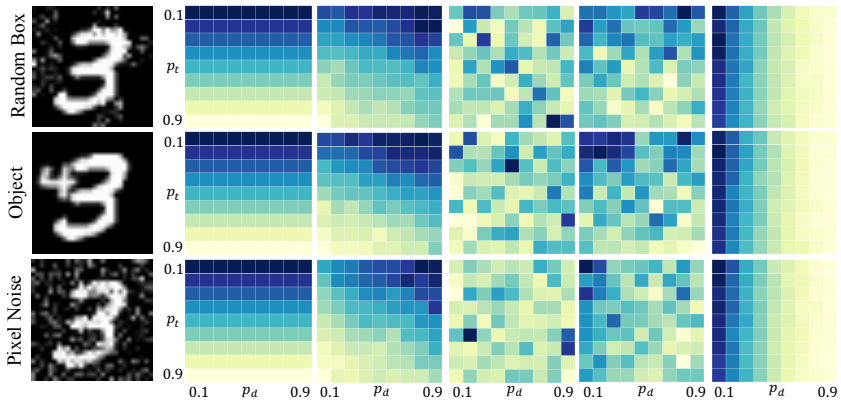

Figure 1. (a) Sample MNIST images corrupted by three types of distraction. (b-f) Heat maps show novelty measures of test images for different retention ratios p_t (vertical coordinate) and distraction probabilities p_d (horizontal coordinate). Gradual variation along the vertical axis indicates that the model correctly detects the strength of novelty, while no variation along horizontal axis means that the model perfectly ignores the novelty of task-irrelevant distractions. Dark blue indicates high novelty values. (b) Heat maps for ideal novelty detection. (c) CB can produce the most similar heat maps to the ideal cases. (d-f) The other baselines does not present consistent gradation along vertical axis.

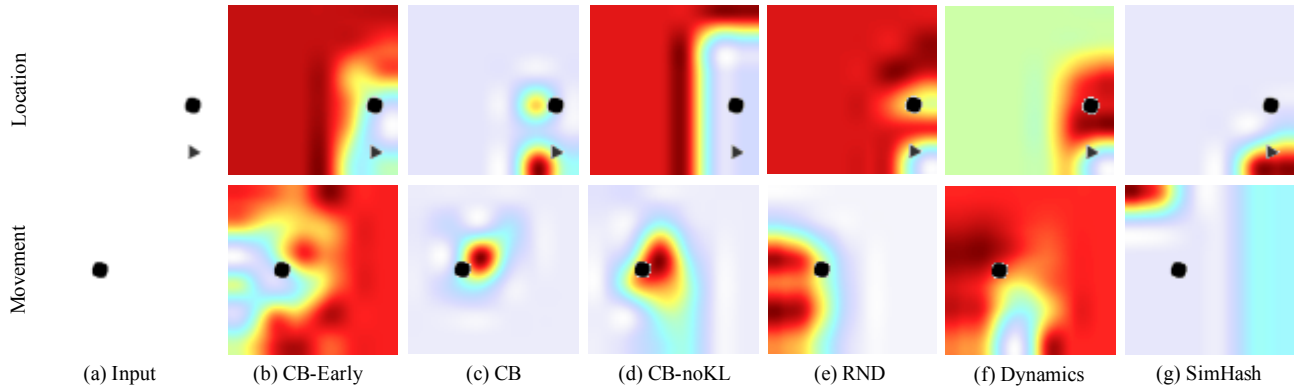

Figure 2. More *Grad-CAM* visualizations of distraction-free observations for the PPO agent that is trained with *Curiosity-Bottleneck* and baselines on *location* and *movement* onset conditions of distraction. We show gradient activation maps of (a) two examples (top and bottom) with two onset conditions including *location* (top row) and *movement* (bottom row). The black circle indicates the agent location, and the dark red color indicates large gradient values in the last CONV layer for the policy. (b) In the early stage, our method has yet to learn task-relevant information. (c-g) *CB* and all the other methods learn to focus on task-relevant information after experiencing more extrinsic rewards.