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]

Abstract

Exploration based on state novelty has brought great success in challenging reinforcement learning problems with sparse rewards. However, existing novelty-based strategies become inefficient in real-world problems where observation contains not only task-dependent state novelty of our interest but also task-irrelevant information that should be ignored. We introduce an information-theoretic exploration strategy named *Curiosity-Bottleneck* that distills task-relevant information from observation. Based on the information bottleneck principle, our exploration bonus is quantified as the compressiveness of observation with respect to the learned representation of a compressive value network. With extensive experiments on static image classification, grid-world and three hard-exploration Atari games, we show that *Curiosity-Bottleneck* learns an effective exploration strategy by robustly measuring the state novelty in distractive environments where state-of-the-art exploration methods often degenerate.

1. Introduction

In reinforcement learning (RL), an agent learns to interact with an unknown environment by maximizing the cumulative reward. In this process, the agent should determine whether to take the best sequence of actions based on previous experiences or to explore different actions in the hope of discovering novel and potentially more rewarding trajectories. This well-known dilemma is often coined as the *exploration-exploitation* tradeoff.

Choosing an appropriate exploration strategy becomes more crucial especially in an environment where observation also contains novel but *task-irrelevant* information[1]. For example, suppose a robot navigating through a crowded street, where it visits known locations (states) while facing various strangers (task-irrelevant novelty). The robot should not classify a state as novel because of strangers; instead, it should ignore the distractions to effectively reach its destination and obtain rewards precisely. Although such situations are widespread in real-world problems (*e.g.* navigating drones in a crowd), many recent exploration strategies for policy optimization (Mohamed & Rezende, 2015; Houthooft et al., 2016; Pathak et al., 2017; Burda et al., 2019a; Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017; Choi et al., 2019) are designed to be effective in environments where observation is well-aligned to the target task such as Atari games (Bellemare et al., 2013). Through a series of experiments, we observe that those approaches are often inaccurate to capture the state novelty when observation contains such novel but task-irrelevant information.

In this work, we propose an information-theoretic approach to measuring state novelty in distractive environments. Our method is *task-specific* in that it learns to identify the target task using sparse extrinsic rewards and filters out task-irrelevant or distractive information from observation when quantifying the state novelty. Motivated by neural network's ability to learn a compressive representation (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017), we propose to quantify *the degree of compression* of observation with respect to the latent representation of a compressive value network, and use it as a surrogate metric for task-specific state novelty as intrinsic reward. The proposed exploration algorithm is referred to as *Curiosity-Bottleneck* since it introduces the information bottleneck (IB) principle (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017; Alemi et al., 2017; Alemi & Fischer, 2018; Alemi et al., 2018b;a) to exploration problems to comprise following properties: (i) encoding an observation of a higher probability to be more compressive in representation and (ii) omitting task-irrelevant information while learning a compressive representation. The degree of compression of observation is estimated by the variational upper-bound of mutual information between observations and learned neural network representations, which can be efficiently computed in a closed

[*]Equal contribution [1]NALBI Inc. [2]Seoul National University, 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] The *task-irrelevant* information refers to ones that affect neither the agent nor the target.

form. Moreover, the *Curiosity-Bottleneck* is integrable with any policy optimization algorithms and naturally scalable to high-dimensional observations. Although there has been an apporach that uses information theoretic approach with action-predictive quality to enhance exploration (Still & Precup, 2012), to the best of our knowledge, this work is the first to introduce the value-predictive information bottleneck for exploration in RL problems.

We perform various qualitative and quantitative experiments in static image classification, customized Grid-world environment named *Treasure Hunt* and three hard-exploration Atari games (Bellemare et al., 2013) including Gravitar, Montezuma's Revenge, and Solaris. We show that the *Curiosity-Bottleneck* accurately captures the state novelty in distractive environments where state-of-the-art methods degenerate due to their over-sensitivity to some unexpected visual information in the environment. We also provide an in-depth analysis of the learned representation and adaptive exploration strategy.

2. Related Work

A majority of task-agnostic exploration strategies in deep RL context quantify the novelty of observation in terms of counts (Bellemare et al., 2016; Ostrovski et al., 2017), pseudo-counts (Bellemare et al., 2016; Tang et al., 2017; Choi et al., 2019), information gain (Houthooft et al., 2016; Chen et al., 2017) prediction error (Schmidhuber, 1991; Stadie et al., 2015; Achiam & Sastry, 2016; Pathak et al., 2017; Haber et al., 2018; Fox et al., 2018; Burda et al., 2019b), or value-aware model prediction (Luo et al., 2019; Farahmand et al., 2017). Despite the significant improvement they have brought on hard exploration tasks, this group of exploration strategies struggles to provide a meaningful metric for exploration when observation contains information that is irrelevant to the target task.

The degeneration in distractive environments partially originates from the task-agnostic objective for intrinsic reward functions. For example, count or pseudo-count based approaches (Tang et al., 2017; Bellemare et al., 2016) encode an observation into a feature space before they allocate the observation to a cluster. Since the feature space is obtained by a deterministic encoder or an autoencoder trained to reconstruct input images, those methods would misallocate an observation to a novel cluster when the observation contains familiar task-related information and novel task-irrelevant information. The same analysis holds for information gain and prediction based approaches. Most of those approaches learn to preserve information about state dynamics (*i.e.* state transition) or inverse-dynamics. However, they are easily deceived by an unpredictable transition of visual stimulus; such phenomenon is called the *Noisy-TV* problem (Burda et al., 2019a). Though some recent methods (Savinov et al., 2019; Burda et al., 2019b) are immune to the Noisy-TV problem, they do not have mechanisms to prioritize task-related information above task-irrelevant one.

Exploration methods in the temporal-difference learning (*e.g.* deep Q-learning) can provide a natural way of incorporating task-relatedness into exploration. Many exploration strategies in this group rely on the principle of *optimism in the face of uncertainty* (Lai & Robbins, 1985). It encourages an agent to explore by choosing an action that yields some uncertainty about the action-value estimates. Classical examples utilize upper confidence bound (Auer et al., 2002) and Thompson sampling (Thompson, 1933) for the stochastic sampling of actions. Recent algorithms incorporate these ideas with finer uncertainty approximations, to be applicable to extremely large state-spaces with deep exploration (Osband et al., 2016; Chen et al., 2017; O'Donoghue et al., 2018; Fortunato et al., 2018). Although they provide a way to indirectly incorporate state novelty to the target task via the minimization of overall uncertainty, there is no explicit mechanism to prune out the uncertainty caused by task-irrelevant perturbations. Another limitation is that their algorithmic (*e.g.* the temporal-difference learning) or architectural (*e.g.* Bayesian neural network) assumptions hinder extension to policy optimization algorithms.

Therefore, it is desirable to have an exploration approach that not only takes advantage of plug-and-play novelty measures but also is capable of filtering out task-irrelevant information by identifying the target task and learning to exclude distractions from its representation.

3. Preliminaries of Information Bottleneck

We introduce some background on information bottleneck (IB) principle (Tishby et al., 2000) and variational information bottleneck (VIB) (Alemi et al., 2017). Our *Curiosity-Bottleneck* is closely related to VIB since it learns compressive yet informative representation using VIB framework, which is key to quantifying task-specific state novelty.

Let the input variable X and the target variable Y be distributed according to some joint data distribution $p(x, y)$. The IB principle provides an objective function to obtain a compressive latent representation Z from the input X while maintaining the predictive information about the target Y:

$$\min -I(Z;Y) + \beta I(X;Z) \tag{1}$$

where $I(\cdot;\cdot)$ is mutual information (MI) and $\beta \geq 0$ is a Lagrange multiplier. The first term in Eq.(1) ensures the latent representation Z to be predictive about the target, while the second term forces Z to ignore irrelevant information from the input X. As a consequence, the learned representation generalizes better, is robust to adversarial attack (Alemi et al., 2017), is invariant to nuance factors (Achille

& Soatto, 2018a), and prevents weight over-fitting (Alemi et al., 2018b; Achille & Soatto, 2018b; Vera et al., 2018).

Alemi et al. (2017) propose a variational approximation of IB that is intuitively applicable to supervised learning problems. The VIB can derive variational lower bounds of the two MI terms in the IB objective. First, minimizing the upper bound of $-I(Z;Y)$ is equivalent to optimizing a standard supervised learning objective:

$$\begin{aligned} -I(Z;Y) &= -\int p(z,y)\log\frac{p(y|z)}{p(y)}dzdy \\ &\leq -\int p(z,y)\log\frac{q(y|z)}{p(y)}dzdy \\ &= \mathbb{E}_{z,y}[-\log q(y|z)] - H(Y), \end{aligned} \tag{2}$$

where $q(y|z)$ is a variational approximation of $p(y|z)$, and the inequality holds because $\mathrm{KL}[p(Y|Z)\|q(Y|Z)] \geq 0$. The entropy of label $H(Y)$ can be ignored since it is often independent of the objective optimization.

For the second term $\beta I(X;Z)$, we minimize the upper bound of $I(X;Z)$ by optimizing the KL-divergence between the posterior $p(Z|X)$ and a variational approximation $r(Z)$ of the marginal distribution $p(Z)$:

$$\begin{aligned} I(X;Z) &= \int p(z,x)\log\frac{p(z|x)}{p(z)}dzdx \\ &\leq \int p(z,x)\log\frac{p(z|x)}{r(z)}dzdx \\ &= \mathrm{KL}[p(Z|X)\|r(Z)], \end{aligned} \tag{3}$$

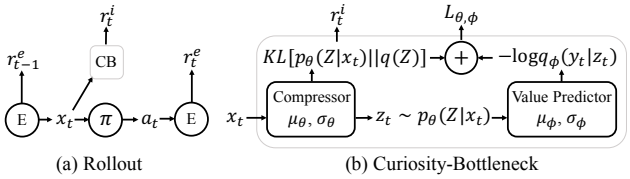
where the inequality holds because $\mathrm{KL}[p(Z|X)\|r(Z)] \geq 0$.

Although Peng et al. (2019) apply the VIB to RL problems, they focus on improving the discriminator of generative adversarial networks in imitation learning tasks. To the best of our knowledge, this work is the first to utilize VIB's capability of learning compressive representation and detecting out-of-distribution data (Alemi et al., 2018a) for exploration in RL problems. We propose the *Curiosity-Bottleneck* which ignores the task-irrelevant information (*i.e.* distractions) by using $\mathrm{KL}[p(Z|x)\|q(Z)]$ as the novelty measure.

4. Approach

In Section 4.1, we introduce an information-theoretic approach for learning a compressor model named *Curiosity-Bottleneck* (*CB*). *CB* can quantify task-specific novelty from observation. In Section 4.2, we describe the novel behavior of *CB* that leads to adaptive exploration respective to the agent's competence in the task. In Section 4.3, we describe how to plug our method into policy optimization algorithms.

Fig.1 shows the overview of our approach. We assume a standard RL setting where an agent interacts with environment E by getting an observation x_t, executing an action a_t sampled from its current policy π and receiving extrinsic reward r^e_t and intrinsic reward r^i_t. The role of *CB* is to compute the intrinsic reward.



(a) Rollout (b) Curiosity-Bottleneck

Figure 1. Overview of *Curiosity-Bottleneck* (*CB*). (a) An agent interacts with environment E by getting an observation x_t, executing an action a_t sampled from its current policy π and receiving extrinsic reward r^e_t and intrinsic reward r^i_t generated by *CB*. (b) In *CB*, the compressor represents the observation x_t in a latent space according to a posterior distribution $p_\theta(Z|x_t)$. The value predictor takes the representation z_t as input and predicts the target value y_t. The KL-divergence, which is the per-instance approximation of $I(Z;X)$, quantifies the *degree of compression* of x_t with respect to the learned compressor. It becomes the intrinsic reward r^i_t. The prediction error $-\log q_\phi(y_t|z_t)$ with the KL-divergence forms the objective of *CB*, $\mathcal{L}_{\theta,\phi}$. *CB* allows task-specific exploration in a distractive environment since it lets the model discard as much information from x_t as possible via the KL-divergence and retain information that is useful to predict y_t via the prediction error.

4.1. The *Curiosity-Bottleneck*

The key to the *CB* is to obtain a compressor model $p_\theta(Z|X)$ whose output representation Z satisfies the three desiderata.

- Minimize the average code-length of observation X to obtain a meaningful novelty measure. It is based on Minimum Description Length (MDL) principle (Rissanen, 1978), which describes a one-to-one correspondence between a code length function and a probability distribution; it encodes a rare observation to a lengthy code and a common observation to a shorter one. This criterion motivates us to minimize the entropy $H(Z)$ that can be seen as an average code length of a random variable (Cover & Thomas, 2006).
- Discard as much information about observation X as possible to exclude task-irrelevant information. This motivates us to disperse $p_\theta(Z|X)$ by maximizing the entropy $H(Z|X)$.
- Preserve information related to target variable Y to include the meaningful information for the task. In our setting, Y is a value estimate since extrinsic rewards indirectly define the task in RL problems. This criterion can be addressed by maximizing mutual information $I(Z;Y)$.

For optimizing the above three desiderata, we derive an

objective function for our compressor model as

$$\min_{\theta} -I(Z;Y) + \beta I(X;Z), \tag{4}$$

where we use the definition of mutual information (MI) $I(X;Z) = H(Z) - H(Z|X)$. β is a non-negative coefficient that trades off the relative importance of compression and relevance to the task.

The MI between the input variable and the code has been often used as a metric for quantifying the degree of compression (Cover & Thomas, 2006). We thus use the per-instance mutual information $i(x;Z)$ as the novelty metric for observation x; that is, $i(x;Z) = \int_z p(z|x) \log \frac{p(x,z)}{p(x)p(z)} dz$ becomes our intrinsic reward function where $I(X;Z) = \int_x p(x) i(x;Z) dx$. However, $I(X;Z)$ is intractable in general; instead, we estimate its variational upper bound.

Interestingly, Eq.(4) has the same form of IB objective as discussed in Section 3. Hence, a tractable variational approximation to the objective is derivable by plugging Eq.(2)–(3) to Eq.(4):

$$\mathcal{L}_{\theta,\phi} = \mathop{\mathbb{E}}_{x,y} [-\log q_\phi(y|z) + \beta \mathrm{KL}[p_\theta(Z|x) \| q(Z)]], \tag{5}$$

where q indicates a variational distribution, z is sampled from posterior $p_\theta(Z|x_n)$ and θ and ϕ denote the parameters of the compressor and the value predictor respectively as presented in Fig.1 (b). Finally, we can represent our intrinsic reward function for observation x_n in a KL-divergence term:

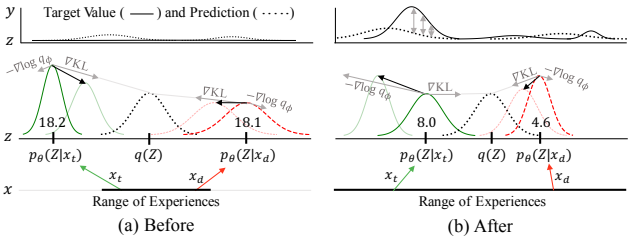
$$r_i(x_n) = \mathrm{KL}[p_\theta(Z|x_n) \| q(Z)]. \tag{6}$$

Using the KL-divergence that approximates $I(Z;X)$ as a novelty measure is also supported by (Alemi et al., 2018a), which show that $\mathrm{KL}[p_\theta(Z|x_n) \| q(Z)]$ itself is a sound uncertainty metric for out-of-distribution detection.

In practice, we assume a Gaussian distribution for $q_\phi(y|z) = N(\mu_\phi(z), \sigma^2)$. We use a simple fully-connected layer that outputs the mean $\mu_\phi(z) \in \mathbb{R}$ of y. We set a constant variance σ^2 so that $\log q_\phi(y|z)$ in Eq.(5) reduces to the mean-squared error (*i.e.* a standard value loss).

We also assume a Gaussian distribution for both compressor output distribution $p_\theta(z|x) = N(\mu_\theta(x), \sigma_\theta(x))$ and variational prior $q(z) = N(0, I)$. The compressor network consists of a standard three-layer convolutional neural network followed by an MLP that outputs both the mean $\mu_\theta(x) \in \mathbb{R}^K$ of z and the diagonal elements of covariance matrix $\sigma_\theta(x) \in \mathbb{R}^K$. We use the reparameterization trick (Kingma & Welling, 2014) to sample $z = \mu_\theta(x) + \epsilon \sigma_\theta(x)$ in a differentiable way with an auxiliary random variable $\epsilon \sim N(0, I)$. In this setting, the intrinsic reward is computed in a closed form as

$$r_i(x) = \frac{1}{2} \sum_{k}^{K} \mu_{\theta,k}^2(x) + \sigma_{\theta,k}^2(x) - \log \sigma_{\theta,k}^2(x) - 1. \tag{7}$$



(a) Before (b) After

Figure 2. Illustration of adaptive exploration of our method. Suppose x_t is task-relevant observation, x_d is task-irrelevant one. Solid lines and dotted lines on the top row are the target values and the predictions of *CB*, respectively. The black arrow indicates the sum of gradients of two terms in an objective function in Eq.(5) that transforms the posterior to a different shape depicted by the blurry curve at the end of the black arrow. The numbers within the posterior curves are the intrinsic rewards for the observation. (a) Before having sufficient experience of receiving external reward signals, $\mathrm{KL}[p_\theta(Z|x) \| q(Z)]$ pulls posteriors to the marginal $q(Z)$, while the value prediction loss $-\log q_\phi(y|z)$ makes relatively little contribution. (b) The loss $-\log q_\phi(y|z)$ largely contributes to shaping meaningful posteriors after collecting sufficient experiences.

4.2. Adaptive Exploration

One outstanding property of *CB* is adaptive exploration respective to the agent's competence in the target task. Induced by changing the balance between KL-divergence term and negative log-probability term in the objective function of Eq.(5), *CB* automatically shifts its exploration strategy from the task-identification phase to the task-specific exploration phase. Fig.2 illustrates how our method adaptively calibrates intrinsic reward by identifying the target task. Both task-relevant observation x_t and distractive observation x_d are mapped to Gaussian posteriors $p_\theta(Z|x)$ on the middle row in Fig.2. The KL-divergence term always reduces the intrinsic reward for observation x by forcing the posterior $p_\theta(Z|x)$ to collapse to the marginal $q(Z)$ as we denote using gray-colored arrow and ∇KL. The negative log-probability term often increases the intrinsic reward for x by encouraging $p_\theta(Z|x)$ to be a meaningful posterior in order to accurately predict the target value that is built from the previous experiences of external rewards. The two terms together change the intrinsic reward of an observation by transforming the posterior to a different shape which is indicated by blurry posterior at the end of the black arrow. Specifically, changes in the target values result in two distinct exploration phases.

Task-identification. In RL problems with sparse rewards, an agent often has no experience of receiving extrinsic reward signals at the early training steps (See Fig.2 (a)). Then, the target values are zero for all observations and the value predictor achieves an arbitrarily small prediction loss (*i.e.* negative log-probability) simply by collapsing model param-

eters to zero. Such constant target values are illustrated by a solid line in the top row and the predicted values are flat as a dotted line. In this situation, the negative log-probability term contributes little to making different intrinsic rewards among observations.

For this reason, the KL-divergence term determines the landscape of the intrinsic reward function over observations. The posterior $p_\theta(Z|x)$ of frequently seen observations become closer to the marginal $q(Z)$. Thus, no matter whether x is task-relevant or not, if x is novel, the KL-divergence induces a high reward, resulting in a task-agnostic prediction-based exploration strategy. Hence, the agent should explore a wide range of the observation space, including distractive regions, until it receives enough extrinsic reward signals (*i.e.* identifying the target task) so that negative log-probability can make a meaningful contribution to the posterior shaping.

Task-specific exploration. After collecting sufficient extrinsic rewards, *CB* gradually calibrates intrinsic rewards by considering relevance to the target task as illustrated in Fig.2 (b). That is, the prediction loss pushes $p_\theta(Z|x)$ to have different shapes from the prior $q(Z)$ in order to construct accurate mappings to the target values and increase $r^i(x) = \mathrm{KL}[p_\theta(Z|x)\|q(Z)]$ at x. As a result, *CB* allocates high intrinsic rewards to observations that satisfy two joint conditions of rareness and task-relevance. Extensive analysis using *Grad-CAM* (Selvaraju et al., 2017) in Section 5.2 visualizes this behavior more clearly.

4.3. Plugging into Policy Optimization Algorithms

CB can be plugged into any RL algorithms that use intrinsic reward functions. *CB* scales well to large parallel environments that require processing a large number of samples, since it is simple to implement and requires a single forward pass to the compressor network to compute intrinsic rewards. In this work, we mainly use the Proximal Policy Optimization (PPO) (Schulman et al., 2017) with two value heads to combine the intrinsic reward with the extrinsic reward as Burda et al. (2019b) suggested. We also adopt the same normalization schemes of (Burda et al., 2019b) for the intrinsic reward and observation. Algorithm 1 shows the overall picture of our method, where we omit the details of normalization, hyperparameters and PPO algorithms for readability. More details can be found in the supplementary file and the code which is available at http://vision.snu.ac.kr/projects/cb.

Algorithm 1 *Curiosity-Bottleneck* with PPO

Given current time step t_0, the number of rollouts N, the number of optimization steps N_{opt}.
for $t = t_0$ **to** $t_0 + N$ **do**
  Sample $a_t \sim \pi(a_t|x_t)$
  Sample $x_{t+1}, r^e_t \sim p(x_{t+1}, r^e_t|x_t, a_t)$
  Calculate $r^i_t \leftarrow \mathrm{KL}[p_\theta(Z|x_t)\|q(Z)]$
end for
Calculate returns R^e and advantages A^e for r^e
Calculate returns R^i and advantages A^i for r^i
$y_n \leftarrow R^e_n$ where $n \in \{1, \dots, N\}$
for $j = 1$ **to** N_{opt} **do**
  Optimize PPO agent
  Optimize θ and ϕ w.r.t. $\mathcal{L}_{\theta,\phi}$ in Eq.(5)
end for

5. Experiments

We design three environments to inspect different aspects of our *CB* method. First, we perform static classification tasks on MNIST (LeCun & Cortes, 2010) and Fashion-MNIST (Xiao et al., 2017) to see whether the *CB* intrinsic reward of Eq.(6) is a consistent novelty measure that can ignore various visual distractions irrelevant to the target label (Section 5.1). Regardless of the task simplicity, this experiment evaluates the *CB*'s ability to detect state novelty while isolating environment-specific factors. Second, we test on the *Treasure Hunt* as a customized grid-world environment to inspect the explorative behavior when observation contains distractive information (Section 5.2). We visualize the internal representation of our model by using the recent network interpretation method *Grad-CAM* (Selvaraju et al., 2017). We also highlight that *CB* adaptively calibrates its exploration strategy according to the agent's proficiency to the target task. Finally, we test the scalability of our method with hard-exploration games in the Atari environment (Section 5.3) using NAVER Smart Machine Learning (NSML) platform (Sung et al., 2017; Kim et al., 2018).

For comparison, we choose four baseline exploration strategies for policy optimization. As prediction-based methods, we select the random network distillation (*RND*) and the dynamics model (*Dynamics*) proposed by Burda et al. (2019b). The intrinsic reward for *RND* is the mean-squared error between two output features of a fixed encoder and a predictor network. *Dynamics* uses the mean-squared error between the two features for future observation. An encoder directly maps future observation to a feature and the predictor predicts the feature of future observation from the current one. For the two models, we use the code [2] released by the original authors. As the count-based method, we choose PPO-SimHash-BASS (*SimHash*) that uses hand-crafted feature transformation named BASS (Naddaf, 2010) within the SimHash framework (Tang et al., 2017). *Simhash* discretizes observation according to a hash function and uses the accumulated visitation count to calculate the intrinsic reward. Finally, we test a non-compressive variant of our method *CB-noKL*, which is a value function that has the same architecture as *CB*, to highlight that the explicit com-

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

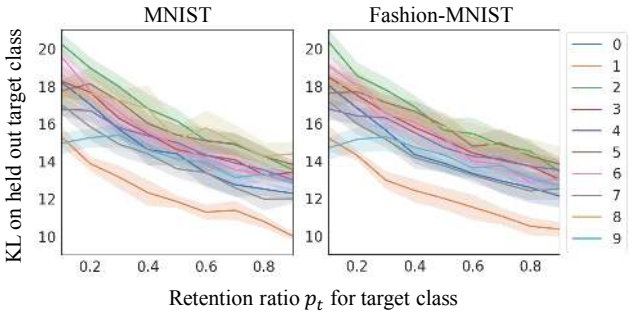


Figure 3. Novelty detection on MNIST and Fashion-MNIST. Each curve visualizes test KL-divergence $\mathrm{KL}[p_\theta(Z|X)\|q(Z)]$ on held-out target class examples over the proportions of training examples of the target class. The number in the legend indicates the target class. We draw the curves on average over 10 random seeds.

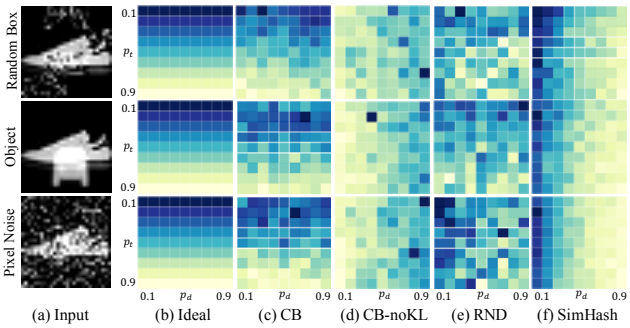


Figure 4. (a) Sample Fashion-MNIST images corrupted by three types of distraction. (b-f) Heat maps show novelty measures of test images for different retention ratios p_t (vertical coordinate) and distraction probabilities p_d (horizontal coordinate). Dark blue indicates the higher novelty values. (b) Heat maps for ideal novelty detection. (c-f) *CB* reproduces more similar heat maps to the ideal cases than the other baselines.

pression is the key to success of our task-specific novelty measure. *CB-noKL* is trained without the KL-divergence term in the objective of Eq.(5) (*i.e.* optimizing only with the cross-entropy loss). We use the posterior uncertainty $\sigma_\theta(x)$ instead of $\mathrm{KL}[p_\theta(Z|x)\|q(Z)]$ as the intrinsic reward.

5.1. Static Image Classification

We show that our *CB*'s intrinsic reward is a valid metric for state novelty. We perform static image classification where observation X is an image and target Y is the class label. We describe the details of the classifier in supplementary file. In order to make the target class rare compared to the other classes, we randomly select a target class and discard $1 - p_t$ proportions of images in the target class. Then we compare novelty metric values for different retention ratios $p_t \in \{0.1, \ldots, 0.9\}$. Fig.3 shows that the KL values of test examples of the held-out target class monotonically decrease as training examples of the target class increase, presenting that *CB* correctly measures the state novelty.

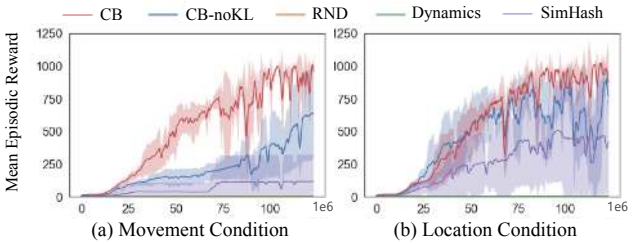
We then validate how robust our method is in the presence of task-irrelevant visual information. As done in a previous work (Zhang et al., 2018), we add various noises to visual inputs to simulate task-irrelevant information. We consider three types of visual distractions (See examples in Fig.4 (a)): (i) **Random Box** (first row) simulates the case where distractions are introduced in vastly various configurations. A random number of small 7×7 boxes appear in random positions. Each box is filled with pixel-wise noise $\eta_{i,j} \sim N(0, 0.3)$ to hinder neural networks from trivially memorizing the box. (ii) **Object** (second row) simulates facing unfamiliar objects. We add a 12×12 resized image patch of a different class to a randomly chosen position. (iii) **Pixel noise** (last row) simulates sudden sensory noise. It adds pixel-wise noise $\eta_{i,j} \sim N(0, 0.3)$ to observation. In all types, the distractions are introduced with a Bernoulli probability of $p_d \in \{0.1, \ldots, 0.9\}$. Note a smaller p_d makes the distraction more novel. We expect our model to correctly ignore such distractions irrelevant of the target task.

Fig.4 visualizes the variation of novelty measures according to the retention ratio p_t and the distraction probability p_d on Fashion-MNIST dataset. The intensity of each cell in the heat map indicates the average novelty values of test images measured by different exploration models. We train each model separately for all combinations of p_t, p_d and distraction types to fill the heat map. Test images are chosen from unseen images in the target class after corrupted by distractions. Ideally, the novelty detection method needs to generate the heat maps in Fig.4(b). That is, the variation should be gradual along the vertical axis, meaning that the model correctly detects the strength of novelty, and no variation should be along the horizontal axis, meaning that the model perfectly ignores the novelty of task-irrelevant distractions. Our *CB* method in Fig.4(c) produces the most similar heat maps to the ideal cases for all kinds of corruptions. On the other hand, the other baseline models (d-f) fail to provide consistent novelty metrics since they have no vehicle to process task-relevance in observation selectively. Note that we exclude *Dynamics* since they are not applicable to the static task. (*i.e.* it assumes temporal dependence).

Quantitative analysis on the heat maps makes clear distinction of *CB* from the other baselines. We introduce a novel evaluation metric, *Signal-to-Distraction Ratio* (SDR) score, which evaluates the robustness of a novelty measure to distractive information. A higher *SDR* score indicates that a novelty measure is more tolerant to distractive information, though exact formula and details of *SDR* score are deferred to the supplementary file. Table 1 shows that *CB* significantly outperforms the task-agnostic baselines on MNIST and Fashion-MNIST datasets for all three distraction types.

Table 1. SDR scores for *CB* and baseline models on MNIST and Fashion-MNIST for three types of distraction.

DATA	MODEL	BOX	NOISE	OBJECT
MNIST	*CB*	**2.57**	**1.76**	**2.82**
	CB-noKL	1.11	0.85	0.80
	RND	2.29	0.57	2.18
	SimHash	0.06	0.06	0.05
FASHION	*CB*	**4.97**	**1.78**	**3.09**
	CB-noKL	0.39	0.26	0.24
	RND	1.44	0.53	1.70
	SimHash	0.22	0.07	0.22



(a) Movement Condition (b) Location Condition

Figure 5. Comparison of mean episodic returns between *CB* and baselines with 5 randon seeds in Treasure Hunt environment where *Random Box* distractions are generated by two onset conditions.

5.2. Treasure Hunt

We test *CB* in a grid-world environment that requires exploration under distraction. We also provide an in-depth analysis of learned representation and exploration strategy. The environment is designed to evaluate the following capabilities of each method: (i) learning from temporally correlated samples collected by the agent, (ii) exploring efficiently until a reward signal is discovered, (iii) identifying the target task from sparsely received reward signals and (iv) ignoring visually novel but task-irrelevant distractions after collecting sufficient reward signals.

In this environment, the agent should explore until it earns the target item, which cannot be seen unless the distance between them becomes less than a certain threshold. Once the agent takes an item, it receives an extrinsic reward and the next item is created in another random location. In the example of Fig.6 (a), the agent is shown as a black circle and the target item is a black triangle but hidden in Fig.6 (a) since the agent is not close enough to it. Each episode terminates when the agent runs for $3,000$ steps. An effective exploration strategy for the agent is to explore throughout the map, undisturbed by distractions.

The distraction, visualized as gray noisy boxes in Fig.6 (a), is the *Random Box* type in Section 5.1; a random number of boxes appear in random positions. We experiment two different onset conditions for the distraction generation: (i) movement condition: distraction occurs when the agent remains stationary for a specific length of steps and (ii) location condition: when the agent is within a certain range from any corner on the map. These conditions allure the agent to the corners of the map or to immobility; hence they hinder the agent's exploration.

Fig.5 compares the maximum episodic rewards of our method and baselines with 5 random seeds for 122M rollouts. Our method significantly outperforms the other baselines in both onset conditions for distraction; *CB* learns to explore efficiently by filtering out such distractive information, while the other baselines often stops moving (movement condition) or stay near the corner (location condition).

Visualization using Grad-CAM. We compare the exploration strategies by visualizing the learned representation of their policy networks using *Grad-CAM* (Selvaraju et al., 2017). Fig.6 illustrates the gradient activation maps for the last CONV layer of the PPO agent in *Treasure Hunt* environment with both *location* and *movement* distraction onset conditions. Agents are trained with different exploration methods (b-g) for 10K updates of parameters. We also present activation maps for non-distractive observations in supplementary file.

When the agent has little experience of receiving extrinsic rewards during early phase of training (Fig.6 (b)), *CB* encourages the agent to take any novel visual information into consideration. We denote this as *CB-Early* since it shows the behavior of a premature agent that had less than 150 updates. After experiencing enough extrinsic reward throughout episodes, the agent with *CB* learns to ignore task-irrelevant distractions (Fig.6 (c)); the gradient values on the distraction regions are small while those on the useful regions to the target task are large (*e.g.* current agent locations or likely locations of target items). On the other hand, the agents with baseline exploration methods still count much on the distraction as novel information by assigning high gradient values on the distraction regions (Fig.6 (d-g)).

5.3. Hard Exploration Games

We evaluate the proposed method for visually complicated hard exploration games of Atari including Gravitar, Montezuma's Revenge, and Solaris. Experiments run for up to 327M rollouts (40K updates of parameters with 64 parallel environments). We measure the mean episodic returns of our method against baselines. All three games require extensive exploration in order to receive a sparsely distributed extrinsic reward. As the observations from Atari games are well-aligned to the target tasks (Bellemare et al., 2013), we introduce *Random Box* distraction used in previous sections to the observations. We set the distraction to occur independently in the environment with a Bernoulli probability of $p_d = 0.1$ since it is hard to localize or track the agent's

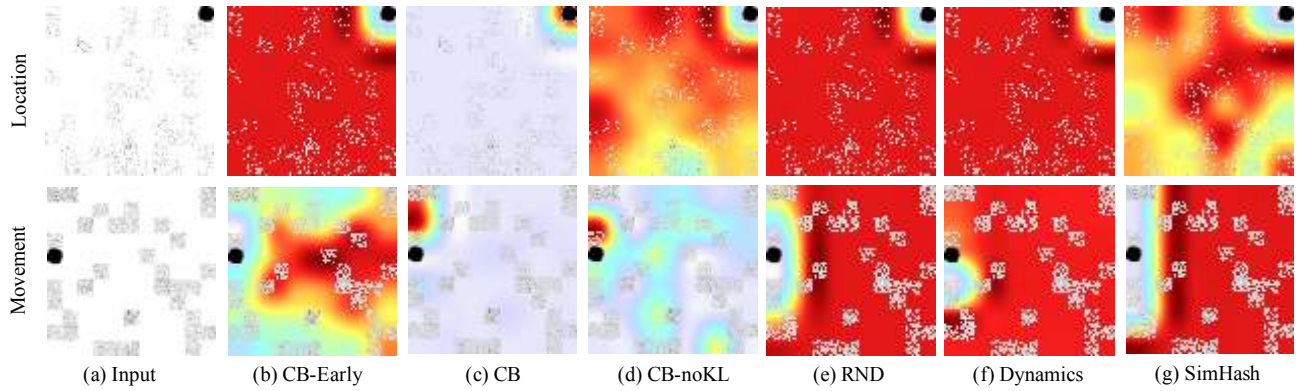


Figure 6. Grad-CAM visualization for the PPO agent that is trained with *CB* and baselines with two onset conditions for *Random Box* distraction. We show gradient activation maps of (a) two examples (top and bottom) that are corrupted by the task-irrelevant distractions. The black circle indicates the 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 encourages the policy to take distractions into account because they are novel but not yet fully determined to be task-irrelevant. (c) As experiencing more extrinsic rewards, the policy with *CB* learns selectively from the information that is useful for the task. The gradient values on the distraction regions are small while those on the useful regions to the target task are large (*e.g.* current agent locations or likely locations of target items). (d-g) Baselines still consider distractive information as novel ones by assigning high gradient values on the distraction regions.

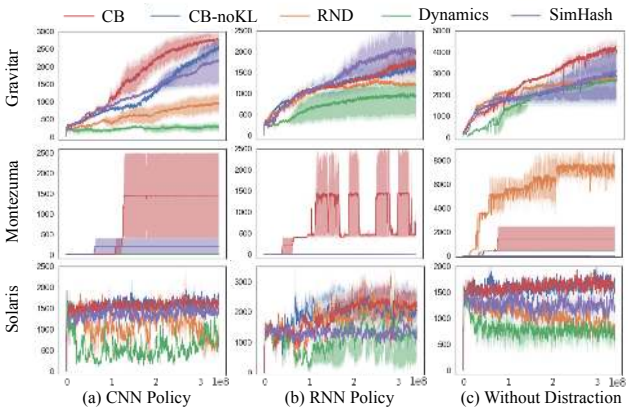


Figure 7. Mean episodic returns on three Atari games with two random seeds for 327M rollouts (40K iterations). *CB* outperforms baselines with both CNN and RNN based policy (a-b) and is still competitive without distraction (c).

movement in Atari environments.

Fig.7 shows that our method consistently outperforms strong baselines on distractive Atari games. A recurrent policy is often recommended to deal with partial observability in hard exploration problems. We thus test all methods with recurrent policy on the same distractive environment, but it does not improve performance much as in Fig.7 (b). Interestingly, our method turns out to be a competitive exploration strategy even when observation does not contain task-irrelevant distractions (Fig.7 (c)). Note that *RND* is the current state-of-the-art exploration strategy in Montezuma's Revenge.

6. Conclusion

We introduced a task-specific exploration method named *Curiosity-Bottleneck* that distills task-relevant information from observation based on the information bottleneck principle. Our internal reward is quantified as the compressiveness of observation with respect to the learned representation of an auxiliary value network. Our analysis and visual interpretation suggested that *Curiosity-Bottleneck* adaptively calibrated the goal of exploration from task-identification to task-specific exploration. A series of experiments on static classification, customized grid-world, and Atari environments confirmed that our method robustly measured the state novelty, filtering out task-irrelevant or distractive information, while previous strong baseline models often failed to disregard distractions and resulted in weaker performance. Improving our method on non-distractive environments and finding an adaptive scheduling for β, which determines the balance between compression and preservation of information, are important directions for future work.

Acknowledgements

This work is collaborated with Clova AI Research, NAVER Corp. This work is supported by Brain Research Program through the NRF of Korea (2017M3C7A1047860) and Creative-Pioneering Researchers Program through Seoul National University.

References

Achiam, J. and Sastry, S. Surprise-based intrinsic motivation for deep reinforcement learning. *NeurIPS DRL Workshop*, 2016.

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018a.

Achille, A. and Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2897–2905, 2018b.

Alemi, A. A. and Fischer, I. GILBO: one metric to measure them all. In *NeurIPS*, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *ICLR*, 2017.

Alemi, A. A., Fischer, I., and Dillon, J. V. Uncertainty in the variational information bottleneck. *UAI UDL Workshop*, 2018a.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In *ICML*, 2018b.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *NeurIPS*, 2016.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning. In *ICLR*, 2019a.

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

Chen, R. Y., Sidor, S., Abbeel, P., and Schulman, J. UCB exploration via Q-ensembles. *CoRR*, abs/1706.01502, 2017. URL http://arxiv.org/abs/1706.01502.

Choi, J., Guo, Y., Moczulski, M., Oh, J., Wu, N., Norouzi, M., and Lee, H. Contingency-aware exploration in reinforcement learning. In *ICLR*, 2019.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006.

Farahmand, A.-M., Barreto, A., and Nikovski, D. Value-aware loss function for model-based reinforcement learning. In *AIStats*, 2017.

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C., and Legg, S. Noisy networks for exploration. In *ICLR*, 2018.

Fox, L., Choshen, L., and Loewenstein, Y. DORA the explorer: Directed outreaching reinforcement action-selection. In *ICLR*, 2018.

Haber, N., Mrowca, D., Wang, S., Li, F., and Yamins, D. L. Learning to play with intrinsically-motivated, self-aware agents. In *NeurIPS*, 2018.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. VIME: variational information maximizing exploration. In *NeurIPS*, 2016.

Kim, H., Kim, M., Seo, D., Kim, J., Park, H., Park, S., Jo, H., Kim, K., Yang, Y., Kim, Y., et al. NSML: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

Luo, Y., Xu, H., Li, Y., Tian, Y., Darrell, T., and Ma, T. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *ICLR*, 2019.

Mohamed, S. and Rezende, D. J. Variational information maximisation for intrinsically motivated reinforcement learning. In *NeurIPS*, 2015.

Naddaf, Y. *Game-independent AI Agents for Playing Atari 2600 Console Games*. University of Alberta, 2010.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty bellman equation and exploration. In *ICML*, 2018.

Osband, I., Blundell, C., Pritzel, A., and Roy, B. V. Deep exploration via bootstrapped DQN. In *NeurIPS*, 2016.

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. Count-based exploration with neural density models. In *ICML*, 2017.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. In *ICLR*, 2019.

Rissanen, J. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Savinov, N., Raichuk, A., Vincent, D., Marinier, R., Pollefeys, M., Lillicrap, T., and Gelly, S. Episodic curiosity through reachability. In *ICLR*, 2019.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991.

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.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. URL http://arxiv.org/abs/1703.00810.

Stadie, B. C., Levine, S., and Abbeel, P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015. URL http://arxiv.org/abs/1507.00814.

Still, S. and Precup, D. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, Sep 2012.

Sung, N., Kim, M., Jo, H., Yang, Y., Kim, J., Lausen, L., Kim, Y., Lee, G., Kwak, D., Ha, J.-W., et al. NSML: A machine learning platform that enables you to focus on your models. *arXiv preprint arXiv:1712.05902*, 2017.

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. #exploration: A study of count-based exploration for deep reinforcement learning. In *NeurIPS*, 2017.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, 2015.

Tishby, N., Pereira, F. C. N., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. URL http://arxiv.org/abs/physics/0004057.

Vera, M., Vega, L. R., and Piantanida, P. Compression-based regularization with an application to multitask learning. *J. Sel. Topics Signal Processing*, 12(5):1063–1076, 2018.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL http://arxiv.org/abs/1708.07747.

Zhang, A., Wu, Y., and Pineau, J. Natural environment benchmarks for reinforcement learning. *arXiv preprint arXiv:1811.06032*, 2018. URL http://arxiv.org/abs/1811.06032.