LECO: Learnable Episodic Count for Task-Specific Intrinsic Reward

Anonymous Author(s)
Affiliation
Address
email

Abstract

Episodic count has been widely used to design a simple yet effective intrinsic motivation for reinforcement learning with a sparse reward. However, the use of episodic count in a high-dimensional state space as well as over a long episode time requires a thorough state compression and fast hashing, which hinders rigorous exploitation of it in such hard and complex exploration environments. Moreover, the interference from task-irrelevant observations in the episodic count may cause its intrinsic motivation to overlook task-related important changes of states, and the novelty in an episodic manner can lead to repeatedly revisit the familiar states across episodes. In order to resolve these issues, in this paper, we propose a learnable hash-based episodic count, which we name LECO, that efficiently performs as a task-specific intrinsic reward in hard exploration problems. In particular, the proposed intrinsic reward consists of the episodic novelty and the task-specific modulation where the former employs a vector quantized variational autoencoder to automatically obtain the discrete state codes for fast counting while the latter regulates the episodic novelty by learning a modulator to optimize the task-specific extrinsic reward. The proposed LECO specifically enables the automatic transition from exploration to exploitation during reinforcement learning. We experimentally show that in contrast to the previous exploration methods LECO successfully solves hard exploration problems and also scales to large state spaces through the most difficult tasks in MiniGrid and DMLab environments.

1 Introduction

In reinforcement learning (RL), tackling the problem of sparse rewards still remains a great challenge. Many works have been proposed to develop dense synthetic reward, which is generally called intrinsic reward, to provide the motivation for exploration in the absence of extrinsic rewards from the environment [5, 24, 1, 26, 13, 36, 31, 10, 38, 27, 35]. Among them, count-based exploration [29, 4, 23, 22, 31, 21, 5], that encourages the RL agent to visit novel states, is one of the most popular classes for intrinsic motivation since it is simple and effective in many tasks, and more importantly it is theoretically justified [29, 19]. In general, the state novelty can be measured within a single episode or through an entire task across episodes. In recent RL problems, environments are often changed over episodes or even procedurally generated within the same task [7, 3, 20, 8, 16]. Therefore, per-episode novelty by *episodic count* has been widely adapted in modern hard exploration problems [1, 26, 36].

However, in high-dimensional and/or continuous state spaces, state counts need to be approximated in an abstract space since most states are likely to be unique and novel in the original space, leading to insufficient exploration [4, 23, 26, 1]. In addition, computation of episodic count at every time step based on the episodic memory could be burdensome if an episode lasts for a long time and

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

therefore requires fast hashing [31]. Previous works have discretized the state space to obtain the hash code [31, 14, 9] for a fast pseudo-count, but they have not extensively applied it to recent difficult exploration problems such as procedurally generated and partially observable environments.

Moreover, existing count-based exploration methods have mostly employed task-agnostic state novelty, and thereby task-irrelevant count may disturb its intrinsic reward bonus to faithfully reflect task-related information [18, 2, 17, 38]. This problem could be severe in exploration by per-episode count. For example, task-wise familiar states across episodes can have high novelty in each episode, which confines overall exploration into smaller areas, while some states that are highly related to extrinsic rewards need to be repeatedly revisited even in a single episode for obtaining extrinsic rewards more densely.

In this work, we propose a hash-based episodic count that is learned to provide task-specific intrinsic motivation in solving hard exploration problems. The proposed **L**earnable **E**pisodic **CO**unt (LECO) produces task-specific intrinsic rewards that are efficiently and practically applicable to high-dimensional states, long-horizon episodes, procedurally generated environments, and highly sparse extrinsic rewards. In specific, the proposed intrinsic reward is composed of the episodic count-based novelty and the task-specific modulation. Our episodic count makes use of a vector quantized variational autoencoder (VQ-VAE) [32] to automatically obtain the discrete state codes for a fast counting. In addition, we exploit a learnable modulator to control the episodic novelty in relation to task-specific extrinsic rewards. The proposed LECO also adaptively shifts the RL phase from exploration to exploitation. Experimental results on the most difficult exploration tasks in procedurally generated environments of MiniGrid [7] and DMLab [3] show that LECO efficiently solves such tasks and outperforms previous state-of-the-art exploration methods.

Our main contributions can be summarized as follows:

- A novel learnable episodic count is proposed to efficiently provide task-specific intrinsic motivation to the RL agent under complex environments with sparse rewards.
- A learnable hash code based on VQ-VAE as well as a learnable reward modulator are combined to maximize extrinsic rewards during RL.
- On various hard exploration benchmark tasks, the proposed LECO demonstrates its feasibility in solving diverse tasks and consistent performance improvements over previous exploration methods.
- Several ablation experiments show the benefits from the use of learnable hash codes as well as the task-specific reward modulation.

2 Related Work

Count-based exploration. Count-based exploration was first proposed in tabular tasks [29], which promotes the agent to reduce its uncertainty by visiting unfamiliar states corresponding to low counts. However, it is basically ineffective in high-dimensional state spaces since every state would be different. Therefore, several recent works have utilized count-based exploration for high-dimensional environments. For instance, pseudo counts are derived from density models using context tree switching [4] or PixelCNN [23]. More recently, RND [5] has approximated the state visitation count by the difference between a random fixed target network and a trained predictor network. RND has also been used in other methods for measuring the state novelty [1, 36, 35]. On the other hand, a number of algorithms have transformed the high-dimensional state to the discretized hash code for fast visitation counting by a hash table [31, 10]. Similarly, we obtain the episodic count by the discretized hash code. However, we exploit learnable hash codes by VQ-VAE that is trained to approximate the state density. In addition, different from [14] that directly produces a global code by VQ-VAE, we concatenate local codes as in the original VQ-VAE for generating a single state code.

Task-specific intrinsic motivation. While count-based exploration methods have typically utilized the state novelty in a task-agnostic manner, curiosity-based methods have tried to consider task-related information for intrinsic motivation by learning the environment dynamics [24, 27, 18, 26]. They have developed intrinsic rewards based on the prediction error upon the agent's actions. Among them, recently, a number of approaches have exploited the task-dependent state novelty using the information bottleneck in order to ignore task-irrelevant distractions [18, 2, 17]. Meanwhile, there

have been a number of attempts to directly learn the intrinsic reward function to ultimately optimize the task-specific extrinsic rewards [38, 37, 34]. Their algorithms are generally performed under the meta-gradient framework. LECO also adopts the idea of learnable rewards and the gradient-based bi-level optimization [38] to construct a task-specific modulator for intrinsic rewards. However, unlike those works that do not work well with sparse extrinsic rewards, LECO takes a full advantage of intrinsic motivation by the task-agnostic state novelty based on the episodic count before obtaining the extrinsic rewards.

Exploration for procedurally generated environments. One of the important challenging RL problems is robust exploration in a procedurally generated environment where the environment (partially) changes between episodes, and hence the agent needs to explore a very large state space and have strong generalization ability. Many recent RL benchmarks deal with procedurally generated environments [7, 3, 20, 8, 16], and accordingly state-of-the-art exploration algorithms have tried to improve sample efficiency in such environments. RIDE [26] has used a change between successive states on a latent space as an intrinsic reward while AGAC [13] has encouraged action diversity by an adversarial policy. NovelD [36] has exploited regulated difference of the novelty between consecutive states, and MADE [35] has devised exploration by maximizing the deviation from previously explored regions. AMIGo [6] has proposed a teacher that generates count-based intrinsic goals. In contrast to these handcrafted or task-agnostic intrinsic rewards, LECO learns to leverage task-level information to guide the count-based exploration especially at the episode level due to the procedurally generated environment. Moreover, LECO also shows improved performances on DMLab tasks with higher-dimensional partial observations.

3 Learnable Episodic Count via Task-Specific Modulation

3.1 Notation

Our problem setting follows a typical RL problem described using Markov Decision Process defined by a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, and a transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$. An RL agent interacting with the environment samples its actions from a policy, which is an action distribution function $\pi \rightarrow \mathcal{P}(\mathcal{A})$, and receives a reward, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, upon state transition. The objective of the agent is to learn a policy that maximizes the expected sum of rewards $G_t = \mathbb{E}_\pi \left[\sum_{k=t}^{\infty} \gamma^k r_{t+k+1}\right]$, where $\gamma \in [0, 1]$ is a discount factor.

Among various sub-problems that RL faces, balancing the trade-off between exploration and exploitation still remains a fundamental problem [30]. Especially, efficient exploration is important when dealing with sparse rewards. One approach to solving such problem is to design an algorithm that directly generates exploring behavior. More specifically, a widely adapted method is to design an intrinsic motivation through a form of a reward such that the reward which the agent receives at time t is given by $r_t = r_t^e + \alpha r_t^i$, where r^e is the extrinsic reward from the environment, r^i is the intrinsic reward designed for exploration, and α is a hyperparameter for scaling [26].

3.2 LECO Intrinsic Reward

The use of state visitation count for intrinsic rewards based on the state novelty has been a core concept in several previous studies [26, 1, 36, 13, 28, 15] due to its simplicity, task-agnostic property, and theoretical basis. However, such state-novelty measured from visitation counts monotonically decreases as more experiences accumulate, equally for both important and unimportant states. Thus, an agent would be discouraged to visit even the task-important states after a certain period of time, and the exploration towards those states vanishes [1]. Furthermore, an intrinsic bonus based on the state visitation could lead the agent to only pursue any state novelty, especially in the partially-observable and sparse reward environments. For example, imagine a task in which the goal is to find a colored key hidden in a box and unlock a door that has the same color as the key. If the box changes its color to a random color every time when it is opened, the agent would repeatedly open and close the box just to obtain the intrinsic bonus. In this task, the agent should be encouraged to find the door after obtaining the key and discouraged to repeat task-irrelevant actions. We directly build upon this idea by decomposing the intrinsic reward into the state-novelty and task-specific modulation, to balance exploration and exploitation, respectively.

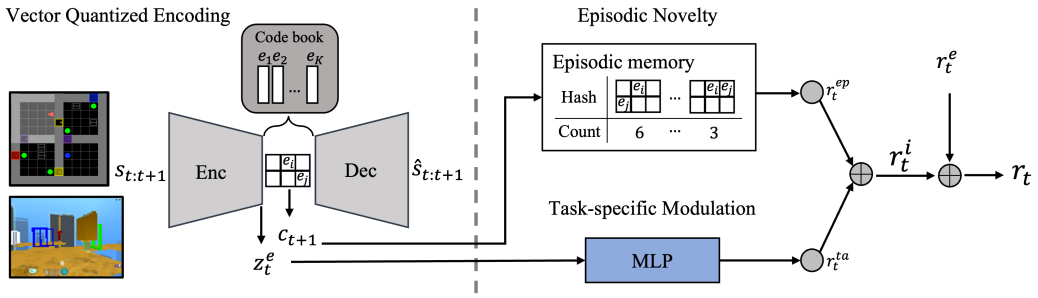


Figure 1: The schematic diagram of the intrinsic reward in LECO. The states at time t and $t+1$ are encoded and quantized to the embedding vector z_t^e and hash c_{t+1}. Then, each is fed into the task-specific modulator and episodic counter, respectively, to produce the task-specific modulation r_t^{ta} and episodic intrinsic reward r_t^{ep}, which are then combined to produce LECO's intrinsic reward r_i^t.

We design our intrinsic motivation to encourage the discovery of state novelty that is adaptively controlled to maximize the extrinsic return, hence the name **L**earnable **E**pisodic **CO**unt for task-specific motivation. Specifically, we define the general form of intrinsic reward as a sum of two parts,

$$r_t^i(a_{t-1}, s_t, a_t, s_{t+1}) = r_t^{\text{ep}}(s_{t+1}) + \lambda r_t^{\text{ta}}(a_{t-1}, s_t, a_t), \tag{1}$$

where $r_t^{\text{ep}}(s_{t+1}) = 1/\sqrt{N_{\text{ep}}(s_{t+1})}$ is the intrinsic reward from the episodic state-novelty, $N_{\text{ep}}(s)$ is the episodic visitation count of state s, and r^{ta} is the task-specific modulation with λ as its weighting factor. The first term measures the episodic novelty by state visitation count within the current episode that is initialized to zero at the beginning of each episode. The second term learns to modulate the episodic novelty in the first term according to whether the state is beneficial or harmful in achieving the task-specific objective. Note that the previous action a_{t-1} is included as the input to reward calculation to provide extra transition information. A schematic diagram of the proposed intrinsic reward can be found in Figure 1.

LECO comprises of two submodules: 1) a learnable hash that encodes input observations to discrete hash codes for episodic counts and 2) a modulator that maximizes the task-specific extrinsic rewards by adaptively scaling the episodic novelty. Here, the modulator additively scales the episodic novelty. In comparison to multiplicative scaling, the additive modulation allows the episodic novelty to be retained solely when the modulator does not give rise to enough signal due to the scarcity of extrinsic rewards.

3.3 Episodic Count via Vector Quantized Hashing

Previous researches have extended classic state visitation counts [29] to high-dimensional state spaces [4, 31]. A generalization of the classic count-based approach by locality-sensitive hashing (LSH) on the continuous embedding space was proposed [31] where the latent space is obtained by an autoencoder (AE). The use of a hash table in their approach enables counting the occurrences in a constant time, taking the advantage of scalability, especially in modern distributed RL systems [12, 11, 25]. However, such LSH does not take into account the spatial information due to its data independency [33]. As an alternative to AE, we propose to utilize vector quantized variational autoencoder (VQ-VAE) [32] that directly optimizes the discrete compressed representation of a state, producing a hash code without LSH.

In LECO, a state s_t is mapped to a hash code c_t by vector quantized hashing based on VQ-VAE comprised of an encoder and a decoder with θ_{enc} and θ_{dec} as their parameters, respectively. In specific, VQ-VAE formulates the hash function as quantization operation $z^q(s_t; \theta_e)$ to a pre-defined codebook $\theta_e \in R^{K \times D}$ where K is the size of the discrete latent space, and D is the dimensionality of each latent embedding vector e_i. The state feature $z^e(s_t)$ is mapped to the nearest embedding $z^q(s_t) = e_k$, where $k = \text{argmin}_j \|z^e(s_t) - e_j\|_2$. The spatial feature $z^e = f_{\text{enc}}(s_t; \theta_{\text{enc}})$ is obtained from the CNN encoder of VQ-VAE. The hash code is then defined as the sequence of indices $c_t = [k_i]_{i=0}^{w \times h}$, where w and h are width and height of the spatial feature, respectively.

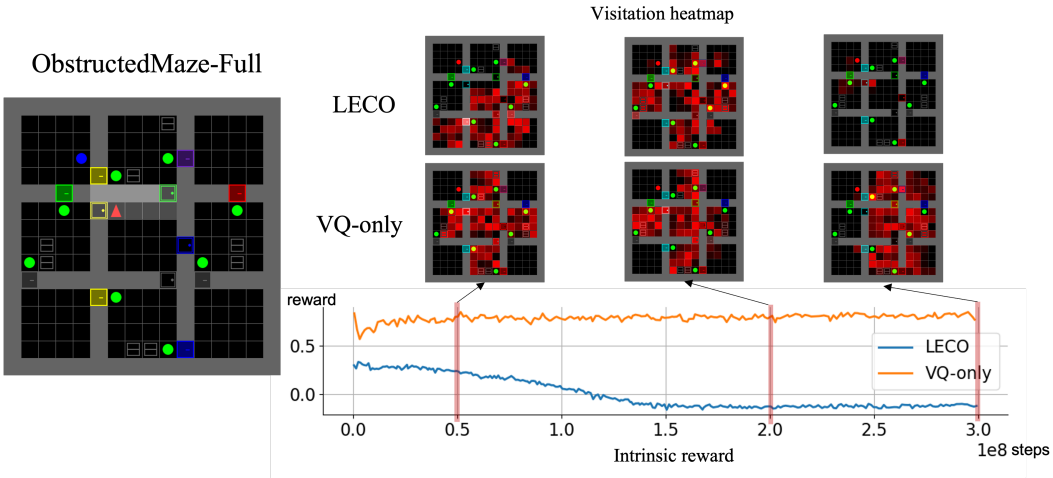


Figure 2: `ObstructedMaze-Full` Environment (**Left**). The changes in the coverage of visitations (**Right Top**) as the training progresses show that LECO modulates the episodic reward (**Right Bottom**) and so the agent focuses on the task-specific objective. On the other hand, the agent with intrinsic motivation from the episodic count only, labeled as VQ-only, remains distracted by the episodic state-novelty even in the late training phase.

Following the conventional VQ-VAE training, the hash function is updated according to the following gradient

$$\nabla_{\theta_{\text{vq}}}\mathcal{J}^{\text{vq}} = -\sum_{t=0}^{T}\nabla\Big(\log(s_t|z^q(s_t)) + \|\mathbf{sg}[z^e(s_t)] - z^q(s_t)\|_2^2 + \|z^e(s_t) - \mathbf{sg}[z^q(s_t)]\|_2^2\Big), \quad (2)$$

where $\mathbf{sg}$ stands for stopgradient operator, T is an unroll length of the agent, and $\theta_{\text{vq}} = \{\theta_{\text{enc}}, \theta_e, \theta_{\text{dec}}\}$.

3.4 Task-Specific Intrinsic Modulation

In a sparse reward environment, when an extrinsic reward signal has little or no presence, having a task-agnostic intrinsic motivation can be significantly beneficial in guiding the agent to efficiently explore the environment until the discovery of more extrinsic signals. However, a task-general intrinsic reward such as the inverse of the episodic count may become a distracting signal in the later exploitation since it is not reduced with the progression of learning. We thereby propose to include task-specific modulation that aims to ultimately maximize the extrinsic return G_t in the intrinsic reward formulation of LECO. The task-specific modulation can be seen as a dynamic factor for controlling the exploration and exploitation, based on the extrinsic rewards earned from the environment.

In order to specify the task-relevant motivation, several methods have been proposed such as reaching subgoals generated by the agent itself [6, 10]. Although those methods are capable of solving hard exploration tasks like grid maze or Montezuma's Revenge from the Atari games, generating goals in high-dimensional environments that do not provide meta-data (e.g. agent position) still remains infeasible.

On the other hand, LIRPG [38] has proposed a learning framework of parametric intrinsic rewards similar to that of the meta-gradient of policy gradient based on the extrinsic return. We leverage this framework in optimizing our modulator to maximize the extrinsic reward. In this framework, the policy parameters θ is updated using policy gradient $\nabla_\theta\mathcal{J}^{e+i}$, where $\mathcal{J}^{e+i} = \mathbb{E}_\theta[\sum_{t=0}^{\infty}\gamma^t(r_t^e + \alpha r_t^i)]$ and the intrinsic reward is constructed from a parametric model $r_t^i = r_t^{\text{ep}} + r_t^{\text{ta}}(a_{t-1}, s_t, a_t; \theta_{\text{ta}})$ with θ_{ta} as the modulator parameters. Then, the modulator is updated using the gradient $\nabla_{\theta_{\text{ta}}}\mathcal{J}^e$ where $\mathcal{J}^e = \mathbb{E}_{\theta'}[\sum_{t=0}^{\infty}\gamma^t r_t^e]$ and $\theta' = \theta + \eta\nabla_\theta\mathcal{J}^{e+i}$ is the updated policy parameters, affected by our intrinsic reward, with the step size η. Namely, the intrinsic module parameters θ_{ta} is updated by computing the gradient by the chain rule such that $\nabla_{\theta_{\text{ta}}}\mathcal{J}^e = \nabla_{\theta'}\mathcal{J}^e\nabla_{\theta_{\text{ta}}}\theta'$. This is the well-known bi-level optimization popularly used in meta-learning.

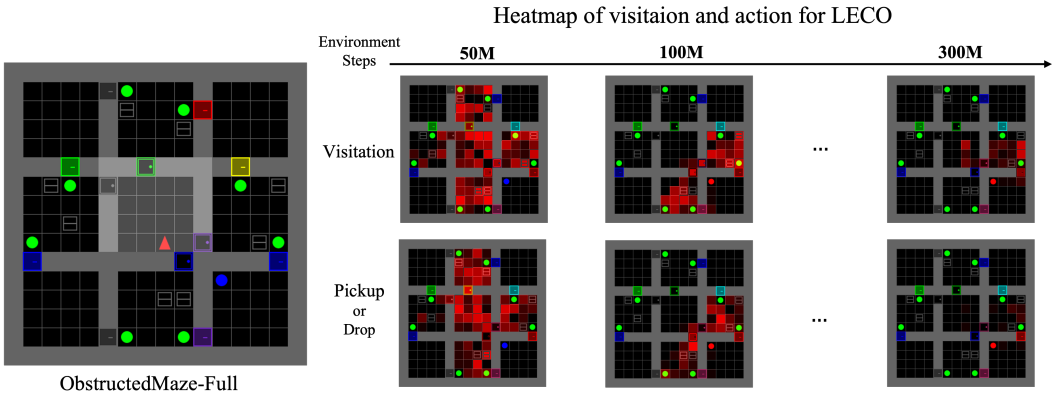


Figure 3: `ObstructedMaze-Full` Environment (**Left**), Visitation heatmaps (**Right Top**) and Action heatmaps (**Right Bottom**): LECO successfully solves this task from 100M steps, and does not repeat meaningless pickup & drop actions to obtain the state novelty at 300M steps.

Model for task-specific modulation. After executing an action a_t a task-specific modulation $r^{\text{ta}} \in (-1, 1)$ is modeled as

$$r^{\text{ta}}(a_{t-1}, s_t, a_t; \theta_{\text{ta}}) = \tanh\left(\mathbb{1}_{a_{t-1}}{}^{\top} f\big(z_t^e, a_t; \theta_{\text{ta}}\big)\right), \tag{3}$$

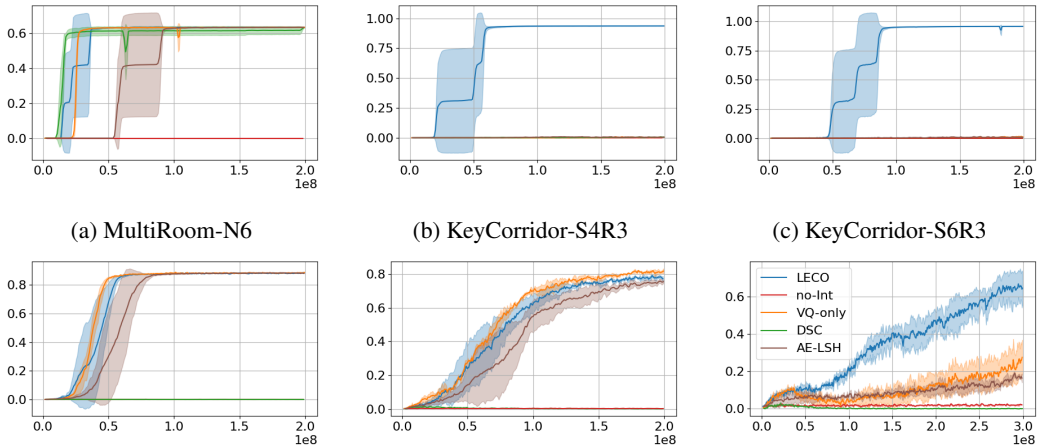
where θ_{ta} is the parameters of r^{ta}, $\mathbb{1}_a$ is the one-hot vector for action a, $f : \mathbb{R}^D \to \mathbb{R}^{|\mathcal{A}|}$ is a MLP network, and $z_t^e = z^e(s_t; \theta_{\text{enc}})$ is shared with VQ-VAE based encoding to make the modulator efficient in training.

4 Experiments

The key concept of LECO is the learnable and modulated episodic count. In other words, compared to a naive episodic count that assigns the state novelty proportional to its visitation count within an episode, LECO aims to produce low intrinsic rewards to task-irrelevant states regardless of its episodic count. We demonstrate such behavior by comparing the mean intrinsic rewards obtained by LECO to those from exploration with just the episodic count-based intrinsic reward in Figure 2. The results show that while the episodic novelty remains high, modulation in LECO decreases the intrinsic reward such that the explorative behavior has vanished towards the end of training and the agent focuses on the task-specific exploitation. Apart from the novelty based on the agent's location, in an environments where agent can pickup and drop objects, such actions can create task-irrelevant novel states, further complicating the search space. The task-specific modulation of LECO can mitigate those task-irrelevant behaviors as shown in Figure 3.

Environments. As LECO operates on multiple components, we design the experiments to thoroughly cover various aspects, especially the adaptiveness and the scalability. Since the real world as an environment is nonstationary and requires high-dimensional perception, it is crucial for an agent to efficiently explore the constantly changing environment, even at a large scale. Thus, we first demonstrate the explorative performance of LECO in procedurally generated environments of MiniGrid [7]. Then, we further extend the experiments on its performance to large scale environments of DMLab tasks.

Baselines. In order to present the contribution of each component of our proposed method, we compare LECO to its variants that incrementally add each component to a baseline without the intrinsic motivation which we denote as *no-Int* for the rest of the section. Similarly, a variant with only the vector quantization episodic counts is denoted as *VQ-only*. The no-Int baseline is an IMPALA-based agent [12], with minor architectural modifications according to the environments. Details on the base model and the architectures can be found in the supplementary material. Other baselines include *DSC* (Down Sampled Cell) [10], *AE-LSH* [31], *NovelD* [36], and *AGAC* [13], where



(a) MultiRoom-N6 (b) KeyCorridor-S4R3 (c) KeyCorridor-S6R3

(d) ObstructedMaze-1Q (e) ObstructedMaze-2Q (f) ObstructedMaze-Full

Figure 4: Comparison of LECO and its variations on six selected MiniGrid tasks. The x-axis represents total number of frames used for training and y-axis represents the score. The shaded area represents a range of standard deviation over 3 runs with different random seeds.

DSC uses a non-learnable hash for state compression and AE-LSH uses a learnable hash. NovelD and AGAC are state-of-the-art exploration methods for procedurally generated environments. Here, for a fair assessment of LECO we implement the episodic count state of NovelD and AGAC using the VQ hashing as same as LECO for all of MiniGrid and DMLab tasks.

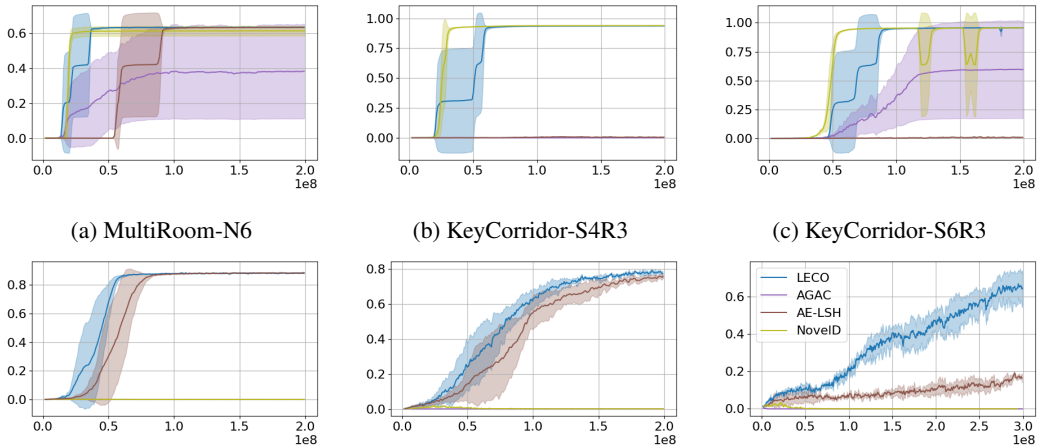
Experimental Setup. In MiniGrid, LECO was trained on A100x2 GPUs with a batch size of 768 for 18 hours. In DMLab, we used V100x8 GPUs with a batch size of 576 for 8 hours. We used the unroll length $T = 96$ for all tasks and the same LSTM-based policy networks for LECO and all baselines. The hyperparameters, model architectures, and detailed training settings are provided in the supplementary material.

4.1 MiniGrid

MiniGrid [7] is a gridworld RL environment specifically developed for testing exploration algorithms under sparse rewards, partial observations, and procedurally generated environments. A typical world in MiniGrid consists of MxN tiles, where each tile is assigned to be objects such as floor, wall, lava, door, key, ball, box, and goal. The agent deployed navigates through the world, collecting objects and solving sub-tasks such as opening a door using a key, to reach a goal placed somewhere in the world. Of the many types of environments included, we test the performance of LECO in three types of MultiRoom, KeyCorridor, and ObstructedMaze environments. More specifically, we use total of six configurations of those environments namely `MiniGrid-MultiRoom-N6`, `MiniGrid-KeyCorridorS4R3`, `MiniGrid-KeyCorridorS6R3`, `MiniGrid-ObstructedMaze-1Q`, `MiniGrid-ObstructedMaze-2Q`, and `MiniGird-ObstructedMaze-Full`, roughly ordered in increasing level of difficulty. The choices were made to reflect a wide range of difficulty including the most difficult task, while evaluating on similar tasks used by the previous baselines.

We emphasize that for all tasks the baselines and LECO only use the original partial input observations, in which a 7x7x3 tensor given as the partially-observable grid cell is the input observation for both the policy network and intrinsic module as in [6]. This setting is more challenging than the other settings such as the partial observation for the policy and full observation for the intrinsic module as in [26, 36, 28], and the partial observation for the policy and meta information such as the agent's position for the intrinsic module as in [13].

We first compare LECO to its variations as shown in Figure 4. The variations include no-Int, DSC, AE-LSH, and VQ-only. The results show that LECO solved all six tasks while VQ-only solved three, `MultiRoom-N6`, `ObstructedMaze-1Q`, `ObstructedMaze-2Q`, and showed slow



(a) MultiRoom-N6 (b) KeyCorridor-S4R3 (c) KeyCorridor-S6R3

(d) ObstructedMaze-1Q (e) ObstructedMaze-2Q (f) ObstructedMaze-Full

Figure 5: Comparison of LECO to baselines on six selected MiniGrid tasks. The x-axis represents total number of frames used for training and y-axis represents the score. The shaded area represents a range of standard deviation over 3 runs with different random seeds.

learning trend in `ObstructedMaze-Full`. On the other hand, compared to DSC and AE-LSH, a similar method of using just the episodic count as intrinsic reward but basing the quantization on heuristic discretization, VQ-only outperforms DSC as DSC only solved one out of the six tasks. While AE-LSH solved same number of tasks as VQ-only, the scores of VQ-only are higher in all tasks. As shown by the results of no-Int, a policy without any intrinsic reward is unable to solve any of the six tasks. Combining all comparisons indicates that 1) episodic count enables training in sparse reward exploration tasks with procedurally generated environments, 2) vector quantization-based counting is more effective than a heuristic cell discretization, and 3) the task-specific modulation of LECO further improves the performances of the episodic count-based approach.

Moreover, we compare the performance of LECO with other baselines of AE-LSH, AGAC, and NovelD on the same six MiniGrid tasks. As shown in Figure 5, only LECO solves all six tasks including the ObstructedMaze-Full. Here, it is noted that the original implementations of AGAC and NovelD used the agent's position and the full observations, respectively, for the episodic counts in their methods while we used the original partial observations in our implementations of AGAC and NovelD. To the best of our knowledge, LECO is the first to achieve average score of over 0.75 on the ObstructedMaze-Full with completely partial observation setting within 300M steps.

Noisy-TV on MiniGrid. Provided by [26] is a variation of MultiRoom task that brings the concept of Noisy-TV problem into the MiniGrid environment setting, named `MiniGridMiniGrid-MultiRoomNoisyTV-N7-S4`. In this version of the task, the color of the ball changes to a random color when the agent performs a specific action, creating a visual novelty irrelevant to the objective of the task, which is to reach the goal [26]. In Figure 6, we present the learning curves of LECO and baselines on this Noisy-TV environment. The result shows that all methods that use the episodic intrinsic reward, including LECO, are able to successfully solve the noisy environments.

4.2 DMLab

DMLab [3] is a learning environment consisting of various tasks, each involving different visuals and objectives in a simulated 3D world. We consider DMLab for its partial-observability that naturally arises from first-person perspective and its complexity from image-based observation with fine details and randomly generated levels to demonstrate the scalability of LECO. We compare LECO to no-Int, episodic count methods based on AE-LSH, DSC and VQ, and NovelD on two individual exploration-heavy tasks.

We select two specific tasks from DMLab, `lasertag_three_opponents_small` and `lasertag_three_opponents_large`, in order to show LECO's capability to solve explo-

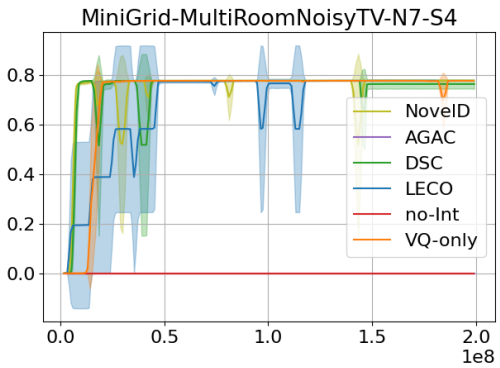
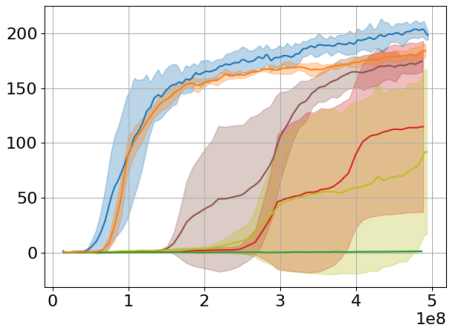
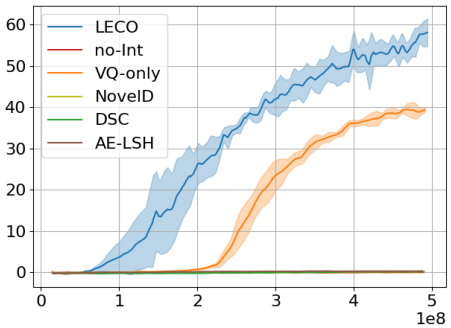


Figure 6: Performance comparison on Noisy-TV environment. The x-axis represents total frames used for training and y-axis represents the score. The shaded area represents a range of a standard deviation over 3 runs of different random seeds.



(a) lasertag_three_opponents_small



(b) lasertag_three_opponents_large

Figure 7: Learning curves on two selected DMLab tasks. The x-axis represents total number of frames used for training and y-axis represents the human normalized score. The shaded area represents the range of a standard deviation over 3 runs of random seeds.

ration tasks at a large scale. These tasks are known to be difficult to solve without advanced exploration strategy and are characterized as sparse reward problems. The `small` and `large` stand for the size of the map, which determines the reward sparsity. In Figure 7, we present the learning curves of LECO and baselines on these tasks. Each task will have its own independent agent, but the architectures and hyperparameters remain consistent across these tasks. As shown in Figure 7, LECO outperforms all of the baselines in the selected DMLab tasks.

5 Conclusion

This paper proposes a learnable episodic count to produce task-specific intrinsic rewards for hard exploration problems including high-dimensional state spaces and procedurally generated environments. In specific, the hash code based on VQ-VAE is used for computing the episodic state count and the corresponding state novelty, and this episodic novelty is regulated by the task-specific modulator that is trained to maximize the extrinsic rewards. Experimental results demonstrate that the proposed task-specific exploration based on the episodic count significantly outperforms the previous state-of-the-art exploration methods on the most difficult tasks of MiniGrid and DMLab. However, the proposed intrinsic reward also has a number of limitations in solving tasks that require intensive leveraging of the inter-episode novelty and in dealing with somewhat difficult bi-level optimization. For future works, we will extend our work to combine the inter-episode novelty and devise a more efficient training of task-specific intrinsic reward to further improve the scalability.

References

[1] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*, 2020.

[2] Chenjia Bai, Lingxiao Wang, Lei Han, Animesh Garg, Jianye Hao, Peng Liu, and Zhaoran Wang. Dynamic bottleneck for robust self-supervised exploration. In *International Conference on Neural Information Processing Systems*, 2021.

[3] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab. *CoRR*, abs/1612.03801, 2016.

[4] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 2016.

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

[6] Andres Campero, Roberta Raileanu, Heinrich Kuttler, Joshua B. Tenenbaum, Tim Rocktäschel, and Edward Grefenstette. Learning with {amig}o: Adversarially motivated intrinsic goals. In *International Conference on Learning Representations*, 2021.

[7] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[8] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International Conference on Machine Learning*, 2020.

[9] Dane Corneil, Wulfram Gerstner, and Johanni Brea. Efficient model–based deep reinforcement learning with variational state tabulation. In *International Conference on Machine Learning*, 2018.

[10] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[11] Lasse Espeholt, Raphaël Marinier, Piotr Stanczyk, Ke Wang, and Marcin Michalski. Seed rl: Scalable and efficient deep-rl with accelerated central inference. In *International Conference on Learning Representations*, 2020.

[12] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

[13] Yannis Flet-Berliac, Johan Ferret, Olivier Pietquin, Philippe Preux, and Matthieu Geist. Adversarially guided actor-critic. In *International Conference on Learning Representations*, 2021.

[14] Petros Giannakopoulos, Aggelos Pikrakis, and Yannis Cotronis. Neural discrete abstraction of high-dimensional spaces: A case study in reinforcement learning. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1517–1521, 2021.

[15] Xiao Jing, Zhenwei Zhu, Hongliang Li, Xin Pei, Yoshua Bengio, Tong Che, and Hongyong Song. Divide and explore: Multi-agent separate exploration with shared intrinsic motivations, 2022.

[16] Arthur Juliani, Ahmed Khalifa, Vincent-Pierre Berges, Jonathan Harper, Hunter Henry, Adam Crespi, Julian Togelius, and Danny Lange. Obstacle tower: A generalization challenge in vision, control, and planning. *CoRR*, abs/1902.01378, 2019.

[17] Jaekyeom Kim, Minjung Kim, Dongyeon Woo, and Gunhee Kim. Drop-bottleneck: Learning discrete compressed representation for noise-robust exploration. In *International Conference on Learning Representations*, 2021.

[18] Youngjin Kim, Wontae Nam, Hyunwoo Kim, Ji-Hoon Kim, and Gunhee Kim. Curiosity-bottleneck: Exploration by distilling task-specific novelty. In *International Conference on Machine Learning*, 2019.

[19] J. Zico Kolter and Andrew Y. Ng. Near-bayesian exploration in polynomial time. In *International Conference on Machine Learning*, 2009.

[20] Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. In *International Conference on Neural Information Processing Systems*, 2020.

[21] Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. Count-based exploration with the successor representation. In *Thirty-fourth AAAI conference on artificial intelligence*, 2020.

[22] Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. In *Twenty-sixth International Joint Conference on Artificial Intelligence*, 2017.

[23] Georg Ostrovski, Marc G. Bellemare, Aaron van den Oord, and Remi Munos. Count-based exploration with neural density models. In *International Conference on Machine Learning*, 2017.

[24] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

[25] Aleksei Petrenko, Zhehui Huang, Tushar Kumar, Gaurav Sukhatme, and Vladlen Koltun. Sample factory: Egocentric 3d control from pixels at 100000 fps with asynchronous reinforcement learning. In *International Conference on Machine Learning*, 2020.

[26] Roberta Raileanu and Tim Rocktäschel. Ride: Rewarding impact-driven exploration for procedurally-generated environments. In *International Conference on Learning Representations*, 2020.

[27] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *International Conference on Learning Representations*, 2019.

[28] Mathieu Seurin, Florian Strub, Philippe Preux, and Olivier Pietquin. Don't do what doesn't matter: Intrinsic motivation with action usefulness. In *Proc. of IJCAI 2021*, 2021.

[29] Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[30] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[31] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2753–2762, 2017.

[32] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *International Conference on Neural Information Processing Systems*, 2017.

[33] Rongkai Xia, Yan Pan, Hanjiang Lai, Cong Liu, and Shuicheng Yan. Supervised hashing for image retrieval via image representation learning. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[34] Lunjun Zhang, Bradly C. Stadie, and Jimmy Ba. Learning intrinsic rewards as a bi-level optimization problem. In *Thirty-sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

[35] Tianjun Zhang, Paria Rashidinejad, Jiantao Jiao, Yuandong Tian, Joseph E. Gonzalez, and Stuart Russell. Made: Exploration via maximizing deviation from explored regions. In *International Conference on Neural Information Processing Systems*, 2021.

[36] Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E Gonzalez, and Yuandong Tian. Noveld: A simple yet effective exploration criterion. In *Advances in Neural Information Processing Systems*, 2021.

[37] Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado van Hasselt, David Silver, and Satinder Singh. What can learned intrinsic rewards capture? In *International Conference on Machine Learning*, 2020.

[38] Zeyu Zheng, Junhyuk Oh, and Satinder Singh. On learning intrinsic rewards for policy gradient methods. In *International Conference on Neural Information Processing Systems*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We have discussed the limitations and corresponding future works in the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [No] This paper will not create any negative social impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will release the code and data upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] More details including hyperparameters are included in appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We include the standard deviation of the results with respect to multiple runs.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We used only the benchmark simulators.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] There is no personal data in our experiments.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]