랜덤 네트워크 증류에서 망각을 활용한 재탐색

안성재[O], 배성호, 황효석
경희대학교 소프트웨어융합대학
{danielahn010, shbae, hyoseok}@khu.ac.kr

Re-exploration by Leveraging Forgetting in Random Network Distillation

SungJae Ahn[O], Sung-Ho Bae, Hyoseok Hwang
College of Software, Kyung Hee University
{danielahn010, shbae, hyoseok}@khu.ac.kr

Abstract

Continual reinforcement learning requires sustained exploration to discover environmental changes. However, existing approaches either rely on reactive exploration triggered by encountered changes or require costly mechanisms for active re-exploration. We propose Forgetful Random Network Distillation (FRND), a simple approach that exploits *forgetting* in the RND predictor network to restore intrinsic rewards for long-unvisited state-action pairs. We argue that unlike forgetting in value functions, which can have unpredictable outcomes, forgetting in the RND predictor offers a clear benefit: restoring exploration bonuses. Rather than relying on incidental forgetting, FRND further guides predictor parameters toward their initial values, allowing exploration bonuses to recover without explicitly tracking recency or sampling from the state space. Our experiments show that FRND effectively captures visitation recency within network weights, promoting re-exploration and rapid adaptation to environmental changes. More broadly, our work showcases that forgetting can be harnessed as a useful mechanism for representing uncertainty about past experiences in continual reinforcement learning.

1. Introduction

Exploration in reinforcement learning (RL) guides agents to learn and adapt to their environment. Although stationarity is not an inherent assumption of RL, the field has been focused on finding optimal policies in such environments. Continual reinforcement learning (CRL) instead focuses on adaptation to non-stationary environments, where past experience can become outdated and exploration must continue. However, existing deep RL exploration approaches are misaligned with the needs of CRL, as they often reduce exploration as agents gain experience with the environment.

To maintain exploration in changing environments, reactive approaches use changes observed along the agent's current path to explore [1, 2]. However, as illustrated by the classical *shortcut problem* [3], these agents may fail to discover changes beyond their reach, making active re-exploration essential [4]. To encourage such re-exploration, prior work tracks tabular visitation recency [3] or estimates state entropy over recent experiences [5, 6]. While prioritizing regions that have not been visited recently is a natural choice for CRL, these approaches introduce complexities in high-dimensional environments. Another line of work attempts to induce optimism for long-unvisited states by adjusting learned values or model uncertainty [7, 8, 9, 10]. Still, globally adjusting state-action values or beliefs is intractable in deep RL settings.

Our starting intuition is that *catastrophic forgetting*, **when properly guided within a suitable neural network**, can efficiently represent uncertainty about regions that may have changed. While one could attempt to induce this forgetting in a value network, the outcome of forgetting is unpredictable (e.g., action values might arbitrarily increase or decrease) and selecting the "pulling" value requires assumptions about task rewards. We instead focus on intrinsic rewards, which are separate from task rewards. When intrinsic rewards are defined by prediction errors, forgetting can predictably increase these bonuses and promote *re-exploration*. However, in joint-training methods such as ICM [11], determining the direction in which to amplify prediction errors is non-trivial.

In this paper, we propose Forgetful Random Network Distillation (FRND), which utilizes forgetting in the RND predictor network [12] to restore intrinsic rewards for previously visited state-action pairs. Our main hypothesis is that a shift in the visitation distribution causes the predictor to forget previously visited state-action pairs, restoring their exploration bonuses, while ongoing learning maintains low bonuses for recently visited pairs. Rather than relying on incidental forgetting, we further guide it by regularizing the predictor parameters toward their initial values using L2 Init [13]. Since the target network remains fixed, the predictor's initial parameters define a clear direction for forgetting: moving toward them reliably restores the prediction errors that served as exploration bonuses. Our experiments show that FRND effectively captures visitation recency in network weights and promotes re-exploration, enabling adaptation to environmental changes.

2. Related Work

2.1 Random Network Distillation

RND [12] is a pseudo-count-based exploration method that uses the prediction error between a trainable predictor network and a fixed, randomly initialized target network as an intrinsic reward. As the predictor learns from visited states, their bonuses decrease, directing exploration toward unfamiliar states. While previous work has sought to mitigate forgetting in the RND predictor [14], we deliberately guide this forgetting to renew exploration bonuses.

2.2 L2 Init

L2 Init [13] is a regularization method for preserving *plasticity* of neural networks in continual supervised learning by pulling network parameters toward their initial values. We instead use it to guide *forgetting* in the RND predictor toward its initial parameters, a direction expected to restore its exploration bonuses.

2.3 Exploration in Non-stationary Environments

Reactive exploration methods [1, 2] leverage prediction errors to explore observed changes, but they often fail to discover changes

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW(2023-0-00042) supervised by the IITP(Institute of Information & Communications Technology Planning&Evaluation) in 2026.

outside their current reach. To drive agents to seek out changes, prior work attempts to induce optimism for long-unvisited regions by adjusting learned values or model uncertainty [7, 8, 9, 10]. Because globally adjusting all entries is intractable in deep RL, they resort to state sampling [10] or penalizing executed entries [9] to induce relative optimism. However, sampling states in high-dimensional spaces is impractical, and applying penalty terms to executed rewards is highly sensitive to reward scale. By leveraging forgetting in the predictor network, FRND bypasses these issues and autonomously induces optimism for long-unvisited entries. Instead of decaying learned components, one could periodically reinitialize value estimates or policies to renew exploration [15]. A common drawback of such reset approaches is their limited ability to leverage shared task structure.

3. Forgetful Random Network Distillation

FRND uses a fixed random target network f and a trainable predictor $\hat{f}_\theta$, both taking a state-action pair as input. To induce desirable forgetting, we train the predictor on currently visited state-action pairs with L2 Init [13] regularization:

$$\mathcal{L}_t(\theta) = \mathbb{E}_{(s,a)\sim\mathcal{D}_t}\left[\frac{1}{d}\|\hat{f}_\theta(s,a) - f(s,a)\|_2^2\right] + \lambda\|\theta - \theta_0\|_2^2, \quad (1)$$

where $\mathcal{D}_t$ is the current visitation data, d is the output dimension, θ_0 represents the initial predictor parameters, and λ controls the regularization strength.

The prediction error in the first term is scaled by a coefficient and serves as an intrinsic reward. We expect this bonus to promote reactive exploration after observable changes by assigning high values to novel state-action pairs, as in RND. Furthermore, as the policy converges and the visitation distribution narrows, we expect the predictor to forget state-action pairs outside the current distribution, thereby restoring exploration bonuses for them.

4. Experiments

We designed our experiments to address these three questions:

- Does forgetting in the predictor network reliably increase prediction errors, and how does guiding the network toward its initial weights strengthen this effect (see Figure 1)?
- How effectively does FRND encourage re-exploration and adaptation to environmental changes compared with existing methods (see Figure 3)?
- Do FRND's exploration bonuses reflect visitation recency, and how do they differ from those of other approaches (see Figure 4)?

4.1 Recovery of Prediction Error

To examine the first question, we alternately train an RND predictor to match its target network on MNIST digits 0 and 1 for 2,000 and 8,000 updates, respectively, and measure error recovery on digit 0. We vary the regularization strength of L2 Init and compare it against RND with increased learning rates. Figure 1 shows that stronger L2 Init regularization restores prediction errors faster and to higher levels. Notably, even with a modest regularization strength of $\lambda = 0.0001$, L2 Init achieves an asymptotic recovery level comparable to that of RND with a 100-fold larger learning rate. These results suggest that predictor forgetting can reliably restore exploration bonuses and that guiding the predictor toward its initial weights further enhances this recovery.

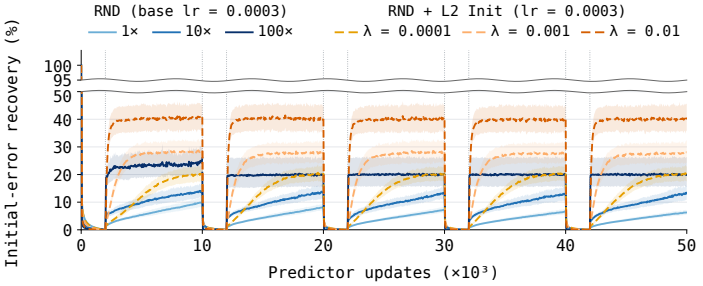


Figure 1. Prediction-error recovery on MNIST digit 0 during alternating training on digits 0 and 1. Lines and shaded bands show the mean and 95% bootstrap confidence interval over 10 seeds.

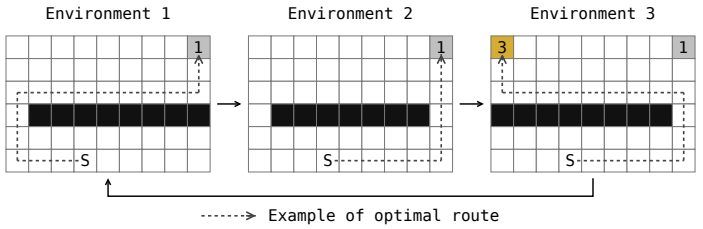


Figure 2. We extend Sutton's shortcut problem by adding Environment 3 and repeatedly cycling through Environments $1 \to 2 \to 3 \to 1$, with changes every 3,000 steps. Environment 2 opens a shortcut to the original goal with reward 1. Our added environment introduces a goal with reward 3, making the retained shortcut suboptimal and introducing non-stationarity in the reward function.

4.2 Adaptation and Re-exploration

We evaluate FRND in a 6×9 gridworld extending Sutton's *shortcut problem* [3] (see Figure 2). Since the previously optimal route remains valid after each change (except from Environment 3 to 1), agents must re-explore state-action pairs to discover better alternatives. We use Dyna-Q [3] as the common backbone for all methods, with 50 planning updates per real interaction, greedy action selection with random tie-breaking, a Q-learning rate of $\alpha = 1$, and a discount factor of $\gamma = 0.95$. We use a gridworld to clearly isolate shortcut changes and tabular Q-values to avoid policy network plasticity issues and enable comparison with Dyna-Q+. Although our experiments focus on a gridworld, we note that FRND can be directly integrated with deep RL algorithms.

We compare FRND with Q−, Q+ [3], Count, RND [12], ICM [11], RE3 [6], SsVD [10], and Reset + Count. These baselines incorporate different exploration or forgetting mechanisms into the planning process. Q+ assigns exploration bonuses proportional to the square root of the number of steps since each state-action pair was last visited. RE3 estimates novelty using nearest-neighbor distances in a random encoder's output space over recent experiences. SsVD samples states, decays all action values at those states, and pulls them toward a predefined reference value. Reset + Count combines periodic Q-value resets with a count-based exploration bonus. We reset both Q-values and visitation counts at oracle change points.

For methods that use neural networks, we one-hot encode grid positions in 54 dimensions and the four actions (UP, DOWN, LEFT, and RIGHT) in 4 dimensions. We concatenate these vectors and apply observation normalization using running statistics for FRND, RND, and RE3. We tune hyperparameters over one cycle (9,000 steps) for all methods except Q− and select the configuration with the highest mean cumulative reward across 10 tuning seeds. We then evaluate all methods for 100 cycles (900,000 steps) using 50 separate seeds.

Figure 3(a) shows that FRND achieves higher cumulative reward than all baselines except Q+, which explicitly tracks visi-

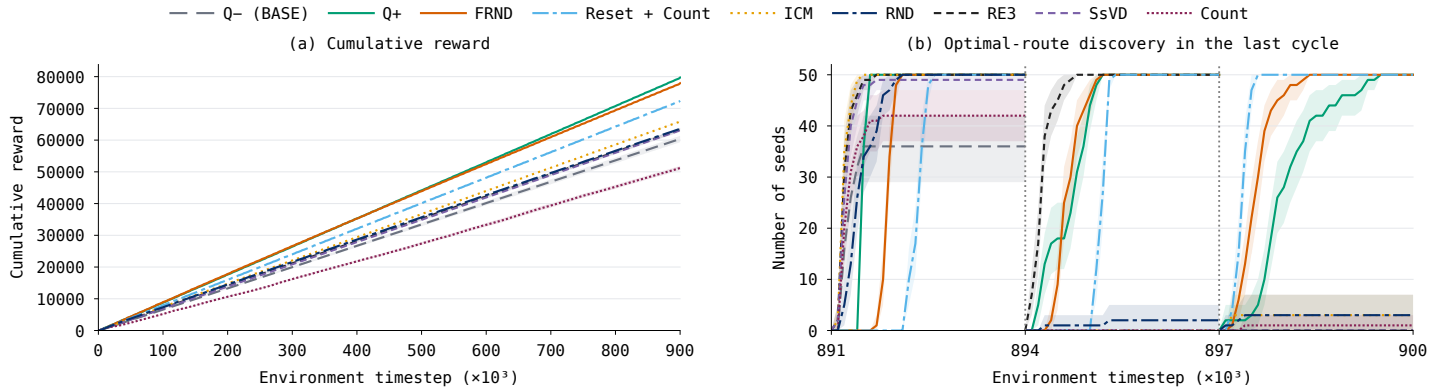


Figure 3. Cumulative reward and optimal-route discovery in the last cycle. Lines and shaded bands show the mean and 95% bootstrap confidence interval over 50 seeds.

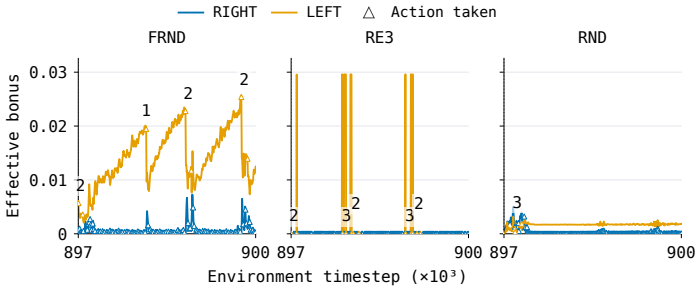


Figure 4. Exploration bonuses and actions taken at the start state during the final 3,000 steps, using each method' s highest-return seed. Numbers indicate counts of grouped LEFT actions.

tation recency. As shown in Figure 3(b), FRND rapidly discovers the optimal route after each change in the final cycle. Reset + Count also performs well but falls behind Q+ and FRND in cumulative reward. This may be because resetting previously learned components prevents the agent from leveraging knowledge of the shared task structure (e.g., UP is often better than DOWN). Notably, Count performs worse than the BASE agent, possibly because its bonuses sustain unproductive exploration even for recently visited state-action pairs. These results highlight the importance of visitation recency in balancing re-exploration and exploitation, and demonstrate that FRND can achieve this balance without explicit tracking.

To further examine whether FRND's exploration bonuses reflect visitation recency, we track the effective bonuses for RIGHT and LEFT at the start state during the final 3,000 steps. These steps correspond to Environment 3, where RIGHT is optimal, but we want the agent to regularly retry LEFT. Figure 4 shows that FRND and RE3 repeatedly restore the bonus for LEFT after it is taken, whereas RND does not. One possible explanation is that RND's predictor settles into a stable parameter region due to the shared predictor–target architecture, while FRND actively guides it toward its initial parameters. RE3 does encourage revisiting LEFT, but its bonuses exhibit sharp spikes and repeated action selections due to its buffer size (1,000) and k-NN estimation ($k = 3$). In contrast, FRND retries LEFT only once or twice at a time, followed by gradual bonus recovery reflecting visitation recency.

5. Conclusion

We introduced FRND, which leverages forgetting in the RND predictor to recover exploration bonuses and promote re-exploration in changing environments. Our experiments demonstrate FRND's effectiveness in adapting to environmental changes and the value of actively guiding forgetting. More broadly, our results suggest that forgetting can be a useful means of expressing uncertainty about past experiences in CRL.

References

[1] C. A. Steinparz et al. Reactive exploration to cope with non-stationarity in lifelong reinforcement learning. In Conference on Lifelong Learning Agents, pages 441–469. PMLR, 2022.

[2] S. Kessler et al. The effectiveness of world models for continual reinforcement learning. In Conference on Lifelong Learning Agents, pages 184–204. PMLR, 2023.

[3] R. S. Sutton. Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In Proceedings of the 7th International Conference on Machine Learning (ICML), pages 216–224, 1990.

[4] L. P. Kaelbling et al. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4:237–285, 1996.

[5] T. Zhang et al. Made: Exploration via maximizing deviation from explored regions. Advances in Neural Information Processing Systems, 34:9663–9680, 2021.

[6] Y. Seo et al. State entropy maximization with random encoders for efficient exploration. In International Conference on Machine Learning, pages 9443–9454. PMLR, 2021.

[7] P. Dayan et al. Exploration bonuses and dual control. Machine Learning, 25(1):5–22, 1996.

[8] W. C. Cheung et al. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In International Conference on Machine Learning, pages 1843–1854. PMLR, 2020.

[9] S. Dey et al. Continual optimistic initialization for value-based reinforcement learning. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, 2024.

[10] F. Störck et al. Space-sampled value decay: Forgetting mechanisms for non-stationary deep reinforcement learning. arXiv preprint arXiv:2606.11797, 2026.

[11] D. Pathak et al. Curiosity-driven exploration by self-supervised prediction. In International Conference on Machine Learning, pages 2778–2787. PMLR, 2017.

[12] Y. Burda et al. Exploration by random network distillation. In International Conference on Learning Representations, 2019.

[13] S. Kumar et al. Maintaining Plasticity in Continual Learning via Regenerative Regularization. In Proceedings of the 3rd Conference on Lifelong Learning Agents (CoLLAs), volume 274, pages 410–430, 2025.

[14] J. Hwang et al. Neuro-inspired fragmentation and recall to overcome catastrophic forgetting in curiosity. In Intrinsically-Motivated and Open-Ended Learning Workshop @NeurIPS2023, 2023.

[15] W. Mao et al. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In International Conference on Machine Learning, pages 7447–7458. PMLR, 2021.