Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Doctoral Consortium Track

Towards High-Level Intrinsic Exploration in Reinforcement Learning

Nicolas Bougie[1,2*], **Ryutaro Ichise**[2,1]
[1]The Graduate University for Advanced Studies, Sokendai
[2]National Institute of Informatics, Tokyo
{nicolas-bougie, ichise}@nii.ac.jp

Abstract

Deep reinforcement learning (DRL) methods traditionally struggle with tasks where environment rewards are sparse or delayed, which entails that exploration remains one of the key challenges of DRL. Instead of solely relying on extrinsic rewards, many state-of-the-art methods use intrinsic curiosity as exploration signal. While they hold promise of better local exploration, discovering global exploration strategies is beyond the reach of current methods. We propose a novel end-to-end intrinsic reward formulation that introduces high-level exploration in reinforcement learning. Our curiosity signal is driven by a fast reward that deals with local exploration and a slow reward that incentivizes long-time horizon exploration strategies. We formulate curiosity as the error in an agent's ability to reconstruct the observations given their contexts. Experimental results show that this high-level exploration enables our agents to outperform prior work in several Atari games.

1 Introduction

In recent years, reinforcement learning (RL) methods have led to remarkable successes in a wide variety of tasks, such as game playing [Mnih *et al.*, 2013] and robot control [Lillicrap *et al.*, 2015]. Despite their exciting results in environments with complex goals, rich sensory inputs, exploring when rewards are sparse or delayed remains an open problem in RL.

Following similar curious behaviors in animals, one solution to this problem is to let the agent generate its own intrinsic exploration bonus (i.e. curiosity-driven learning). For instance, count-based exploration [Bellemare *et al.*, 2016] keeps visit counts for states and incentivizes the exploration of novel states. Another solution [Houthooft *et al.*, 2016] relies on information gain to assess novelty. A line of work generates a bonus based on the inability to predict the future [Pathak *et al.*, 2017]; but may push the agent to seek out states with stochastic transitions. This issue has motivated several recent works [Burda *et al.*, 2018; Savinov *et al.*, 2019]. In spite of their ability to deal with local exploration - capturing the consequences of short-term decisions, they may get stuck in local optima and only detect local novelty.

In this paper, we propose a novel exploration bonus that can deal with high-level exploration. Capturing the consequences of actions on long-time horizon allows us to overcome the known "local optima" issues of prior work - when the agent cannot compensate easy immediate rewards or deceptive rewards and therefore learns a sub-optimal policy. To this end, we combine a fast reward dealing with local novelty and a slow reward that assesses global novelty. In our formulation, intrinsic rewards are based on the reconstruction errors of the observations given their *contexts*. The intuition behind this approach is that changing how are created the contexts of the fast, and slow reward models, results in flexible exploration behaviors. We benchmark our approach on a set of hard exploration tasks from Atari games, and compare our method with state-of-the-art curiosity methods.

2 Method

The proposed method (FaSo) is an effective way to incentivize high-level exploration. Our curiosity formulation decomposes the intrinsic reward r_t^e into a fast reward r_t^{fast} and a slow reward r_t^{slow}. Fast rewards assess the novelty of each state, to deal with local exploration. One the other hand, slow rewards r_t^{slow} encourage global exploration behaviors.
We sum-up our exploration bonus with the task reward: $r_t = r_t^e + r_t^i = r_t^e + [\alpha r_t^{fast} + \beta r_t^{slow}]$. We formulate intrinsic rewards as the error in an agent's ability to reconstruct observations given their contexts (context-driven curiosity).

2.1 Context-Driven Curiosity

We propose a measure of intrinsic motivation formulated as the quality of the agent to reconstruct an observation given its context. The module takes the current observation's *context* as its input and reconstructs the original image. The discrepancy between the reconstructed image and the actual image then serves as the intrinsic reward. In this work, the *context* of an observation refers to a version of it with one or more regions missing, noisy, or corrupted.

Context Creation

We introduce a simple method to extract the context of an observation based on *image downsampling*. The original image

*Contact Author

of size $w \times w$ is downscaled using a nearest-neighbor interpolation to a smaller image of size $\frac{w}{K} \times \frac{w}{K}$ and then upscaled to the original size, introducing small artifacts. The hyperparameter K controls the amount of artifacts.

Reward Calculation

We now formulate the procedure to assign the intrinsic reward. This involves the prediction error of a *reconstructor* network trained to reconstruct an observation given as input the observation's context. Formally, let s_t be the original observation at time t and s_t^* its context. The reconstructor network $R_\phi : s^* \mapsto s$ parameterized by ϕ, takes the context of an observation and reconstructs the original image s_t. This prediction will have some errors that can be measured using a distance function. We can now assign an intrinsic bonus r^{ib} that penalizes states difficult to reconstruct via:

$$r^{ib}(s_t) = ||s_t - R_\phi(s_t^*)||_2 \quad (1)$$

2.2 Fast and Slow Rewards

Although a context-driven curiosity bonus can improve local exploration, such as how to interact with a particular object; global exploration is beyond the reach of a single curiosity-based reward. Moreover, previous works [Burda *et al.*, 2018; Savinov *et al.*, 2019] found that recent curiosity models will fail to balance the loss of immediate extrinsic reward and tend to exhaust their curiosity quickly. This paper introduces a different approach where the intrinsic bonus r_t^i is the combination of two distinct context-driven rewards. They are estimated by two distinct context-driven curiosity models which reconstruct the original observation s given its fast s_f^* and slow s_s^* contexts respectively, $R_\phi^f : s_f^* \mapsto s$ and $R_\psi^s : s_s^* \mapsto s$. Thus, the overall intrinsic reward is:

$$r_t^i = \alpha||s_t - R_\phi^f(s_f^*)||_2 + \beta||s_t - R_\psi^s(s_s^*)||_2 \quad (2)$$

The key difference is how to generate s_f^* and s_s^* to achieve exploration behaviors with different ranges of time horizons. Let K_{fast} the parameter used to create the fast contexts, we typically used $K_{fast} \times 2$ or $K_{fast} \times 4$ to create the slow contexts - slow contexts are more corrupted. When large regions of images are missing or corrupted (i.e. slow contexts), similar observations of a region of the state space have nearly identical contexts. Therefore, a partial exploration of this region enables a satisfactory reconstruction. Since the agent aims to maximize this prediction error, it drives the agent to seek out more diverse and novel regions of the state space (i.e. global exploration) to notably increase the prediction errors. In contrast, contexts of fast rewards are nearly unique which entails that slightly deviating from previous policies - visiting novel states is sufficient to significantly increase the reconstruction error; encouraging to locally explore.

3 Experimental Results

We evaluated the FaSo on two difficult exploration Atari 2600 games from the Arcade Learning Environment (ALE): Montezuma's Revenge and Private Eye. The results are shown in Table 1. On Montezuma's Revenge and Private Eye, our model outperforms prior approaches that mainly deal with local exploration. It might be related to the very fact that slow rewards are large enough to encourage the agent to discover and visit new rooms. As a result, FaSo explores a larger number of rooms as compared to RND. It suggests that high-level exploration is crucial for exploring in sparse environments.

	Maximum Mean Score (at convergence)	
Method	Montezuma's Revenge	Private Eye
PPO+EC [Savinov *et al.*, 2019]	8,025	9,244
RND [Burda *et al.*, 2018]	8,152	8,666
PPO+ICM [Pathak *et al.*, 2017]	329	485
PPO+FaSo	**8,951**	**11,150**

Table 1: Final score of FaSo and baselines on Atari games. We report the results achieved over total 600M steps of training (10 seeds).

4 Conclusion

In this paper, we present a novel curiosity-based intrinsic formulation, which introduces high-level exploration to solve challenging sparse-reward problems. Our method leverages two streams of intrinsic rewards to achieve flexible exploration behaviors. We presented an evaluation on two Atari games and found that it exceeds baseline agents in terms of overall performance and convergence speed. As future work, we would like to dynamically weight local and global exploration during training. We are also willing to test our approach on different environments, as well as the method in the absence of task reward.

References

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *NIPS*, pages 1471–1479, 2016.

[Burda *et al.*, 2018] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[Houthooft *et al.*, 2016] Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *NIPS*, pages 1109–1117. 2016.

[Lillicrap *et al.*, 2015] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Ioannis Graves, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

[Savinov *et al.*, 2019] Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *ICML*, 2019.