Curiosity-driven Exploration by Self-supervised Prediction

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell

University of California, Berkeley

1. Introduction

Reinforcement learning algorithms aim at learning policies for achieving target tasks by maximizing rewards provided by the environment. However, in many real-world scenarios, rewards extrinsic to the agent are extremely sparse or missing altogether, and it is not possible to construct a shaped reward function. This is a problem as the agent receives reinforcement for updating its policy only if it succeeds in reaching a pre-specified goal state.

Motivation/curiosity [7, 10] have been used both to explain the need to explore the environment and discover goal states, but also, more generally, as a way of learning new skills which might come handy for pursuing rewards in the future. Most formulations of intrinsic reward can be grouped into two broad classes: 1) encourage the agent to explore "novel" states [1, 3, 6] or, 2) encourage the agent to perform actions that reduce the error/uncertainty in the agent's ability to predict the consequence of its own actions (i.e. the agent's knowledge about the environment) [2, 5, 8, 9, 11].

This work belongs to the broad category of methods that generate an intrinsic reward signal based on how hard it is for the agent to predict the consequences of its own actions, *i.e.* predict the next state given the current state and the executed action. However, we manage to escape most pitfalls of previous prediction approaches with the following key insight: we only predict those changes in the environment that could possibly be due to the actions of our agent or affect the agent, and ignore the rest. That is, instead of making predictions in the raw sensory space (e.g. pixels), we transform the sensory input into a feature space where only the information relevant to the action performed by the agent is represented. We learn this feature space using self-supervision – training a neural network on a proxy inverse dynamics task of predicting the agent's action given its current and next states. Since the neural network is only required to predict the action, it has no incentive to represent within its feature embedding space the factors of variation in the environment that do not affect the agent itself. We then use this feature space to train a forward dynamics model that predicts the feature representation of the next state, given the feature representation of the current state and the action. We provide the prediction error of the forward dynamics model to the agent as an intrinsic reward to encourage its curiosity.

(a) learn to explore in Level-1 (b) explore faster in Level-2

Figure 1: Discovering how to play *Super Mario Bros* **without rewards**. (a) Using only curiosity-driven exploration, the agent makes significant progress in Level-1. (b) The gained knowledge helps the agent explore subsequent levels much faster than when starting from scratch. Watch the video at http://pathak22.github.io/noreward-rl/

2. Curiosity-driven Exploration

Our agent is composed of two subsystems – a reward generator that outputs a curiosity-driven intrinsic reward signal and a policy that outputs a sequence of actions to maximize that reward signal. Let the intrinsic curiosity reward generated by the agent at time t be r_t^i and the extrinsic reward be r_t^e. The policy sub-system is trained to maximize the sum of these two rewards $r_t = r_t^i + r_t^e$. We represent the policy $\pi(s_t; \theta_P)$ by a deep neural network with parameters θ_P. Given the agent is in state s_t, it executes the action $a_t \sim \pi(s_t; \theta_P)$ sampled from the policy. θ_P is optimized to maximize the expected sum of rewards,

$$\max_{\theta_P} \mathbb{E}_{\pi(s_t;\theta_P)}[\Sigma_t r_t] \quad (1)$$

Our Intrinsinc Curiosity Module (ICM) can potentially be used with a range of policy learning methods; in the experiments here, we use asynchronous advantage actor critic policy gradient (A3C) [4]) for learning a policy. Instead of hand-designing a feature representation for every environment, our aim is to come up with a general mechanism for

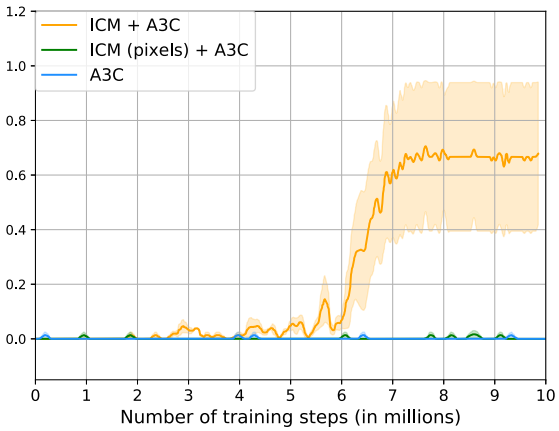

Figure 2: Comparing the performance of the vanilla A3C agent with no curiosity against the proposed curious A3C agent on a hard exploration task in VizDoom-fixed (c.f. Section 3) indicates that the curiosity based intrinsic reward signal helps solving the task.

learning feature representations such that the prediction error in the learned feature space provides a good intrinsic reward signal. We propose that such a feature space can be learned by training a deep neural network with two sub-modules: the first sub-module encodes the raw state (s_t) into a feature vector $\phi(s_t)$ and the second sub-module takes as inputs the feature encoding $\phi(s_t), \phi(s_{t+1})$ of two consequent states and predicts the action (a_t) taken by the agent to move from state s_t to s_{t+1} (i.e. the inverse dynamics model). In order to generate the curiosity reward, we train another neural network that takes as inputs a_t and $\phi(s_t)$ and predicts the feature encoding of the state at time step $t+1$ (i.e. $\hat{\phi}(s_{t+1})$). The curiosity reward, r_t^i is set to $\|\phi(s_{t+1}) - \hat{\phi}(s_{t+1})\|_2$.

3. Results

We qualitatively and quantitatively evaluate the performance of the learned policy with and without the proposed intrinsic curiosity signal in two environments, *VizDoom* and *Super Mario Bros*. Three broad settings are evaluated: a) sparse extrinsic reward on reaching a goal; b) exploration with no extrinsic reward; and c) generalization to novel scenarios. In *VizDoom*, generalization is evaluated on a novel map with novel textures, while in *Mario*, it is evaluated on subsequent game levels.

Sparse Reward: We evaluated the sparse external reward cases on VizDoom, where the agent is always spawned at a fixed room which is ~ 350 steps away from the goal under an optimal policy. A long sequence of directed actions is required to reach the goals from these rooms, making these settings hard goal directed exploration problems. Figure 2 shows that while the baseline A3C agent fails to solve the task, the curious A3C agent (ICM + A3C) is able to learn the task quickly. In other experiments [1], we show that our agent is superior to VIME [2] and robust to uncontrollable environment dynamics.

No Reward Setting: In order to test if our agent can learn a good exploration policy, we trained it on *Mario* without any rewards from the environment. We then evaluated how much progress it made (for *Mario*) in this setting. To our surprise, the no-reward Mario agent can learn to cross over 30% of Level-1[1]. The agent received no reward for killing or dodging enemies or avoiding fatal events, yet it automatically discovered these behaviors. To the best of our knowledge, this is a first demonstration where the agent learns to act with relatively complex visual imagery directly from pixels without any extrinsic rewards.

Generalization to Novel Scenarios: In Mario, we show that policies learnt by maximizing only curiosity reward on Level-1 outperform policies learnt on Level-3 on Level-3 itself. In VizDoom, we show that the ICM agent pre-trained only with curiosity on the training maps learns faster and achieves higher reward than a ICM agent trained from scratch to jointly maximize curiosity and the external rewards on the testing map.

References

[1] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *NIPS*, 2016. 1

[2] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *NIPS*, 2016. 1, 2

[3] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *NIPS*, 2012. 1

[4] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016. 1

[5] S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *NIPS*, 2015. 1

[6] P. Poupart, N. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete bayesian reinforcement learning. In *ICML*, 2006. 1

[7] E. L. Ryan, Richard; Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 2000. 1

[8] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: Proceedings of the first international conference on simulation of adaptive behavior*, 1991. 1

[9] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2010. 1

[10] P. J. Silvia. Curiosity and motivation. In *The Oxford Handbook of Human Motivation*, 2012. 1

[11] S. P. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *NIPS*, 2005. 1

[1]Full paper (ICML 2017) available at http://pathak22.github.io/noreward-rl/