Journal of Machine Learning Research 23 (2022) 1-8 Submitted 1/21; Revised 5/22; Published 9/22

A Survey on Intrinsically Motivated Curiosity Driven Reinforcement Learning

Kris Frasheri KFRASHER@UWATERLOO.CA
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, 200 University Ave W, Canada

Abstract

Reinforcement Learning (RL) has emerged throughout the past decade as a prominent method for training agents to solve complex tasks. Often these agents are trained with extrinsic rewards, however in many real-world contexts extrinsic rewards are sparse or non-existent. In cases where extrinsic rewards cannot be feasibly modelled, Curiosity Driven Learning (CDL) in RL can leverage curiosity to generate intrinsic rewards for agents to learn from. This paper introduces two components of CDL in RL, novelty-based and uncertainty-based agents. Both methods are defined and explored throughout this paper through a survey on modern, popular and emerging solutions to CDL in RL. CDL algorithms and methods are analyzed and compared, where both utility, limitations and future research are discussed. As a result, this work provides a deeper insight into intrinsically motivated agents in existing literature alongside future CDL research that would benefit from future investigations.

1. Introduction

In the past few decades reinforcement learning (RL) has emerged as a popular method for training agents to perform complex tasks. In RL, agents learn by interacting with their environment to receive rewards or punishments with the intention of training the agent policy to maximise the cumulative rewards granted. Traditionally, the rewards presented are extrinsic to the agent and specialised to the environment they are defined within. Previous works demonstrate that the success of an RL agent is attributed to the density and shape of the reward function present in their experiments. Designing a well-shaped reward function is a notoriously difficult engineering problem, and often fails to extend in industrial scenarios. In the real-world rewards extrinsic to the agent are often sparse or non-existent, making it infeasible to construct a shaped reward function. Previous research into blind RL often relies on agents receiving rewards by stumbling into a specified goal state, for instance through random exploration, failing to extend into larger and complex environments (Amin et al., 2021). An alternative to shaping extrinsic rewards is to supplement agents with dense intrinsic rewards, those being rewards generated by the agent's themselves not supplied through the environment.

The intuition behind intrinsic reward shaping is to replicate human exploration and curiosity within their environments in discovering novel states. Examples of intrinsic rewards include developing upon the notion of "curiosity" where prediction errors are substituted

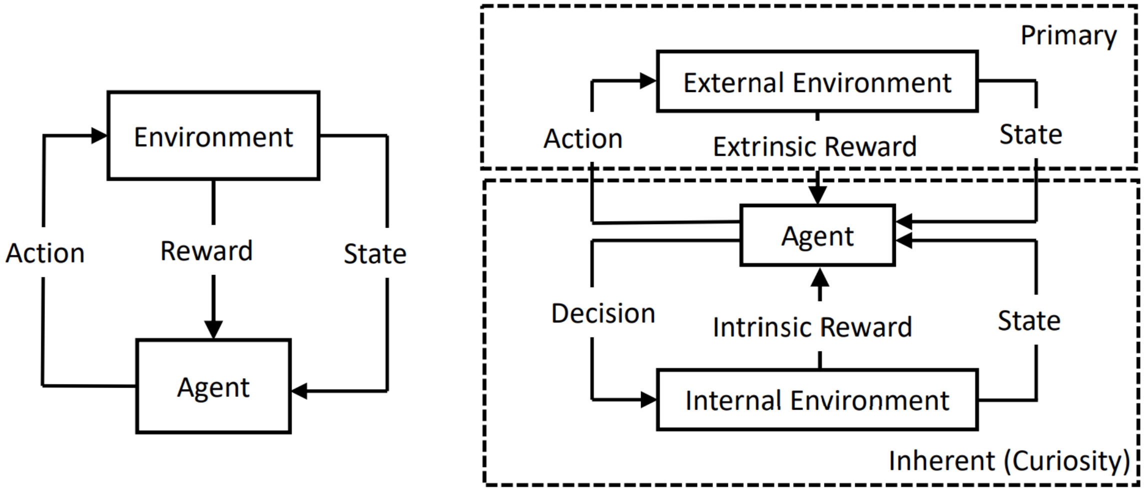

Figure 1: Agent-Environment interactions for traditional RL models (left) and Curiosity Driven Learning RL models (right)

as reward signals and “novelty” which discourages agents from re-visiting previously seen states. The author’s of Pathak et al. (2017a), introduce the ideology of Curiosity Driven Learning (CDL) in RL, where intrinsic rewards are provided to agents to encourage exploration and learning new skills to be applied in their current environment to reach new states or transferred to unseen scenarios. Through this use of intrinsic rewards, CDL has the potential to improve the quality of learning. By exploring its environment and discovering new and interesting states and transitions, the agent can learn more about the underlying mechanisms of its environment, which can lead to better predictions and control. This can be particularly useful in complex and dynamic environments, where traditional RL struggles to learn effectively. With the growing complexity of real-world problems, CDL presents a solution to the limitations of extrinsic reward based models in RL.

The purpose of this paper is to provide a comprehensive survey which reviews modern CDL algorithms and heuristics proposed in RL. Specifically, the goal is to present an exploration of existing intrinsically motivated CDL approaches in RL through identifying their utility and limitations in addressing complicated RL problems in addition to insight for future research. The survey will begin with a brief RL background into curiosity driven RL, followed by an investigation into CDL models that are designed to reward “novelty” and “uncertainty”. A comparison of the benefits and limitations of each proposed CDL model will be investigated. Finally, a conclusion of our analysis and perspective into future research will be presented.

2. Curiosity Driven Learning

In traditional RL, agents are designed to learn through interacting with their environment and maximise the corresponding extrinsic rewards granted. In Figure 1, the left diagram illustrates simply how at every time step t an agent will provide a state action pair (

s_t, a_t)to receive a reward r_t and observe the next state s_{t+1} with the transition model $P(s_{t+1}|s_t, a_t)$. The Markov Decision Process (MDP) assumes that the transition model $P(s_{t+1}|s_t, a_t)$ depends solely on the current state action pair (s_t, a_t), where agents should take the action a_t only based on their current state s_t. In the right diagram of Figure 1, the traditional RL architecture is separated into two components being the primary feedback and inherit curiosity. The reward r_t is separated into r_t^e, representing the extrinsic rewards provided by the external environment and r_t^i representing the intrinsic rewards of the internal environment. Extrinsic rewards are designed to be the primary feedback provided by the external environment whereas intrinsic rewards are the modelled "curiosity" of the agent from the internal environment. In Burda et al. (2018) the authors note that the actions and state can be discrete or continuous, with the states themselves being able to take on a multitude of complexities. For example, states can be represented as high-dimensional visual observations o_t such as pixels or states can be made compact in dimensionality through being random feature representations presented by a convolutional neural network. The representation of states will impact the complexity of the underlying RL problem, where regardless of form the goal of an agent is to learn an optimal policy π to maximise the expected rewards from their environment.

CDL techniques in RL enable researchers to address a plethora of challenging problems presented in traditional RL systems. In traditional RL problems, agents rely on extrinsic rewards r_t^e to learn from their environment, often lacking or ultimately not being designed to consider intrinsic rewards r_t^i. They have been shown to work with a high degree of efficacy when extrinsic rewards are explicitly and continuously given, however face adversity converging to an optimal π when these rewards are sparsely or are not provided. One problem in RL is that models may require a large amount of compute to aggregate trajectories $\tau_t = (s_t, a_t, r_t, s_{t+1})$ into a replay buffer potentially hindering the convergence to an optimal policy. An additional problem faced by RL today is also designing a model that is able to sample efficiently from high-dimensional observational inputs, such as pixel images (Savinov et al., 2018a). The proven ability for CDL agents to extract impactful features from high-dimensionality input with a limited emphasis on collecting trajectories is one of the main reasons for the surge in CDL methods to address RL problems.

In the following sections we will survey RL literature that incorporates a variety of designs in developing artificial curiosity. The proposed CDL agents are all designed to solely consume intrinsic rewards r_t^i, omitting any bias extrinsic rewards r_t^e to enable an investigation into the utility of these agents in real-world-like scenarios.

3. Intrinsic Motivation

The following sections will identify and analyse popular and emerging CDL based RL agents from the past decade. The authors of Pathak et al. (2017a) describe that most formulations of intrinsic rewards r_t^i can be aggregated into two broad classes. The first class is "Novelty", where rewards are designed to encourage agents to explore new and "novel" states. The second is "Uncertainty", where rewards are designed to encourage agents to take actions that minimise the error in the agent's ability to predict the consequence of their own actions. Throughout the following sections we will be investigating CDL models from both of these categories that aim to address the limitations of traditional RL problems, observing the efficacy, efficiency and short-comings of each category.

3.1 Novelty-Based Curiosity

By designing novelty-based rewards encourages agents to explore novel states and learn new behaviours within their environment. Traditionally this was done by keeping count of state and action visitations directly measuring the novelty of a state, where intrinsic rewards were designed to dissuade agents from re-visiting the same state frequently. For example, multi-armed bandit strategies such as the Upper Convergence Bound (UCB) Auer et al. (2002) and Thompson Sampling Agrawal and Goyal (2012) are examples of an agent being incentivized to choose under-explored actions to explore their environment. A limitation of this approach is that the state action pairs need to be discrete and countable, failing to scale up towards a continuous domain. To address this problem, the author Bellemare et al. (2016) proposed a count based exploration function where intrinsic rewards would be based on the pseudo-count function

$$\hat{N}_t(s) = \frac{p_t(s)(1 - p'_t(s))}{p'_t(s) - p_t(s)}, \tag{1}$$

As opposed to the number of occurrences $N_t(s)$ for a state s, where $p_t(s)$ and $p'_t(s)$ (recoding probability) are respectively defined as $p_t(s) = P(S_{t+1} = s | S_1 = s_1, \cdots, S_t = s_t)$ and $p'_t(s) = P(S_{t+2} = s | S_1 = s_1, \cdots, S_t = s_t, S_{t+1} = s)$.

The author's leveraged a reward function proportional to $\hat{N}_n(s)^{-\frac{1}{2}}$ with $r^i_n = \alpha(\hat{N}_n(s) + 0.01)^{-\frac{1}{2}}$ (where α is a scaling hyper parameter). This reward function allowed the agents to demonstrate an exploratory behaviour while allowing the authors to tune it's degree of exploration to the environment presented. In their experimentation it was discovered that the intrinsic rewards presented in their Pseudo-Count algorithm resulted in significant exploration performed in various Atari games Mnih et al. (2015). A limitation in their approach however was that the agent did not assume the action space to be continuous, with states being acquired directly from the games. In real-world scenarios the agents should be designed to learn from high-dimensional input states such as through raw pixels to extract valuable information. A constraint in their approach is the density model p may become significantly expensive should the state dimension increase drastically, presenting scalability issues in their design. Additionally, it is uncertain whether or not this approch would perform well in a partially observable environment as all evaluation was done with full observability into their environments.

To avoid the exploding density representation issues in the Pseudo-Count model, the authors Savinov et al. (2018a) proposed that novelty can be measured through reach ability. Given an encoded state e from a visual observation o, the estimated environment steps to take from the experienced states stored in a memory buffer M to the given e can be used to describe the degree of novelty for o. Thus, those observations which require substantial efforts to reach will be given high intrinsic rewards for exploration

$$r^i = \alpha(\beta + C(M, e)), \tag{2}$$

where α and β are hyper-parameters, and $C(M, e) \in [0, 1]$ represents the reachability of e. In visually dynamic and noisy scenarios, exploration based models may become subjected to the ”couch potato” issue where an agent may become overwhelmed in a state with a lot of noise (Savinov et al., 2018b). This causes agents to remain static, as the underlying model may be stuck trying to learn from the randomized noise present in the environment. The reachability-based reward function was tested in visually rich 3D environments where agents were shown to outperform other models in dynamic graphical scenarios with a plethora of noise, evading the couch potato issue. Their utilization of encoding high-dimensional input into a compact and learnable lower-level input enabled the reachability-based approach to be effective at encoding high-dimensional input as opposed to the Pseudo-Count algorithm proposed by Bellemare et al. (2016). A limitation of this method however is similar to the Pseudo-Count model where it is uncertain whether or not it may perform well with partially observability, as the model was evaluated with fully observable environments.

3.2 Uncertainty-Based Curiosity

In uncertainty based CDL, agents are designed to take actions to minimize error in their predictions of the consequences of their results. In this way, uncertainty acts as a way for agents to identify surprises from the consequences of their actions and take actions to minimize the errors in their estimations.

One major challenge faced in RL is agents who can learn from an environment with no extrinsic rewards present. Proposed by Pathak et al. (2017b), the Intrinsic Curiosity Model (ICM) incentives agents to actively explore unseen states, where it encourages agents to take actions towards states when it’s expected prediction of the consequence of it’s action are different than what actually happens. Intrinsic rewards are generated by the ICM to provide to the agent, where a feature model ϕ is utilized to extract high-dimensional input from the input state, such as pixel images. An inverse model $\hat{a}_t$ is then utilized to optimize ϕ where a forward model $\hat{\phi}$ then leveraged to obtain the agent’s prediction of the next state. The author’s jointly optimize the components of the ICM for it to extract meaningful state representations. These representations are additionally shown to be robust to noise from the input images, thus providing the intrinsic reward

$$r_t^i = \frac{\alpha}{2} \|\phi(s_{t+1}) - \hat{\phi}_{s_{t+1}}\|_2^2. \tag{3}$$

where α is a hyper parameter for controlling the degree of exploration exhibited by the agent. In their evaluation on the VizDoom and Super Mario Brothers environments the author demonstrate that ICM can sample efficiently directly when learning about their environment from the raw pixel input. The author’s further demonstrate the model’s robustness against noise in the VizDoom environment where the agent is able to successfully complete the maze with noise present in it’s input. Furthermore, this is the first CDL model observed throughout this study where it was possible for the agent to generalize learned results from a transition in environments. This was observed in the level transition in the Super Mario Bros games where the agent trained on one level was able to extend the mechanics of the game it learned to other, never before seen levels. Though it has proven an immense level of precision compared to the novelty-based CDL models we observed, one major limitation exists in ICM’s vanishing rewards problem. Throughout their experiments, the authors observed the vanishing rewards problem, where intrinsic rewards would vanish during training failing to encourage the agent to explore further. This was often observed in larger runs where the longer it took the agent to reach a goal state the more vulnerable it was to experiencing this problem. Additionally, in their results the authors observed different seeds were able to cause the agent to fail to converge or generate undesirable results in training.

Addressing the vanishing rewards problem, Shyam et al. (2018) presents a Bayesian active exploration algorithm, Model-Based eXploration (MAX), which leverages an ensemble of forward models to plan in observing novel events. The author's note that in Pathak et al. (2017b)'s work the agent exploration methods developed were reactive, where an agent learns through accidentally observing novelty in their environment to then be incentivized to continue further exploration. The vanishing rewards problem is claimed to be caused by ICM's over commitment to exploring new states, causing the novelty of unseen states to wear off over each time step. MAX is designed to engage in active exploration, where the agent seeks out novelty through their own internal estimate of action sequences that will lead to novel transitions. The MAX agent calculates the Jenson-Shannon divergence in discrete environments and the Jensen-Rényi divergence in continuous environments of the predicted space of distributions from the resulting one. The maximization of the resulting novelty measure then governs the agent's exploration incentives to pursue predicted actions sequences which may yield the most novel transition states. The authors demonstrating the superiority of MAX to other CDL models in the Half Cheetah, Ant Maze, Continuous Mountain Car and Chain tasks. Overall, MAX was able to resolve the vanishing reward problem seen in ICM by directing it's exploration heuristic to sequences of actions it determined uncertain towards. Limitations in MAX however is that the model makes the assumption of the average utility of a policy being the average utility of the probable transitions when the policy is used. Encountering a subset of those transitions and training the model can change the utility of the remaining transitions, as seen in the Chain experiments MAX was susceptible to looping between pairs of uncertain states rather than visiting other different uncertain states. Furthermore, MAX was also noted to being much more computationally demanding than the other CDL models it was compared against, as it trades off computational efficiency for data efficiency.

4. Analysis

Throughout this survey we have been able to observe the ability novel-based and uncertainty-based CDL models have in addressing challenging problems in RL. Both approaches are able to address complexity issues in extrinsically trained RL systems, being able to leverage intrinsic rewards to explore and exploit their environments. Additionally, both approaches were able to improve the efficiency of RL problems by reducing the dimensionality of input and retrieving valuable results from them in learning the dynamics of their environments. Furthermore, both approaches were able to present solutions which aided in minimizing the computational resources needed in addressing their problems. Throughout our investigation we were able to experience how uncertainty based CDL models often required more computational resources to remain effective in designing error predictability within their environments. MAX and ICM both required large amounts of computational resources to store their state action distributions in modelling uncertainty in their environments as opposed to Pseudo-Counts and Reachability. These computational trade offs however enabled uncertainty based CDL models to become more effective at addressing a wider range of applications within their environments while also being able to quickly converge. For instance, both ICM and MAX were able to perform exceptionally well with high-dimensionality input whereas Pseudo-Counts was unable to scale well with similar inputs. Furthermore, all uncertainty models were able to perform in partially observable environments, whereas novelty-based models were restricted to fully-observable models. In practice novelty based CDL approaches seem to be much lighter weight requiring less computational resources and could be used in larger scale real-world applications that may not require extreme precision of results but faster learning, such as real-time object detection. Uncertainty lends itself better to lower-scale applications where meticulous results and optimal precision are needed. For instance, automated game play testing to automate Quality Assurance tasks may benefit from using uncertainty-based CDL models as their prioritization on uncertainty may lend themselves to discover bugs in a company's game. From initial surveys, uncertainty CDL methods appear to be much more robust to noise or the couch potato problem in high-dimensionality input as opposed to novelty-based models as well.

A few open problems within CDL that currently exist regard developing CDL models that may assist in communicating and developing extrinsic rewards for environments that previously provided no reward signals. For instance, being able to develop a model that may be able to quantify and identify rewards of interest in an environment may assist researchers in identifying quality issues in their tests. Furthermore, it may enable previously extrinsic reward designed RL models to be extendable into sparse signal contexts.

5. Conclusion and Future Works

In this paper we provided an extensive survey into the applications of CDL in RL within the past decade of research. The ability to explore environments without explicitly shaped extrinsic rewards presents a major benefit to scaling CDL agents into real-world applications. Curiosity mechanisms presented additionally demonstrated how agents were able to process high-dimensional and complex inputs to efficiently explore their environments with limited compute resource visibility. Finally we note that novelty and uncertainty based CDL models are able to address similar problems, with different approaches presenting various levels of utility and limitations towards solving RL problems.

Research that may benefit from further investigation would be designing privacy bounded CDL models in RL. In both novelty and uncertainty based CDL models, agents are rewarded for their intrinsic curiosity of their environment, encouraging exploration as much as possible. In real-world applications the privacy concerns of CDL may need to be further researched and developed to prevent the development of agents which may gather sensitive information from clients. In order for CDL models to be applied within real-world applications containing sensitive information, it is critical these agents be designed with privacy in mind while being able to perform in information blocked / obfuscated contexts.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke van Hoof, and Doina Precup. A survey of exploration methods in reinforcement learning. *CoRR*, abs/2109.00157, 2021. URL https://arxiv.org/abs/2109.00157.

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

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. *CoRR*, abs/1606.01868, 2016. URL http://arxiv.org/abs/1606.01868.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos J. Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. *CoRR*, abs/1808.04355, 2018. URL http://arxiv.org/abs/1808.04355.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *CoRR*, abs/1705.05363, 2017a. URL http://arxiv.org/abs/1705.05363.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017b.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *arXiv preprint arXiv:1810.02274*, 2018a.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy P. Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. *CoRR*, abs/1810.02274, 2018b. URL http://arxiv.org/abs/1810.02274.

Pranav Shyam, Wojciech Jaskowski, and Faustino Gomez. Model-based active exploration. *CoRR*, abs/1810.12162, 2018. URL http://arxiv.org/abs/1810.12162.