A Survey on Continual Reinforcement Learning

Jinyuan Sun*
Beijing University of Posts and Telecommunications
Beijing, China
jinyuan@bupt.edu.cn

Abstract

Humans possess a remarkable capacity for continuous learning and adaptation throughout their lifetimes. This ability is often referred to as "never-ending learning," also known as continual learning or lifelong learning. Never-ending learning entails the ongoing development of increasingly complex behaviors and the acquisition of intricate skills to complement those already acquired. It involves the capacity to reapply, adapt, and generalize these abilities to novel situations. In this survey, I delve into the fundamental concept of Continual Reinforcement Learning (CRL) and offer a concise introduction to the world model and the mechanisms employed to bolster an agent's lifelong learning journey.

1 Motivation and Introduction

We aspire to develop a machine that emulates the brain's capacity for perpetual learning, enabling it to consistently acquire new skills without erasing previously learned ones[7]. In the realm of Reinforcement Learning (RL), we employ an agent to engage with unfamiliar environments using a sequence of inputs and actions. While our inspiration draws from human learning, we realize that survival alone may not be the most efficient objective, as it can demand immense computational resources and lack a holistic understanding of the world.

Conventional RL approaches tend to focus on specific tasks, often reduced to the pursuit of predefined goals using a fixed reward function. Regrettably, these methods have a tendency to overlook a wealth of information about the real world, as agents tend to disregard what they perceive as irrelevant to their immediate objectives. This approach risks the loss of valuable insights and the ability to comprehend the world in its entirety. So

2 Related Work

2.1 Reinforcement Learning

Reinforcement learning (RL) investigates the interaction between intelligent agents and their environment. It revolves around the continuous acquisition of an optimal policy, guiding a sequential decision-making process, ultimately leading to the attainment of maximum rewards[18, 8]. The process of reinforcement learning can be described through a Markov Decision Process (MDP), employing a quadruple (A, S, P, R) to represent its essential components.

Here, the action space A signifies the set of actions exerted by the agent upon the environment, the state space S denotes the collection of environmental states, and P signifies the transition probability function. The reward, denoted as R, signifies the feedback provided by the environment in response to actions taken by the agent. The policy, denoted as π, is a mapping from the state space to the action

*https://github.com/Jinyuan-S—*not* for acknowledging funding agencies.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

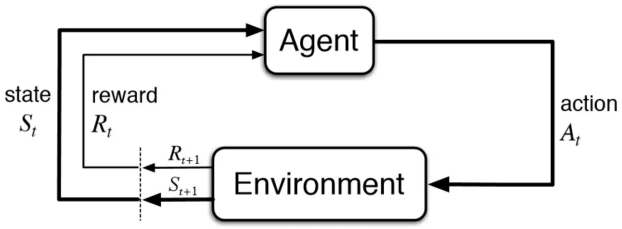

Figure 1: the interaction between agent and environment

space. As depicted in 1, the intelligent agent interacts with the environment, where A_t, S_t, and R_t represent the action, state, and reward at time t, respectively.

Typically, the configuration of states and rewards is closely tied to the specific problem at hand. The fundamental objective of reinforcement learning is to maximize the accumulation of long-term rewards. The cumulative reward is defined as a specific function of the reward sequence. Since the sum of future rewards is often infinite, a common approach is to introduce a discount factor, γ, within the range $\gamma \in [0, 1]$. This factor balances the importance of immediate rewards with future rewards. The cumulative reward from time t onward is expressed as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{\tau=0}^{+\infty} \gamma^{\tau} R_{t+\tau+1} \tag{1}$$

Based on whether the environment is explicitly modeled, reinforcement learning methods can be categorized into model-based reinforcement learning methods and model-free reinforcement learning methods.

Model-based reinforcement learning methods operate under the assumption that the Markov Decision Process (MDP) quadruple corresponding to the task is known in advance, meaning the machine has already modeled the environment. Subsequently, these methods utilize the model to plan actions or select strategies. It's worth noting that these methods are highly sensitive to the quality of the model.

Model-free reinforcement learning methods, on the other hand, do not necessitate the explicit modeling of the environment. Instead, they learn a value function or a policy function through interaction with the environment. Depending on how the agent's strategy is computed, these methods can be classified into three categories: value-based, policy-based, and Actor-Critic methods. Table 1 provides a concise comparison of classic model-free reinforcement learning algorithms.

Table 1: Comparison of Classic Mode-Free Reinforcement Learning Methods

Category	**Value-based**	**Policy-based**	**Actor-Critic**
Mechanism	Guiding the policy through the learning of values	Directly learning the policy	Simultaneously learning both the value function and the policy
RL Algorithms	Q-learning[18], Deep Q-Network(DQN)[13], Double DQN[19], Prioritized Relay DQN[15], Dueling DQN[22]	Policy Gradient/REINFORCE[23], TRPO[16], PPO[17]	Actor-Critic[10, 18], A3C[12], DDPG[11]
Advantages	High sample efficiency, low variance in value function estimation, and resistance to falling into local optima	Applicable to tasks in the current stage where actions have a profound impact on future decisions, such as in games like Go or chess	Low variance in value function estimation, high sample efficiency, and fast overall training speed of the algorithm
Drawbacks	Applicable to tasks with discrete action sets, where the optimal policy is typically deterministic	Requires a substantial amount of sample training, exhibits poor convergence, and is susceptible to converging to local optima	

2.2 Continual Learning

Although machine learning and deep learning have achieved remarkable results in many individual tasks, these models are designed for specific tasks and lack the ability to dynamically adapt to changing environments. They require a pre-prepared dataset that is shuffled randomly to approximate an independent and identically distributed (IID) data distribution before training the model within a finite timeframe. This learning approach in machine learning and deep learning falls under the category of "batch learning mode."

Conversely, human learning is a gradual process, and individuals continue to acquire new information and learn throughout their lifetimes. The acquisition of new knowledge often benefits from existing experiences while also supplementing existing knowledge. This mode of learning in the human brain falls under "continual learning mode." However, in the real world, it is impossible to obtain all the data for a given task at once, and we cannot predict the nature of future tasks and data. Therefore, conventional methods cannot transform training data into an independent and identically distributed (IID) form over the timeline. In other words, the data for each task over time is non-independent and non-identically distributed (Non-IID).

3 Modeling

We represent the agent's variables as 'z' and the inputs from the environment as 'x.' We assume that the agent is organized in a certain way to form a parameterized agent[1]. In this setup, the agent and the environment adhere to a joint distribution: $p_\phi(x, z)$. The agent's parameters are optimized to match a desired target distribution, minimizing the KL divergence between beliefs and the actual:

$$minKL\left[p_\phi(x, z) \;||\; \tau(x, z)\right] \tag{2}$$

where p refers to the actual distribution and τ refers to the target distribution.

3.1 Targets

Factorized targets represent a type of target where the inputs x and the agent variables z are treated as independent. In such scenarios, the agent's preference is to extract as little information as necessary from the environment. This leads to the agent focusing on specific tasks and solving narrow questions but may hinder its ability to continue learning.

expressive targets are a different type of target in which the inputs x and the agent variables z are highly correlated. Agents are inclined to extract more information from the environment and seek to find intersections between these variables. This approach enables agents to engage in unsupervised learning, rare learning, exploration, and control.

In world modeling RL algorithms, expressive targets are typically employed to maximize mutual information.

3.2 Information

Continual Reinforcement Learning (CRL) poses unique challenges for analysis because environments can change rapidly, and the reward function may remain unclear and changeable. To address this issue, Empowerment[9, 14] has been introduced as a metric to evaluate the difficulty of current tasks at each stage of the learning process in an algorithm-agnostic way. The Policy Information Capacity (PIC)[4] is a metric used to measure the mutual information between the episodic return R and the policy parameter Θ. It is denoted as:

$$\mathcal{I}(R; \Theta) = \mathcal{H}(R) - \mathbb{E}_{p(\theta)}\left[\mathcal{H}(R \mid \Theta = \theta)\right] \tag{3}$$

where $p(\Theta)$ refers to the prior distribution of the policy parameter (including network architecture). PIC is employed to gauge the controllability of rewards via parameter sampling, and it can be interpreted as a form of reward empowerment.

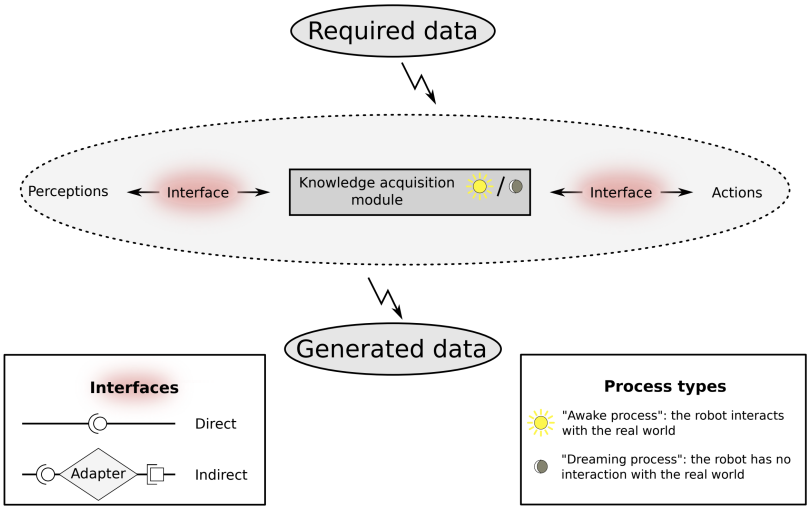

Figure 2: an example agent with a Dreaming[2] process.

4 Methods

4.1 Dreaming Based

To enable the application of planning in unknown environments, it's essential for the agent to acquire knowledge about the dynamics of the environment through interactions. Deep Planning Network (PlaNet)[6] represents a model-based agent that accomplishes this by learning environment dynamics from images and making action decisions via rapid online planning within a latent space. To achieve superior performance, a latent dynamics model with encoders and decoders is used to accurately forecast future rewards for multiple time steps.

Planning from each individual step can still be a resource-intensive process. To address this challenge, a mechanism known as "dreaming" has been introduced in Figure 2. It allows the agent to acquire knowledge about long-term behaviors by backpropagating values through imagined latent trajectories[5].

4.2 Diversity Based

By observing that intelligent creatures can explore their environments and learn useful skills without supervision, a method of creating open-ended algorithms, which generate their own never-ending stream of novel and appropriately challenging learning opportunities is proposed[21]. DIAYN ('Diversity is All You Need')[3] learning useful skills without a reward function and by maximizing an information-theoretic objective using a maximum entropy policy. The model is able to learn a skill that solves the benchmark task despite never receiving the true task reward and can provide a good parameter initialization for downstream tasks.

5 limitations and Challanges

The exploration of social learning within the framework of RL has been somewhat constrained. To improve the model's behavior and align it more closely with human-like actions, it becomes imperative for the model to predict what others think.[20] This is because human learners continually attempt to deduce the contents of other individuals' minds. Incorporating this capability can equip artificial agents with greater potency and flexibility as learners.

Agents Even young children possess intricate mental representations of others' minds, structured causally in a manner that enables them to acquire knowledge from others. In the process of designing these agents, the challenge lies in striking the right balance between inductive bias and learning.

Merely focusing on information gain is insufficient; it necessitates the formulation of a novel type of "reward" that can incentivize and guide their behaviors effectively.

Environments Human learning goes beyond mere actions, labels, or straightforward reward-predictive cues. The environments currently available for machines to learn from are overly simplistic to foster the development of genuine general intelligence. Consequently, in future endeavors, the challenge that remains is how to design and benchmark artificial agent environments that encompass a broader spectrum of data, including emotional and tonal information. These environments need to be richer, more interactive, and dynamically adaptive to facilitate more sophisticated machine learning.

References

[1] Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. *Advances in neural information processing systems*, 31, 2018.

[2] Stephane Doncieux, Nicolas Bredeche, Léni Le Goff, Benoît Girard, Alexandre Coninx, Olivier Sigaud, Mehdi Khamassi, Natalia Díaz-Rodríguez, David Filliat, Timothy Hospedales, A. Eiben, and Richard Duro. Dream architecture: a developmental approach to open-ended learning in robotics, 2020.

[3] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[4] Hiroki Furuta, Tatsuya Matsushima, Tadashi Kozuno, Yutaka Matsuo, Sergey Levine, Ofir Nachum, and Shixiang Shane Gu. Policy information capacity: Information-theoretic measure for task complexity in deep reinforcement learning. In *International Conference on Machine Learning*, pages 3541–3552. PMLR, 2021.

[5] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[6] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[7] Kole Harvey. An open-ended approach to piagetian development of adaptive behavior. *Open Access Library Journal*, 5(3):1–33, 2018.

[8] Sergey Ivanov and Alexander D'yakonov. Modern deep reinforcement learning algorithms. *arXiv preprint arXiv:1906.10025*, 2019.

[9] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

[10] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[11] 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.

[12] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[13] 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.

[14] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[15] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[16] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[19] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[20] Natalia Vélez and Hyowon Gweon. Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current opinion in behavioral sciences*, 38:110–115, 2021.

[21] Rui Wang, Joel Lehman, Aditya Rawal, Jiale Zhi, Yulun Li, Jeffrey Clune, and Kenneth Stanley. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. In *International Conference on Machine Learning*, pages 9940–9951. PMLR, 2020.

[22] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

[23] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.