A Review on Continual Reinforcement Learning

Arya Ebrahimi

Abstract
This literature review tries to offer a comprehensive overview of continual reinforcement learning, encompassing an introduction to its core concepts and an overview of current approaches. The primary goal of this review is to analyze the existing literature and categorize the findings to present a broader view, and finally, it aims to underscore the necessity of integrating diverse studies in reinforcement learning in order to develop effective continual learning agents.

1 Introduction

During the past decade, significant advancements in reinforcement learning have led to achieving superhuman performance in various tasks. However, these approaches are typically specific to particular tasks, lack the capacity to generalize, and frequently demand substantial volumes of data. They are often restricted their focus in some ways. For example, it is often supposed that a complete description of the state of the environment is available to the agent or that the interaction stream is subdivided into episodes[1]. This stands in contrast to human learning, which is continual across their lifespan and has the capability to generalize across multiple tasks. This contrast has led to research efforts aimed at bridging the gap between reinforcement learning agents and the ability to learn continually in human learning, resulting in the development of continual reinforcement learning.

Learning approaches that try to generalize across multiple tasks, like multi-task learning and meta-learning, often focus on generalizing those tasks by utilizing batches of data to train an agent. This approach differs from the human learning process, which occurs sequentially as new data is encountered without immediate access to extensive data batches (Figure 1). Additionally, a continual learning agent should acquire behaviors or skills progressively and further build upon those to develop more complex abilities hierarchically. It is also possible for an agent to invent subtasks or make use of subgoals to reach milestones, which simplifies its further learning. Moreover, a continual agent's learning process should be task-agnostic and enable the generalization to new tasks. Such agents must retain previously acquired abilities and knowledge without catastrophic forgetting and also achieve a balance between stability and plasticity.

However, few works explicitly consider continual reinforcement learning in its entirety, and many of the advances in deep reinforcement learning are yet to be fully investigated in continual settings. Thus, it is crucial to explore the ideas and recent advances of reinforcement learning for our pursuit to create continual reinforcement learning agents. This literature review presents several recent articles, introducing their main ideas and their relation to continual reinforcement learning.

It is also essential to contrast the continuing setting of reinforcement learning with continual learning. Continual learning emphasizes the ever-changing aspect of the world in which the agent needs to adapt continually to the non-stationary dynamics. Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction. While the continuing formulation can incorporate non-stationarity, the never-ending aspect of continuing tasks itself poses unsolved research questions even with the stationary dynamics.[10]

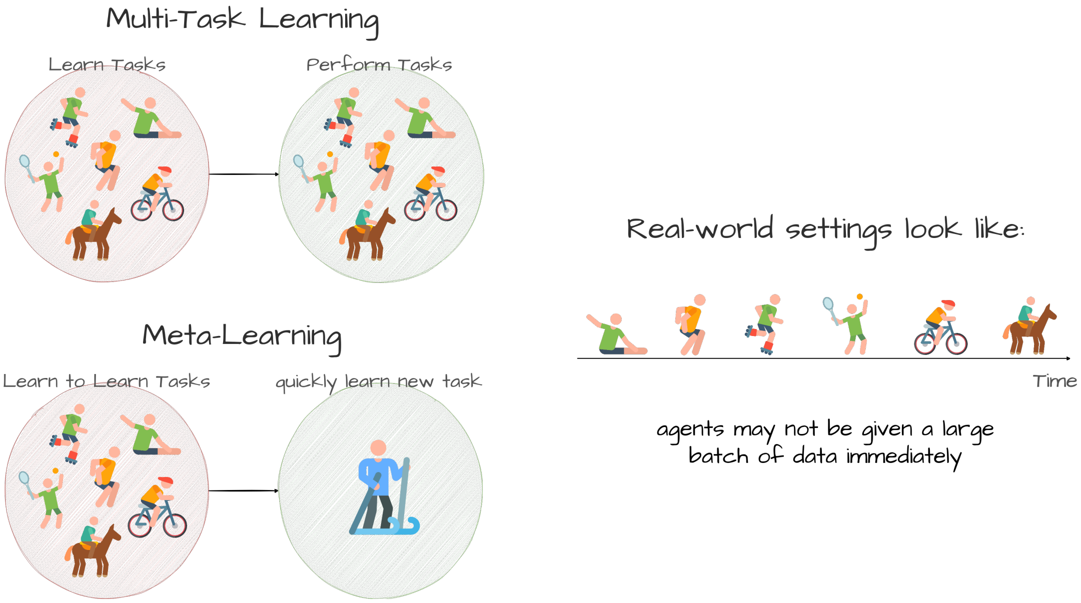

Figure 1: Difference between multi-task learning, meta-learning, and continual learning, which is inspired from Stanford CS330 course.

2 Contributing Approaches

To gain a comprehensive understanding of the ideas and methods aimed at mitigating the challenges of continual reinforcement learning, it is vital to consider recent advancements in reinforcement learning. Therefore, this section categorizes recent methods and provides a brief introduction to each, aiming to present their ideas.

Meta-learning

An essential requirement of continual RL is to acquire new capabilities in a sample efficient manner. Meta-learning is a data-driven approach to improving an agent's learning efficiency. In this setting, an agent first performs meta-training about how to learn to generalize efficiently on a distribution of tasks, and this meta-learning model is transferred to a new task in order to adapt to it quickly. In other words, meta-learning provides an inductive bias for the agent's further training that improves sample efficiency in acquiring new behaviors. For example, in MAML[3], the agent is first trained on multiple tasks to learn initial parameters, which can further be used in another training phase for fast adaptation.

Table 1: Covered methods and their scope.

Method	Meta-learning	Multi-task RL	Model-based RL	Offline RL	Reset-free
ΨΦ-learning[2]	✗	✓	✗	✓	✗
MTRF[4]	✗	✓	✗	✗	✓
VaPRL[12]	✗	✗	✗	✗	✓
ReLMM[14]	✗	✗	✗	✗	✓
OptiDICE[7]	✗	✗	✗	✓	✗
RECON[11]	✗	✗	✗	✓	✗
LiSP[9]	✗	✗	✓	✓	✓
HyperCRL[5]	✗	✗	✓	✗	✗
COMBO[16]	✗	✗	✓	✓	✗
MTSGI[13]	✓	✓	✗	✗	✗

One approach that uses meta-learning is MTSGI[13], which proposes a method that can learn a prior model of task structure from the training tasks and transfer it to the unseen tasks for fast adaptation. It suggests that tasks often consist of multiple subtasks with complex dependencies, which can be considered as subtask graphs. MTSGI infers the common task structure in terms of the subtask graph from the training tasks and uses it as a prior to improve the task inference in testing.

Multi-task RL

Multi-task learning is defined as an inductive transfer mechanism with the key objective to improve generalization performance, which is vital in continual agents. The core objective behind multi-tasking is to follow a learning-to-learn methodology to leverage the domain-related information accumulated by training the individual, related tasks in parallel with a shared representation of the system. In this way, the knowledge that is acquired during each task learning can be utilized and thereby help other tasks be learned better. Multi-task learning improves the overall generalization performance and can be applied across many domains, including reinforcement learning[15].

An approach that utilizes multi-tasking is MTRF[4], aiming to learn manipulation tasks without human interventions. The idea is that in a multi-task setting, some tasks can serve as resets for other tasks, and learning multiple tasks simultaneously enables uninterrupted continuous learning. It jointly learns K different policies π_i for each task. The agent collects a stream of data without any resets in the environment. Given the current state of the environment, a task-graph $G(s) : S \to \{0, 1, ..., K-1\}$ makes a decision once every T time steps on which of the tasks should be executed and trained for the next T time steps. This task-graph decides what order the tasks should be learned and which of the policies should be used for data collection.

Model-based RL and Planning

A continual learning agent must be able to effectively plan for the future by leveraging its acquired knowledge. To this end, several approaches are proposed to learn a model of the environment's dynamic for further planning or mitigating other problems which can be addressed using a model of the environment, such as out-of-distribution problem[16] or reset-free learning[9].

HyperCRL[5] learns a dynamic model using task-conditioned hypernetworks. It consists of a neural network that receives a task encoding as an input and outputs the weights of another neural network, which is the dynamics function and is used in the planning phase. Another approach, COMBO[16], combines offline RL with learning a model of the environment, which can mitigate the out-of-distribution problem in offline methods by generalizing beyond the offline data.

Moreover, humans acquire skills and build on them to solve increasingly complex tasks. A continual learning agent must be able to reuse previously learned skills in new, unseen situations (skill reusability). This is an important ability, especially when new skills can be created on the fly in new situations. A continual RL agent should also have the ability to compose its previous knowledge and skills to perform new ones (skill composition), which enables the agent to exploit what was learned before with greater efficacy[6].

To this end, LiSP[9] learns a set of skills and argues that planning could be a unified solution to a reset-free setting and has two main stages. First, a diverse set of low-level skills are learned in an offline manner using intrinsic rewards, and this set of skills is used for further planning by using model predictive coding (MPC). Whereas RL methods act in the environment according to a parameterized policy, model-based planning methods learn a dynamics model $p(s_{t+1}|s_t, a_t)$ to approximate the transition dynamics and use MPC to generate an action via search over the model.

Offline RL

Offline reinforcement learning[8] is another area that, because of its capability of learning from an offline dataset, has been adapted to continual reinforcement learning problems in which massive online agent-environment interactions are expensive, dangerous, or impractical.

OptiDICE[7] is an offline RL algorithm that eliminates the need to evaluate out-of-distribution actions. It estimates stationary distribution ratios that correct the difference between the data distribution and the optimal policy's stationary distribution. RECON[11] uses a visual sensor to efficiently discover and reach a target image in a previously unseen environment. It utilizes an offline dataset of previous experience transitions to learn a context-conditioned latent goal model from a pair of context and goal images. This context-conditioned model is trained to predict short-range temporal distances to go, as well as the best action towards it.

Additionally, if artificial agents are to be effective in the real world, they will need to thrive in environments populated by other agents. Humans can observe the behavior of other humans and combine the obtained observation with their experiences to quickly learn how to achieve their own goals. This objective is also desirable in continual agents to learn from other agents or even human demonstrations.

To this end, $\Psi\Phi$-learning[2] formalizes and addresses a problem setting in which an agent has access to offline observations and actions drawn from the experiences of other agents interacting with the same environment. However, it has no access to the rewards or goals of these agents, and their objectives and levels of expertise may vary widely, which is also common in real-world settings. To learn the shared features of the environment, $\Psi\Phi$-learning utilizes the successor features framework to capture the environment's dynamic, which is further used to accelerate the reinforcement learning phase.

No Reset!

The objective of lifelong reinforcement learning is to optimize agents that can continuously adapt and interact in changing environments. However, current RL approaches fail when environments are non-stationary, and interactions are non-episodic. To address this problem, some methods consider reset-free settings, in which resets to a fixed start distribution are not viable.

VaPRL[12] formulates persistent reinforcement learning and suggests that learning how to reach a goal g is easier from an initial state s close to it. Then, Knowing how to reach g from s, also makes reaching g from neighbor states of s more straightforward, facilitating incremental movement away from the goal. In other words, subgoals are defined to make learning, especially in sparse reward settings, easier. A curriculum is created, which starts from a state close to the goal and progressively moves towards the initial state distribution. The curriculum $C(g)$ results in the closest state to the initial state distribution such that $V^{\pi}(s, g) \geq \epsilon$. At the beginning of learning, the policy is insufficient, so $C(g)$ selects the states near the goal. As the policy improves, more states satisfy the $V^{\pi}(s, g)$ constraint, so $C(g)$ will select the states closer to the initial state distribution.

Another reset-free approach, ReLMM[14], disentangles learning grasping policy from navigation policy in a mobile manipulation task. It only uses successful grasp rewards for training both policies and does not rely on complex sensory inputs but only a first-person image of the area in front of the robot. In the first stage, the grasping policy is trained ensemble, and after a successful grasp of a ball, the agent (robot) randomly puts it on the ground for further training. At the next stage, the navigation and grasp policies are trained concurrently to learn how to navigate to a ball and then grasp it. After a successful grasp, again, the agent moves to a random location and puts the ball on the ground for further training. The pseudo-reset behavior introduced in this approach minimizes the need for human interventions.

3 Conclusion

This literature review presented recent approaches that address the issues associated with continual reinforcement learning and introduced their ideas. In the future, the review will expand its scope to cover a broader range of methods.

References

[1] David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *arXiv preprint arXiv:2307.11046*, 2023.

[2] Angelos Filos, Clare Lyle, Yarin Gal, Sergey Levine, Natasha Jaques, and Gregory Farquhar. Psiphi-learning: Reinforcement learning with demonstrations using successor features and inverse temporal difference learning. In *International Conference on Machine Learning*, pages 3305–3317. PMLR, 2021.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[4] Abhishek Gupta, Justin Yu, Tony Z Zhao, Vikash Kumar, Aaron Rovinsky, Kelvin Xu, Thomas Devlin, and Sergey Levine. Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6664–6671. IEEE, 2021.

[5] Yizhou Huang, Kevin Xie, Homanga Bharadhwaj, and Florian Shkurti. Continual model-based reinforcement learning with hypernetworks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 799–805. IEEE, 2021.

[6] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

[7] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

[8] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[9] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Reset-free lifelong learning with skill-space planning. *arXiv preprint arXiv:2012.03548*, 2020.

[10] Abhishek Naik, Zaheer Abbas, Adam White, and Richard Sutton. Towards reinforcement learning in the continuing setting. 2021.

[11] Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

[12] Archit Sharma, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Autonomous reinforcement learning via subgoal curricula. *Advances in Neural Information Processing Systems*, 34:18474–18486, 2021.

[13] Sungryull Sohn, Hyunjae Woo, Jongwook Choi, Lyubing Qiang, Izzeddin Gur, Aleksandra Faust, and Honglak Lee. Fast inference and transfer of compositional task structures for few-shot task generalization. In *Uncertainty in Artificial Intelligence*, pages 1857–1865. PMLR, 2022.

[14] Charles Sun, J drzej Orbik, Coline Manon Devin, Brian H Yang, Abhishek Gupta, Glen Berseth, and Sergey Levine. Fully autonomous real-world reinforcement learning with applications to mobile manipulation. In *Conference on Robot Learning*, pages 308–319. PMLR, 2022.

[15] Nelson Varghese and Qusay Mahmoud. A survey of multi-task deep reinforcement learning. *Electronics*, 9:1363, 08 2020.

[16] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.