On-Policy Algorithms for Continual Reinforcement Learning

Anonymous submission

Abstract

Continual reinforcement learning (CRL) is the study of optimal strategies for maximizing rewards in sequential environments that change over time. This is particularly crucial in domains such as robotics, where the operational environment is inherently dynamic and subject to continual change. Nevertheless, research in this area has thus far concentrated on off-policy algorithms with replay buffers that are capable of amortizing the impact of distribution shifts. Such an approach is not feasible with on-policy reinforcement learning algorithms that learn solely from the data obtained from the current policy. In this paper, we examine the performance of proximal policy optimization (PPO), a prevalent on-policy reinforcement learning (RL) algorithm, in a classical CRL benchmark. Our findings suggest that the current methods are suboptimal in terms of average performance. Nevertheless, they demonstrate encouraging competitive outcomes with respect to forward transfer and forgetting metrics. This highlights the need for further research into continual on-policy reinforcement learning.

Introduction The assumption of numerous machine learning algorithms that data follows an independent and identically distributed (i.i.d.) pattern is frequently invalid, given that the real world is in a constant state of flux. Such a shift may occur in the observed data distribution during the training phase, yet it is not accounted for by the aforementioned algorithms. Continual learning (CL) represents a field of study that aims to address these issues. However, its application for solving reinforcement learning (RL) tasks presents considerably greater challenges than those encountered in the context of simple classification problems. The Continual World robotic benchmark, as proposed by Wołczyk et al. (2021), is a specific tool designed for the evaluation of RL agents in CL environments. The primary limitation of this tool is that it relies solely on the off-policy soft actor-critic (SAC) algorithm (Haarnoja et al. 2018). The aim of this study is to investigate whether existing approaches to continual learning can be effectively combined with proximal policy optimization (PPO), a widely used on-policy algorithm (Schulman et al. 2017). Our findings indicate that, while the use of PPO in lieu of SAC typically results in markedly inferior average performance, this is not the case with regard to forward transfer and forgetting metrics. We suspect that this phenomenon may be attributed to the exploitation of off-policy algorithms from a replay buffer, which provides access to data generated by any policy. The obtained results suggest the necessity for further investigation into on-policy continual reinforcement learning (CRL). In subsequent work, we intend to develop CL methods that will address the identified shortcomings.

RL and CRL in Brief Reinforcement learning is a framework for modeling and solving decision-making problems where an agent interacts with a dynamic environment to maximize a cumulative reward over time. In contrast, continual reinforcement learning is a domain where the agent encounters a sequence of tasks (with different environments) rather than a single, isolated one. This setup presents two principal challenges: catastrophic forgetting, whereby the agent loses the ability to perform previous tasks after learning new ones, and transfer learning, which entails applying knowledge from one task to enhance learning in related new tasks. Accordingly, in lieu of an average *performance* (success rate) across all tasks, a more comprehensive evaluation of CRL necessitates the utilization of sophisticated metrics (see Wołczyk et al. (2021)) such as average *forward transfer* (normalized area between training curves of the task in a sequence and detached) and average *forgetting* (difference between success rate on the task at the end of its training and at the end of the entire learning process).

Experiments We perform experiments to compare PPO and SAC in terms of catastrophic forgetting and forward transfer, using a setup similar to the Continual World benchmark (Wołczyk et al. 2021). For this comparison, we create a sequence of $N = 5$ tasks that the agent learns sequentially, without resetting the network parameters when transitioning between tasks. Although each task is evaluated throughout the learning process, it is only trained for $\Delta = 10^6$ steps during its specific interval. We apply PPO and SAC with simple fine-tuning and three different CL methods: L2 regularization, elastic weight consolidation (EWC) (Kirkpatrick et al. 2017), and PackNet (Mallya and Lazebnik 2018). The results of the experiments are shown in Fig. 1 and Tab. 1.

While all CL methods address forgetting in PPO, it is observed that the average performance after training remains markedly inferior to that of SAC. Furthermore, it is noteworthy that PackNet with SAC demonstrates no signs of forgetting, whereas PackNet with PPO displays some degree of

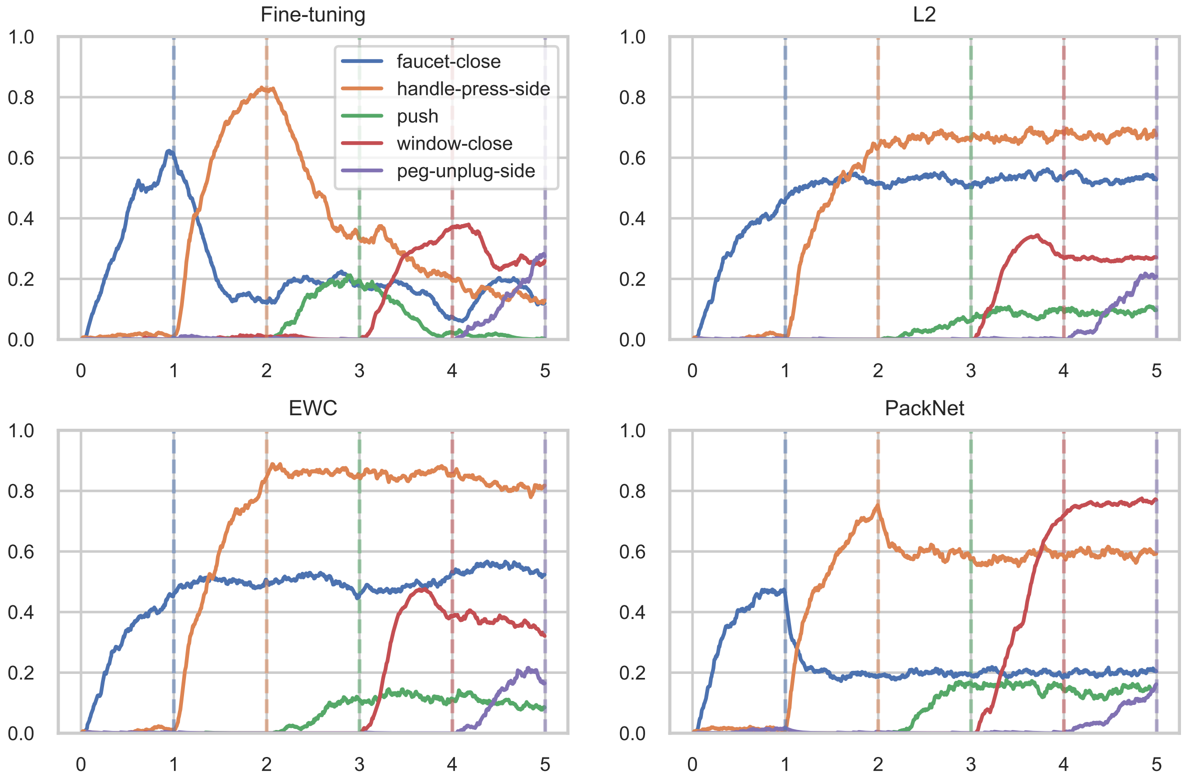

Figure 1: Comparison of the PPO success rate between fine-tuning, L2 regularization, EWC, and PackNet methods, obtained on a sequence of $N = 5$ tasks. Each task is trained on its specific interval for 1 million steps, and its success rate is evaluated throughout the learning process. The vertical dashed lines represent the points where training transitions from one task to the next, indicating the start of a new task.

Method	Performance		Transfer		Forgetting	
	PPO	SAC	PPO	SAC	PPO	SAC
Fine-tuning	0.164	**0.212**	-0.007	**0.458**	**0.271**	0.764
L2	0.348	**0.639**	**-0.138**	-0.886	**-0.016**	0.072
EWC	0.375	**0.817**	**-0.050**	-0.127	**0.015**	0.051
PackNet	0.387	**0.842**	**-0.053**	-0.735	0.073	**-0.011**

Table 1: CLR metric values for the sequence of five tasks. All results are averaged across 20 random seeds.

forgetting. It is postulated that this phenomenon occurs due to the utilisation of the replay buffer for retraining subsequent to the pruning phase, and the PPO training objective is not optimally aligned with the undertaking of multiple gradient steps on the same data. Finally, it was observed that fine-tuning with PPO did not exhibit any forward transfer, indicating that the knowledge gained from previous tasks was not beneficial when training on the current task. Conversely, higher forward transfer was observed in PPO than in SAC when CL methods were employed.

Conclusions The results obtained thus far provide preliminary evidence that on-policy RL algorithms, such as PPO, exhibit distinct behaviors in CL setups compared to off-policy ones. In future work, we aim to identify the precise reasons for these discrepancies and propose CL methods specifically tailored for on-policy algorithms.

References

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Mallya, A.; and Lazebnik, S. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7765–7773.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Wołczyk, M.; Zajac, M.; Pascanu, R.; Łukasz Kuciński; and Miłoś, P. 2021. Continual World: A Robotic Benchmark For Continual Reinforcement Learning. arXiv:2105.10919.