8/10

Evaluating Current Infrastructures for Safe Reinforcement Learning

Kiran Sthanusubramonian
ksthanus@asu.edu
School of Computing & Augmented Intelligence
Arizona State University

Introduction & Problem Statement

Reinforcement Learning (RL) is a powerful class of Machine Learning (ML) algorithms that are useful in various fields such as robotics, game-playing (arcade games, Chess, Go, etc.), and autonomous driving. More recently, Reinforcement Learning proved to be a critical component in building the current world-renowned chatbot in ChatGPT, which uses Reinforcement Learning from Human Feedback models (RLHF) (Chip Huyen 2023).

what will you do with this?

However, applying RL in safety-critical raises questions about the potential side-effects and unintended consequences of these algorithms. This leads to the requirement of open-source and well-maintained infrastructures designed to test new RL algorithms in different safety settings. In 2019, OpenAI introduced a safety starter agent that implements a few classical RL algorithms for benchmarking safety standards in deep reinforcement learning (Ray, Achiam, and Amodei 2019). However, this project was not maintained and built on Tensorflow version 1, which has since been deprecated.

This project will implement and test two newly proposed frameworks that can potentially be used for a wide setting of Safe Reinforcement Learning experiments. The primary motivating publication for this project is implementing benchmarks for Offline Safe Reinforcement Learning proposed in (Liu et al. 2023). This publication covers two major aspects of safe offline Reinforcement Learning: constructing datasets for offline safe RL (in the same vein as D4RL) and tailoring a benchmarking platform for offline RL algorithms on these datasets. As opposed to Markov Decision Processes (MDPs) used in conventional Reinforcement Learning settings, safe Reinforcement Learning is built on Constrained Markov Decision Processes (CMDPs). This forms the basis for the creation of safe RL datasets. This project will primarily focus on implementing and benchmarking Q-Learning-based algorithms to be tested on safe RL datasets. The entirety of this publication is built on top of the OpenAI Gymnasium used for Reinforcement Learning (Brockman et al. 2016).

The second framework/infrastructure of focus is OmniSafe (Jiaming Ji 2023). This infrastructure provides a more comprehensive architecture for Safe Reinforcement Learning, focusing on providing a highly modular interface with high-performance parallel computing capabilities. Furthermore, this infrastructure implements a broader array of algorithm types, including On-Policy and Model-Based RL algorithms, as opposed to just Offline Algorithms presented in our first publication of focus. Deep-diving into the implementation details of both these publications will provide a clear understanding of the current state of safe Reinforcement Learning infrastructures present for research.

Consider what that means... No SIM

Proposed Design & Implementation Plan

Stage-wise Development

The implementation plan for this project will be divided into the following stages:

- **Stage 1:** The first stage of the project will involve a thorough background study of classical Reinforcement Learning algorithms that are implemented in the publications mentioned. Particular focus will be emphasised on model-free and offline Reinforcement Learning algorithms, the primary theme behind the main infrastructures referenced for this project.
- **Stage 2:** The second stage of the project will involve setup and basic experimentation with the codebases of each of the two infrastructures.
- **Stage 3:** The final stage of this project will involve reconciling the various benchmarks proposed between the two infrastructures and evaluating the key differences and shortcomings of the infrastructures based on these benchmarks.
- **Stretch Objective:** Implement a Q-learning-based algorithm not provided in the two infrastructure implementations: Deep Reinforcement Learning with Double Q-Learning (van Hasselt, Guez, and Silver 2015).

Environments & RL Algorithms

The primary environment class for this project will be the **Safety Gymnasium**, an environment class built on the MuJoCo Physics Simulator built for Python with standard and velocity-based environments. This environment is chosen as there is a direct overlap of usage between the two infrastructures of primary reference. The class of RL algorithms to be experimented with will be Q-learning-based algorithms such as Batch-constrained deep Q-Learning (BCQ), Constrained-penalized Q-learning (CPQ), and Critic-Regularized Regression (CCR).

Evaluation Metrics

The primary evaluation metrics are the same as the one proposed in our first publication of reference (Liu et al. 2023). These are the normalized Reward return and normalized Cost return defined as follows:

$$R_{normalized} = \frac{R_\pi + r_{min}}{r_{max} - r_{min}} \times 100 \tag{1}$$

$$C_{normalized} = \frac{C_\pi + \epsilon}{\kappa + \epsilon} \tag{2}$$

where R_π and C_π are the evaluated reward return and evaluated cost return, respectively, of policy π, r_{max} and r_{min} as the maximum and minimum empirical reward return for a particular RL task. In our proposed methodologies, the versatility of our algorithms is tested using Constraint Variation Evaluation, which is achieved using distinct target thresholds defined as κ.

Contributions & Learning Outcomes

This is a solo project. Hence, each aspect of this project will be implemented by myself. I wanted to do this project to thoroughly study Reinforcement Learning and its intricacies in practical implementation scenarios. Learning more about end-to-end RL infrastructures, from Dataset Preparation and Preprocessing to Algorithm Development, will hold me in good stead in my goal of becoming a well-rounded and reliable Machine Learning Engineer.

References

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym. .

Chip Huyen. 2023. RLHF: Reinforcement Learning from Human Feedback. https://huyenchip.com/2023/05/02/rlhf.html. Accessed: 2023-05-02.

Jiaming Ji, B. Z. J. D. X. P. R. S. W. H. Y. G. M. L. Y. Y., Jiayi Zhou. 2023. OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research. *arXiv preprint arXiv:2305.09304*.

Liu, Z.; Guo, Z.; Lin, H.; Yao, Y.; Zhu, J.; Cen, Z.; Hu, H.; Yu, W.; Zhang, T.; Tan, J.; and Zhao, D. 2023. Datasets and Benchmarks for Offline Safe Reinforcement Learning. arXiv:2306.09303.

Ray, A.; Achiam, J.; and Amodei, D. 2019. Benchmarking Safe Exploration in Deep Reinforcement Learning.

van Hasselt, H.; Guez, A.; and Silver, D. 2015. Deep Reinforcement Learning with Double Q-learning. arXiv:1509.06461.