19th International Conference on Innovative Computing, Information and Control (ICICIC2025)
August 26-29, 2025, Kitakyushu, Japan

PS-28: ICICIC2025-202

Causality-Aware Exploration and Model-Based Exploitation for Sample-Efficient Reinforcement Learning

Zhuoya Zhao[1] and Peng Shi[1,2,*]

[1]School of Electrical and Mechanical Engineering, The University of Adelaide, SA 5005, Australia
[2]National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, P. R. China
*Corresponding author: peng.shi@adelaide.edu.au

Introduction

- Reinforcement learning (RL) enables robotic agents to learn policy through trial-and-error interactions with environment.
- RL often struggles with poor sample efficiency, particularly in the environments with sparse rewards.
- Acquiring the massive number of samples needed is often impractical when interactions are costly, particularly in real-world settings.
- RL consists of exploration and exploitation phases that can be optimised for better sample efficiency.
- Incorporating prior knowledge, such as causal awareness and transition dynamics model, can significantly improve the sample efficiency.

Research Questions

- How to develop a novel RL framework to improve sample efficiency in sparse-reward environments by addressing both exploration and exploitation challenges?
- How to incorporate causal awareness into the framework to guide new sample generation for more effective exploration?
- How to leverage transition dynamic model to reinterpret collected samples for exploitation of existing samples?

Methodologies

Counterfactual Data Augmentation (CDA)

- As a form of implicit exploration, it generates synthetic counterfactual samples and thus increases the diversity of samples.
- It is a causality-driven approach that focuses on causally independent relationships.
- Counterfactual samples are synthesised by intervening in tasks-independent variables at a specific state.

Model-based Hindsight Experience Replay (MHER)

- Hindsight experience replay (HER) reinterprets failures as success by relabelling goal with virtual goal that is achieved in the trajectory.
- Based on HER, MHER uses a learned transition dynamics model to predict plausible virtual goals.

Proposed Framework

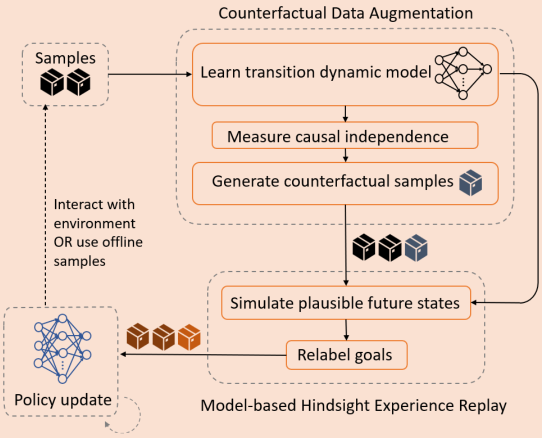

Figure 1. Proposed framework

Preliminary Results

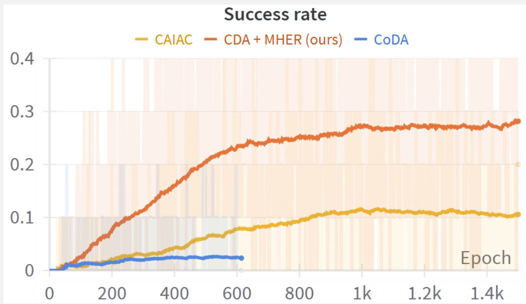

Figure 2. Comparison of the success rate of our framework with CAIAC and CoDA. The simulations are conducted in a low data regime (the dataset only contains 4k episodes).

Conclusion

- We propose an RL framework to improve efficient exploration and exploitation of samples in sparse-reward environments.
- The framework integrates CDA and MHER to improve sample efficiency.
- This approach generates causally valid synthetic samples and efficiently relabels goals.
- Our preliminary experimental results show that our framework outperforms benchmarks in low data regimes.