Proceedings of the DataNinja sAIOnARA 2024 Conference 58-61 DOI: 10.11576/dataninja-1172

Beyond Trial and Error in Reinforcement Learning

Moritz Lange MORITZ.LANGE@INI.RUB.DE
Institute for Neural Computation, Ruhr University Bochum, Germany

Raphael C. Engelhardt RAPHAEL.ENGELHARDT@TH-KOELN.DE
TH Köln, Germany

Wolfgang Konen WOLFGANG.KONEN@TH-KOELN.DE
TH Köln, Germany

Laurenz Wiskott LAURENZ.WISKOTT@RUB.DE
Institute for Neural Computation, Ruhr University Bochum, Germany

Abstract

In this work, we address the trial-and-error nature of modern reinforcement learning (RL) methods by investigating approaches inspired by human cognition. By enhancing state representations and advancing causal reasoning and planning, we aim to improve RL performance, robustness, and explainability. Through diverse examples, we showcase the potential of these approaches to improve RL agents.

Keywords: Reinforcement learning, representation learning, reasoning

1. Introduction

In reinforcement learning (RL), an agent learns to act in an environment to achieve some goal. RL problems are framed as Markov decision processes (MDPs), defined by states ($\mathcal{S}$), actions ($\mathcal{A}$), transition probabilities ($\mathcal{P}$), and rewards ($\mathcal{R}$).

RL algorithms solve RL tasks by learning a mapping $\mathcal{S} \mapsto \mathcal{A}$, i.e. finding suitable actions for states, to maximize accumulated rewards. They can be model-free (e.g. [1–4]) or model-based (e.g. [5, 6]). In model-free algorithms, the agent balances exploration and exploitation while trying actions to learn a value function for state-action pairs, which is then used to sample actions. Model-based RL learns an internal model of the environment, which is used by the agent as a simulator for planning. Such environment models are usually forward models, i.e. they provide agents with the basic reasoning capacity of rolling out hypothetical actions. Model-based RL has the advantage that agents require less actual real-world experience and the disadvantage that they require a reliable model of the environment.

Both kinds of methods rely on trial and error by the agent. Agents use black-box neural networks to map raw inputs, e.g. images, to state-action values without any sophisticated understanding of state information. Networks are trained directly on the RL task of optimising return, without incentives to learn representations that could help reasoning about the environment and its dynamics. Furthermore, the state-action pairs are considered independent and not treated as part of targeted action sequences.

Humans, on the other hand, process abstract representations of information, can contextualize information and reason causally about steps required to reach a goal. In this work, we aim to bridge this gap by showing how cognition-inspired methods can improve performance, sometimes even make tasks possible in the first place, and benefit robustness and explainability.

2. State Representations

We showcase the benefits of appropriate representations with two examples. First, we demonstrate the use of representing location and heading in visual navigation as in Lange et al. [7]. In particular, we use three representation learning methods with a PPO agent [2]: (i) Slow feature analysis (SFA) [8], which is able to extract location and heading from visual input, (ii) principle component analysis (PCA), able to extract heading but not location and (iii) convolutional neural networks (CNNs), the go-to approach in this context, trained on the RL task jointly with the agent. CNNs do not encode either location or heading. Figure 1 shows how SFA outperforms the other representations. For more details, see Lange et al. [7].

Figure 1: Performance of SFA, PCA and two CNN representations on a star maze task with fixed (left) or random (right) goal position. The images on the right show the agent's observation (top) and top view of the maze (bottom; triangle: agent, cube: goal). Image modified from [7].

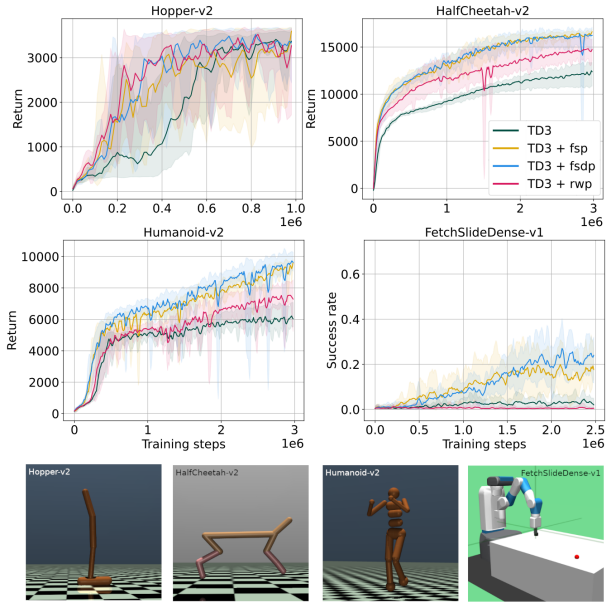

Figure 2: Performance of the TD3 algorithm with fsp, fsdp and rwp representations (see Section 2) on various environments. Image modified from [9].

Figure 2 goes beyond visual navigation. It compares different auxiliary tasks (additional tasks other than reward maximization) for representation learning, in various non-visual environments. According to our findings in Lange et al. [9], which are summarized in Figure 2, forward state (difference) prediction (fsp/fsdp) outperforms reward prediction (rwp) and baseline RL representation learning without any auxiliary task. This suggests that there is a benefit in learning representations that are generally optimized for modeling the dynamics of the environment.

Figure 3: Time evolution of different numbers of interacting objects, all generated with the same trained denoising diffusion model. This environment, Phyre, is from Bakhtin et al. [15].

3. Reasoning and Planning

Models of the environment can be statistical or causal [10]. The former is easier to learn, but the latter is more robust and generalizes better to out-of-distribution situations [11]. Both benefit from representations with high-level, causal variables. Such representation learning methods already exist and are used, for instance, in physical reasoning [12]. However, only recent gradient-based causal discovery methods are efficient and scalable enough for RL. Unfortunately, in ongoing work, we (and others [13]) found that some current approaches might fail due to various natural effects in data distributions. Still, we consider the field of gradient-based causal discovery a promising direction for causal reasoning in RL.

Beyond causality, a smart planning algorithm should be able to plan both forward from a state and backward from a goal to incorporate both constraints. This is necessary to eliminate the need for trial and error. Recently, Janner et al. [14] have made an exciting step in this direction with denoising diffusion models for invertible planning. We extended their work with a model that can handle variable time horizons and numbers of objects during inference, as well as object interactions (see Figure 3). After reintroducing start and goal conditioning from [14], this model will be useful not only for planning: An agent can also use it to reason about past and future, to explore hypothetical options or to learn from mistakes.

4. Conclusion

We illustrated through different examples how informative representations, as well as causal and invertible reasoning have the potential to improve RL agents that often rely on trial and error. Through their alignment with human reasoning, our methods can also provide robustness and explainability.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.
The ongoing work on causal models is conducted jointly with **Tim Schwabe** and **Maribel Acosta** (both TU Munich).
The ongoing work on diffusion models is conducted jointly with **Andrew Melnik** (University of Bielefeld).

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

[3] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[4] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pages 5556–5566. PMLR, 2020.

[5] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7559–7566. IEEE, 2018.

[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] Moritz Lange, Raphael C Engelhardt, Wolfgang Konen, and Laurenz Wiskott. Interpretable brain-inspired representations improve rl performance on visual navigation tasks. In *AAAI workshop eXplainable AI approaches for Deep Reinforcement Learning*, 2024.

[8] Mathias Franzius, Niko Wilbert, and Laurenz Wiskott. Invariant object recognition with slow feature analysis. In *International Conference on Artificial Neural Networks*, pages 961–970. Springer, 2008.

[9] Moritz Lange, Noah Krystiniak, Raphael C Engelhardt, Wolfgang Konen, and Laurenz Wiskott. Improving reinforcement learning efficiency with auxiliary tasks in non-visual environments: A comparison. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 177–191. Springer, 2023.

[10] Judea Pearl. *Causality*. Cambridge university press, 2009.

[11] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[12] Andrew Melnik, Robin Schiewer, Moritz Lange, Andrei Ioan Muresanu, Animesh Garg, Helge Ritter, et al. Benchmarks for physical reasoning ai. *Transactions on Machine Learning Research*, 2023.

[13] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

[14] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022.

[15] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick.

Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.