Online Representation Learning for Reinforcement Learning

Amir Samani
Department of Computing Science
University of Alberta
samani@ualberta.ca

Abstract

Reinforcement learning (RL) is a powerful paradigm that might be a solution for artificial general intelligence [Silver et al., 2021]. An RL agent interacts with its environment by performing actions and receiving observations. This interaction results in what we call the agent's data stream of experience. In many cases, observations only give partial information about the state of the environment, and the agent needs to construct its state based on the data stream of experience [Sutton and Barto, 2018]. The state is fundamental for an RL agent. The state is the input to the agent's policy and value functions. The state is both the input and the output of the environmental model of a model-based RL agent. In the past, domain experts designed the state based on their understanding of the environment and what they thought might be helpful to the agent. Recently, modern deep learning methods based on gradient descent have been applied to RL to learn the state while the agent is performing on the task of interest. Many of these deep learning methods were originally designed for supervised learning problems, which made these methods tricky to be a part of an RL agent. The agent's data stream of experience consists of highly correlated data that require workarounds such as target network and experience replay to perform well on RL problems [Mnih et al., 2015]. While these workarounds achieved significant performance, they have limited the way that RL may be used. With the emergence of popular benchmarks such as Arcade Learning Environment [Bellemare et al., 2013] and MuJoCo [Todorov et al., 2012], the final performance on these benchmarks overshadows the attempt to answer the underlying fundamental research questions about intelligence, such as learning online without storing past data points. Learning online from the most recent experiences enables the agent to perform better in the problems it is currently facing and to adapt to changes in non-stationary environments. Deep learning methods based on gradient descent have been shown to lose their adaptability as their initial randomness in their weights is lost [Dohare et al., 2021]. However, RL is well suited for online learning and dealing with non-stationary environments. RL agents can acquire rich knowledge of the environment by online interaction without storing past data points [Sutton et al., 2011]. Our methods of learning and constructing the state should not undermine these abilities; on the contrary, they should embrace them. Instead of avoiding the issues and finding workarounds for the problems arising by using deep learning methods, we should focus on finding methods that learn the state online without storing past data points. Looking back at the history of artificial intelligence research, we see that utilizing general methods such as search and learning is the most effective [Sutton, 2019]. Perhaps we can start by employing a search method similar to the generate-and-test algorithm [Mahmood and Sutton, 2013]. The search algorithm improves the agent's performance and also becomes better at learning and adapting to the environment in the future. We may start by dealing with the representational problems in isolation. For instance,

Preprint. Under review.

Animal Learning Testbeds [Rafiee et al., 2021] provide benchmarks inspired by animal learning experiments. Each of these benchmarks focuses on a particular problem, such as remembering past events, which helps us better evaluate our methods of learning the state. We need to study various aspects of learning the state directly from the agent's data stream of experience and benchmarks similar to Animal Learning Testbeds enable us to focus on the online setting.

References

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Shibhansh Dohare, A. Rupam Mahmood, and Richard S. Sutton. Continual backprop: Stochastic gradient descent with persistent randomness, 2021.

Ashique Rupam Mahmood and Richard S. Sutton. Representation search through generate and test. In *Proceedings of the 12th AAAI Conference on Learning Rich Representations from Low-Level Sensors*, AAAIWS'13-12, page 16–21. AAAI Press, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL `http://dx.doi.org/10.1038/nature14236`.

Banafsheh Rafiee, Zaheer Abbas, Sina Ghiassian, Raksha Kumaraswamy, Richard S. Sutton, Elliot Ludvig, and Adam White. From eye-blinks to state construction:diagnostic benchmarks for online representation learning. *CoRR*, abs/2011.04590, 2021.

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Richard S. Sutton. The bitter lesson, Mar 2019. URL `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), Taipei, Taiwan, May 2-6, 2011, Volume 1-3*, pages 761–768. IFAAMAS, 2011.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.