Reward-driven State Tabulation for Model-based Reinforcement Learning

1 Context

Reinforcement Learning (RL) has had several achievements such as in the game of Go [4]. These recent achievements were obtained from the combination of RL with deep neural networks, also known as deep RL. The vast majority of deep RL algorithms use their neural architecture to approximate value functions following the scheme of fitted value iteration, as popularised by DQN [3]. While the convergence of fitted value iteration is well understood with linear function approximators, no such guarantees exist in the non-linear case and empirical evidence suggests that the quality of the approximation of value functions is poor in deep RL [2].

Instead of directly approximating value functions, a less researched alternative to scale RL to large state spaces is to use neural networks to aggregate states into fewer, abstract states, such that the abstract Markov Decision Process (MDP) behaves similarly to the ground MDP. As the abstract MDP is of reasonable size, one can directly apply tabular RL algorithms to learn Q-functions, even in high dimensional, image-based environments, such as in [1]. The resulting method is model-based, and the model is given by the empirical estimates of the transition and reward functions of the abstract MDP. However, a recent trend in model-based RL is to not try to capture all the information about the state but only that which helps predict future rewards. For example, in a robotics object manipulation task the model might only need to predict if a glass will break after performing a given action, but not the position of each glass shard after the glass would break, which is significantly harder to predict but not necessarily useful for the task.

In this vein, while the work of [1] uses variational inference to maximize the log-likelihood of future states—a completely reward-free loss that does not filter information contained in the ground states—we will be interested in this internship in only predicting state information that is useful for for predicting future rewards. To do so, the work program of the intern will comprise the following steps i) to reproduce the results of [1] to establish a baseline, ii) to propose a new, reward-driven, loss function to train the neural state aggregator, iii) to devise a model-based RL algorithm operating on the abstract MDP iv) to evaluate the algorithm against the baseline on tasks where states contain superfluous information and investigate whether the neural state aggregator can learn to ignore it.

To apply, send your CV and grades to `riad.akrour@inria.fr` and `matheus.medeiros-centa@inria.fr`. For more internship proposals from Inria Scool, please see `https://team.inria.fr/scool/job-offers/`.

References

[1] Dane Corneil, Wulfram Gerstner, and Johanni Brea. "Efficient Model-Based Deep Reinforcement Learning with Variational State Tabulation". In: *International Conference on Machine Learning (ICML)*. 2018.

[2] Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. "A Closer Look at Deep Policy Gradients". In: *International Conference on Learning Representations (ICLR)*. 2020.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, et al. "Human-level control through deep reinforcement learning". In: *Nature* (2015).

[4] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* (2016).