Introspective Reinforcement Learning and Learning from Demonstration

Extended Abstract

Mao Li
University of York
York, United Kingdom
ml1480@york.ac.uk

Tim Brys
Vrije Universiteit Brussel
Brussels, Belgium
timbrys@vub.ac.be

Daniel Kudenko
University of York
York, United Kingdom
National Research Acad. University of the Russian Academy of Sciences
St Petersburg, Russia
JetBrains Research
St Petersburg, Russia
daniel.kudenko@york.ac.uk

ABSTRACT

Reinforcement learning is a paradigm used to model how an autonomous agent learns to maximize its cumulative reward by interacting with the environment. One challenge faced by reinforcement learning is that in many environments the reward signal is sparse, leading to slow improvement of the agent's performance in early learning episodes. Potential-based reward shaping is a technique that can resolve the aforementioned issue of sparse reward by incorporating an expert's domain knowledge in the learning via a potential function. Past work on reinforcement learning from demonstration directly mapped (sub-optimal) human expert demonstrations to a potential function, which can speed up reinforcement learning. In this paper we propose an introspective reinforcement learning agent that significantly speeds up the learning further. An introspective reinforcement learning agent records its state-action decisions and experiences during learning in a priority queue. Good quality decisions will be kept in the queue, while poorer decisions will be rejected. The queue is then used as demonstration to speed up reinforcement learning via reward shaping. An expert agent's demonstrations can be used to initialise the priority queue before the learning process starts. Experimental validations in the 4-dimensional CartPole domain and the 27-dimensional Super Mario AI domain show that our approach significantly outperforms state-of-the-art approaches to reinforcement learning from demonstration in both domains.

ACM Reference Format:
Mao Li, Tim Brys, and Daniel Kudenko. 2018. Introspective Reinforcement Learning and Learning from Demonstration. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018*, IFAAMAS, 3 pages.

1 INTRODUCTION

One of the main challenges Reinforcement Learning [9] faces is dealing with the sparsity of rewards that is present in many tasks. A reinforcement learning agent will only very slowly learn to solve a task through trial-and-error if the feedback it receives in the form of rewards is given sparsely. This lack of feedback, and thus gradient, makes the agent explore the task uniformly at random, unless special mechanisms are implemented that guide the agent's exploration.

One way of tackling this problem that is often taken is to include prior (heuristic) knowledge that can bias the agent towards what are *a* priori believed to be good states, or good behaviour. Prior knowledge can come in the form of rules defined by a domain expert [5], demonstrations provided by such an expert [7][10][8][4], etc.

Conversely, a second stream of thought, orthogonal to the former, tries to leverage internal knowledge and experiences in the current task to generate internal rewards to bias exploration in the absence of immediate external rewards. Approaches relying on intrinsic motivation [1, 6] attempt to leverage the ideas of curiosity and uncertainty to achieve more intelligent and effective ways of exploration.

This paper falls in the second category, as we develop a technique that allows the agent to bias its exploration by generating internal shaping rewards based on its own successful previous experiences in the task.

2 INTROSPECTIVE RL

Typically, in complex environments, rewards must be observed many times before the agent acquires a significant behavioural bias towards pursuing those rewards. In the Introspective Reinforcement Learning approach we propose, we leverage these experiences to include a more explicit bias, by shaping the reward function, rewarding current behaviour that is similar to past behaviour that led to rewards. A different perspective on this is to say that those previous experiences that led to rewards are task demonstrations (provided by the agent itself), which can be used to bias the agent's current exploration in the same fashion as external expert demonstrations might be used. If actual external expert demonstrations are available, these could even be used to initialised the introspective agent's bias.

2.1 Collecting the experiences

Introspective Reinforcement Learning extends RL by adding an experience filter module, and addresses the reward sparsity problem

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

using dynamic reward shaping based on the filtered experiences. The experience filter collects the agent's exploratory behaviour that led to positive outcomes into a priority queue. These experiences are then immediately used as an exploratory bias to speed up the learning process.

Specifically, during a learning episode, every state-action-next state-reward tuple (s_t, a_t, s_{t+1}, r_t) is stored in memory. At the end of the episode, the Q-value $\widehat{q(s,a)}$ for each state-action in this episode is estimated on-policy according to reward r_t.

The state-action pairs and their estimated Q-values are then stored in a limited priority queue with the estimated Q-values being the sort key of the queue. Progressively, poorer Q-value elements of the queue will be removed, and experiences with higher estimated performance levels will remain. Thus, the exploratory bias induced by these experiences will progressively increase in quality.

2.2 Defining the Potential Function

In the manner of Brys et al. [2], we encode state-action information as a potential function using a Gaussian similarity metric over state-action pairs. The assumption is that, if in the agent's past experience, certain state-action pairs led to high rewards, taking the same action in a similar state might lead to similarly high rewards. We modify [2]'s potential function by incorporating the estimated Q-value in the experience buffer:

$$\Phi(s,a) = \rho \max_{(s^d,a)\in PQ} g(s, s^d, \Sigma)\widehat{q}(s^d, a) \qquad (1)$$

with g the similarity metric over states and PQ the priority queue.

Since the agent progressively collects more and more experiences, in principle of higher and higher quality, the potential function will change from episode to episode, making this potential function a dynamic one. Note that theoretical guarantees, i.e. that the optimal policy does not change, hold for dynamic potential-based advice [3].

2.3 Initialising with Demonstrations

Even though the introspective reinforcement learning idea is focused on using the agent's own experiences to bias its learning, it is also completely amenable to receiving an a priori bias from demonstrations provided by an external agent. Such demonstrations can easily be incorporated by putting them through the same process of estimating Q-values and storing them in the priority queue as the experiences collected by the agent itself. When qualitatively good, these demonstrations can prevent the agent from initially filling the priority queue with whatever low-quality random trajectories it executes first, allowing it to be positively guided from the first episode.

3 EXPERIMENTAL VALIDATION

Two domains, CartPole and Super Mario, were selected to demonstrate the strength of the proposed approach in this paper. $Q(\lambda)$-learning and Reinforcement Learning from Demonstration [2] were used as benchmarks to be compared with the learning curve of the proposed approach.

Figure 1 shows the results of comparing plain Q(λ)-learning with introspective RL (without demonstrations) for various λ. While both methods converge to optimal behaviour, the results show that introspection leads the agent to learn significantly faster than the regular Q(λ)-learning agent in every case.

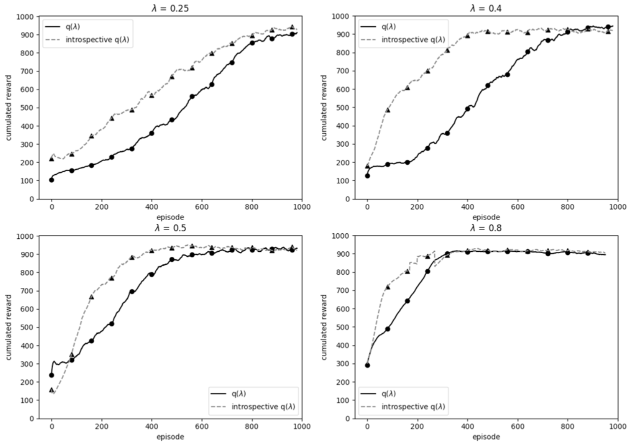

Figure 1: CartPole learning curves of Q(λ)-learning, and Introspective RL for $\lambda \in \{0.25, 0.4, , 0.5, 0.8\}$.

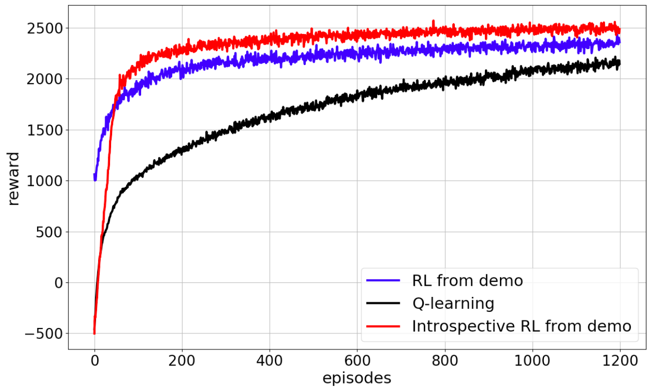

Figure 2: Super Mario Domain learning curves of q(λ)-learning, RLfD, and Introspective RL with demonstration

In another set of experiments, shown in Figure 2, 20 demonstration episodes from a human player (all with a performance score between 400 to 650) are used to initialise the priority queue for introspective RL in Super Mario. The results show that in this highly complex domain, introspective RL with demonstrations outperforms both Reinforcement Learning from Demonstration and regular $Q(\lambda)$-learning.

REFERENCES

[1] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*. 1471–1479.

[2] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. 2015. Reinforcement Learning from Demonstration through Shaping.. In *IJCAI*. 3352–3358.

[3] Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. 2015. Expressing arbitrary reward functions as potential-based advice. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[4] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Andrew Sendonaris, Gabriel Dulac-Arnold, Ian Osband, John Agapiou, et al. 2017. Learning from demonstrations for real world reinforcement learning. *arXiv preprint arXiv:1704.03732* (2017).

[5] Maja J Mataric. 1994. Reward functions for accelerated learning. In *Machine Learning: Proceedings of the Eleventh international conference*. 181–189.

[6] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. *arXiv preprint arXiv:1705.05363* (2017).

[7] Stefan Schaal. 1997. Learning from demonstration. *Advances in neural information processing systems* 9 (1997), 1040–1046.

[8] Halit Bener Suay, Tim Brys, Matthew E Taylor, and Sonia Chernova. 2016. Learning from demonstration for shaping through inverse reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 429–437.

[9] R.S. Sutton and A.G. Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1. Cambridge Univ Press.

[10] Matthew E Taylor, Halit Bener Suay, and Sonia Chernova. 2011. Integrating reinforcement learning with human demonstrations of varying ability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 617–624.