Hierarchical Reinforcement Learning for a Robotic Partially Observable Task

Denis Steckelmacher, Hélène Plisnier, Diederik M. Roijers, and Ann Nowé

Vrije Universiteit Brussel, Brussels, Belgium
{dsteckel,hplisnie,droijers,anowe}@vub.ac.be
http://steckdenis.be/bnaic_demo.mp4

Abstract. Most real-world reinforcement learning problems have a hierarchical nature, and often exhibit some degree of partial observability. While hierarchy and partial observability are usually tackled separately, we illustrate on a complex robotic task that addressing both problems simultaneously is simpler and more efficient. We decompose our complex partially observable task into a set of sub-tasks, in a way that allows each sub-task to be solved by a memoryless option. Then, we implement Option-Observation Initiation Sets [4], that make the selection of any option conditional on the previously-executed option. Our agent successfully learns the task, achieves better results than a carefully crafted human policy, and does so much faster than an recurrent neural network over options [3].

1 Background

Options factor a complex task into simpler sub-tasks. The agent learns a top-level policy that repeatedly selects options, that in turn execute a sequence of actions before returning [5]. In Partially Observable MDPs (POMDPs), the agent does not fully observe the state of the environment. Remembering past observations therefore becomes necessary to behave optimally [2]. While most current approaches to POMDPs rely on recurrent neural networks [1], sometimes used on top of options [3], the experiment in this demonstration shows that Option-Observation Initiation Sets [4] allow a challenging robotic partially observable task to be learned, achieving better results than an expert policy, while providing intuitive explanations of the behavior of the robot.

2 Experiment

This experiment is inspired from a real-world industrial object gathering task, where a robot has to bring objects from two terminals to a central carrier belt, for further processing by other robots. A Khepera III robot has to gather objects from two terminals, separated by a wall, and to bring them to the root, as shown in Figure 1 (a). Objects have to be gathered one by one from a terminal until it becomes empty. When a terminal is emptied, the other one is automatically refilled. The robot therefore has to alternatively gather objects from both terminals. The root is colored in red and marked by a paper QR-code encoding `1`. Each terminal has a screen displaying its color and a `1` QR-code

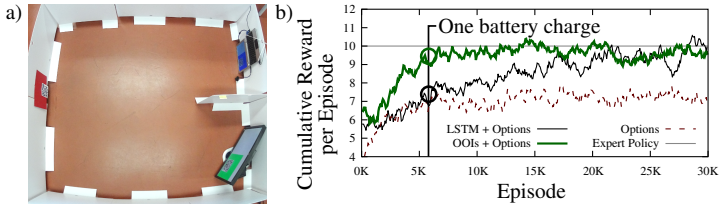

Fig. 1. a) Experimental setup. b) Cumulative reward per episode for different agents (*OOIs + Options* is ours). Our agent outperforms a human expert policy.

when full, `2` when empty. Because the camera mounted on the robot cannot read QR-codes from far away, the state of a terminal cannot be observed from the root, where the agent has to decide to which terminal it will go. This makes the environment partially observable. Fixed options allow the robot to move towards the largest red, green or blue blob in its field of view. The options terminate as soon as a QR-code close enough to be read. The robot has to learn a policy over options that solves the task.

During the demonstration, we will compare the policy learned by the robot with an expert policy, that empties a terminal before switching to the other one (see video[1]). The learned policy obtains returns comparable with the expert policy, and can be learned by the robot in a single battery charge (see Figure 1, b).

Acknowledgments

The first author is “Aspirant” with the Science Foundation of Flanders (FWO, Belgium, 1129317N). The third author is “Postdoctoral Fellow” with the FWO (12J0617N).

References

1. Bram Bakker. Reinforcement learning with long short-term memory. In *Advances in Neural Information Processing Systems (NIPS)*, volume 14, 2001.
2. Long-Ji Lin and Tom M Mitchell. *Memory approaches to reinforcement learning in non-Markovian domains*. Carnegie-Mellon University. Department of Computer Science, 1992.
3. Mohan Sridharan, Jeremy Wyatt, and Richard Dearden. Planning to see: A hierarchical approach to planning visual actions on a robot using POMDPs. *Artificial Intelligence*, 174:704–725, 2010.
4. Denis Steckelmacher, Diederik M. Roijers, Anna Harutyunyan, Peter Vrancx, and Ann Nowé. Reinforcement learning in POMDPs with memoryless options and option-observation initiation sets. *CoRR*, abs/1708.06551, 2017.
5. Richard Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112:181–211, 1999.

[1] `http://steckdenis.be/bnaic_demo.mp4`