Learning Generalized Temporal Abstractions across Both Action and Perception

Khimya Khetarpal
School of Computer Science, McGill University
Mila - Reasoning and Learning Lab
Montreal, Quebec

Introduction

Reinforcement Learning (RL) involves a decision-making agent learning to achieve *goals*, while dealing with *uncertainty and limited data* about its environment (Sutton and Barto 1998). The goal in a RL problem is indicated by an external *reward signal* provided by the environment with which the agent is interacting. The agent learns what to do – how to map situations to actions so as to maximize the cumulative expected reward over time.

The RL problem specification is defined more formally in terms of Markov Decision Processes (MDPs). In a MDP, the state observed and the action taken at time-step t determine the distribution of the state and the immediate reward at time $t+1$. This allows learning fine-grain courses of action. One could however, learn much more rapidly by abstracting away the myriad of details and considering actions of longer duration, which generate longer-range transitions. This serves as a motivation for learning *temporal abstractions* in the framework of RL. Hierarchical Reinforcement Learning (HRL) methods aim to find *closed-loop policies* which have a temporal extent, and can then be used instead (or in addition to) one-step actions. Semi-Markov decision processes (SMDPs) provide a generalized framework for HRL methods, by allowing the amount of time between two decision points to be a random variable (Puterman 1994).

Learning temporal abstractions which are partial solutions to a task and could be reused for other similar or even more complicated tasks is intuitively an ingredient which can help agents to plan, learn and reason efficiently at multiple resolutions of perceptions and time. Just like humans acquire *skills* and build on top of already existing skills to solve more complicated tasks, AI agents should be able to learn and develop skills continually, hierarchically and incrementally over time. In my research, I aim to answer the following question: *How should an agent efficiently represent, learn and use knowledge of the world in continual tasks?* My work builds on the options framework, but provides novel extensions driven by this question.

Theme 1: Learning Temporal Abstractions with Interest Functions

The options framework (Sutton, Precup, and Singh 1999) enables an agent's trajectory to be analyzed in both through the lens of discrete-time transitions or through SMDP-style transitions. Recent research has demonstrated that options can be learned automatically, end-to-end, for a given task (Bacon, Harb, and Precup 2017). Unfortunately, this can result in degenerate solutions, with either one option being used for the entire task, or option duration collapsing to single time steps. This type of degenerate solution is potentially due to a simplifying assumption used in the option-critic: that all options are available in all states.

In order to learn options that represent specialized and meaningful skills for lifelong learning, we revisit the idea of an initiation set, used in the options framework, but through a formulation that is more amenable to learning. We introduce the notion of *interest functions* $I_\omega : S \longrightarrow \mathbb{R}$. The interest function $I_\omega(s)$ is indicative of the availability of an option ω in state s. The idea is inspired by human visual attention: while we engage in any task, each skill employed is specialized in attending to only certain states. For example, a skill such as 'stop if the traffic light is red' is only applicable in states in which a traffic light is present.

We define interest functions in the options framework as follows. The state-value function is defined as:

$$V_\Omega(s) = \sum_\omega \pi_{I_{\omega,z}}(\omega|s) Q_{\Omega,\theta}(s,\omega) \tag{1}$$

where $Q_{\Omega,\theta}$ is the option-value function parameterized by θ, and the probability of an option being sampled in a given state is defined as:

$$\pi_{I_{\omega,z}}(\omega|s) = I_{\omega,z}(s)\pi_{\Omega,\theta}(\omega|s) \Big/ \sum_\omega I_{\omega,z}(s)\pi_{\Omega,\theta}(\omega|s) \tag{2}$$

Here, $I_{\omega,z}(s)$ is the interest function parameterized by z. The agent initially would consider that all options are available everywhere. As learning progresses, we expect that the emerging options will be specialized over *different* state regions. Starting with the option-value function, we can derive the interest function gradient, obtaining the following result:

Theorem 1. *Given a set of Markov options with stochastic, differentiable interest functions $I_{\omega,z}$, the gradient of the*

expected discounted return with respect to z at (s, ω) is:

$$\sum_{s',\omega'} \hat{\mu}_{\Omega}(s', \omega'|s, \omega)\beta_{\omega,\nu}(s')\frac{\partial \pi_{I_{\omega,z}}(\omega'|s')}{\partial z}Q_{\Omega}(s', \omega')$$

where $\hat{\mu}_{\Omega}(s', \omega'|s, \omega)$ is the discounted weighting of the state-option pairs along trajectories starting from (s, ω) sampled from the sampling distribution determined by $I_{\omega,z}$.

We are currently evaluating this approach in simulated environments. But, in order to understand more broadly the advantage of using some form of interest or attention in RL, we have already done some preliminary empirical work which investigates the role of biases acquired from human visual attention in automated RL agents. In (Khetarpal and Precup 2018), we leverage where humans look in an image as an implicit indication of what is salient for decision making. Our goal was to explore how foveating around the regions where humans look impacts the reinforcement learning process, especially focusing on robustness and continual learning. We *hypothesized* that knowing where to look aids continual learning across tasks.

We trained a Visually-Attentive UNREAL agent based on the UNREAL (Jaderberg et al. 2016) agent with varying degrees of foveation. To evaluate the trained models for continual learning, we introduced 3 types of perturbations in the input frames namely; Gaussian noise (easy), tinting of images at random with the same hue (moderate), and tinting of images at random with different hues (difficult). We show empirically that upon encountering flickering in frames at random, the Visually-Attentive UNREAL agent is still able to perform as well as the baseline and is relatively more robust to distractors in both easy and moderate tasks. The project page for this work[1] provides an overview of the results along with links to the code and paper.

Theme 2: Temporal Abstractions across both Perception and Action

Humans are presented with a never ending stream of sensorimotor data in the rich environments in which they live. Similarly, modern RL agents experience rich simulated worlds through their sensors. What our agents *see* forms an important source of information. My goal in this part of the thesis is to investigate the best way for an agent to represent and learn from this sensory stream. A powerful approach that has been studied already is to learn *predictive* (Littman and Sutton 2002) models of the world, such as predicting how objects and interactions with them would change the environment.

Analogous to temporally extended actions, we propose learning temporally extended *perception*. The key idea is to learn temporal abstractions unifying both action and perception. Most of the success in the field of computer vision has relied on static labelled data of multiple classes. However, it is much more natural to learn about the world knowledge through *embodied* interaction with the world. We propose to allow perceptual features to represent multiple time steps, in synchrony with the agent's options.

[1] https://sites.google.com/view/attendbeforeyouact

For example, consider any household environment where a robot is left to navigate and learn about the dynamics of the environment via interacting with it. Consider that we have a set of task descriptions, such as opening a door or picking up an object, each corresponding to a skill. Each task is specified through a pseudo reward function. We would like the agent to develop features which can serve as meaningful pseudo rewards. Can we represent the task specific information in the form of temporal abstractions which carry the notion of both visual features and actions across multiple scales of time and state space granularity? How can we learn a feature embedding which supports learning about useful and diverse skills? One of the most challenging aspects of this problem is – how do we learn temporal abstractions that enable the agent to be a lifelong learner?

Acknowledgements and Timeline

For theme I; the problem formulation and derivation has been completed. The algorithm for the tabular intra-option Q learning has also been laid out. This work is in the experimentation phase and is scheduled to be completed by the time of the conference. The initial direction was suggested by my advisor. I then worked independently on the derivation and algorithm of this approach which were both reviewed by my advisor. For theme II; we are currently brainstorming about the way in which the open questions outlined above could be formulated mathematically in the context of lifelong learning. This part of the thesis is in the problem formulation phase and could benefit from expert input.

References

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In *AAAI*, 1726–1734.

Jaderberg, M.; Mnih, V.; Czarnecki, W. M.; Schaul, T.; Leibo, J. Z.; Silver, D.; and Kavukcuoglu, K. 2016. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*.

Khetarpal, K., and Precup, D. 2018. Attend before you act: Leveraging human visual attention for continual learning. *arXiv preprint arXiv:1807.09664*.

Littman, M. L., and Sutton, R. S. 2002. Predictive representations of state. In *Advances in neural information processing systems*, 1555–1561.

Puterman, M. 1994. Markov decision processes. 1994. *Jhon Wiley & Sons, New Jersey*.

Sutton, R. S., and Barto, A. G. 1998. *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1st edition.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112.