Reinforcement Learning in POMDPs

Matthew Feng

October 29, 2018

1 Logistic Information

Student Matthew Feng

Faculty Supervisor Leslie Kaelbling

Term Fall 2018

Date October 29, 2018

2 Project Overview

Recently, Karkus et al. [2] published the QMDP-net, a fully differentiable neural network with structural priors that encode the Q_{MDP} algorithm for finding approximate solutions to POMDPs. By incorporating the procedure of the Q_{MDP} algorithm into the structure of the network, the network was able to outperform other network architectures on many tasks — in particular those requiring a generalization of demonstrations provided as training examples — by learning a model of the environment that allowed the embedded algorithm to perform well. This marks an interesting shift in planning: rather than developing a planning algorithm around a model, develop a model that will complement the chosen planning algorithm.

Reinforcement learning has also shown increasing promise in recent years with mastery in domains including Atari and Go. The results are impressive and exciting; however, these domains have remained fully observable, limiting the scope of the problems on which the algorithms may be applied. Attempts to extend these approaches to partially observable environments [1] have typically involved introducing general recurrence architectures to the networks, such as connecting LSTMs to convolutional layers, thereby adding capacity to distinguish temporal dependencies.

This project aims to combine the work of Karkus et al. with the advances in model-free reinforcement learning, by embedding planning algorithms as recurrences into previously successful model-free network architectures, such as Deep Q-Networks. In this way, we hope to augment these networks with a capacity for planning and state prediction. The project will focus on finite, discrete state spaces on simple domains such as *GridWorld*, before attempts to scale up to

more complex environments such as *Montezuma's Revenge*, a game requiring substantial planning for success.

I will be working with Peter Karkus in the Learning and Intelligent Systems Group in CSAIL.

3 Personal Role and Responsibilities

I will be working 15 hours a week to design, train and evaluate models that combine planning algorithms with model-free approaches to reinforcement learning. Additionally, I hope to work with Peter to develop a more theoretically sound basis for the designs of the networks (or gain a better understanding of how such theory is gradually developed).

I will also meet with Peter Karkus every Thursday from 2-3pm to review progress over the past week and discuss questions for the following week. I will be responsible for preparing relevant questions and ideas to be discussed during these meetings.

4 Goals

The goal of the project by the end of this term is to have a network architecture that is able to exploit planning in networks originally designed as model-free. Hopefully, such new network architectures will be able to perform better in environments requiring substantial planning than their model-free counterparts.

By the end of this term, I aim to have a much more thorough understanding of both planning algorithms and reinforcement learning (via papers and Barto and Sutton's book on reinforcement learning), as well as a better ability to develop theory regarding the theory of deep learning.

I will be working primarily with *TensorFlow*, and so I also aim to gain a better fluency with the library.

Finally, I hope to be able to interact with the other members in the Learning and Intelligent Systems Group, to be able to discuss ideas and recent papers for a more complete understanding of both the intuition and theoretical details.

5 Personal Statement

I am excited for this UROP because of the ideas behind the project (e.g. learning "intuitive" or useful models), and its application of those ideas to reinforcement learning, a field I have always found exciting and relevant. I have also been very interested in learning what research in machine learning is like, in trying and experimenting with new models and theories. Finally, I am also excited to have a group that I may easily discuss research results with.

References

[1] M. J. Hausknecht and P. Stone. Deep recurrent q-learning for partially observable mdps. *CoRR*, abs/1507.06527, 2015.

[2] P. Karkus, D. Hsu, and W. S. Lee. Qmdp-net: Deep learning for planning under partial observability. *CoRR*, abs/1703.06692, 2017.