Unifying value learning and model learning

Carles Gelada

January 12, 2017

Abstract

Learning a model of the world, that focuses on the relevant information, is a fundamental challenge of artificial intelligence. In this work I present a method of integrating model based and model free reinforcement learning in a simple, end-to-end manner. It naturally integrates with DQN, extending it with the ability to predict the future and learns inner representations that are useful to do both, learn to predict the action-values, and to predict the future action-values. Once trained, planning algorithms can be used together with the learned model to select better actions. The presented architecture has a minimal computational overhead compared to DQN. The proposed model is general, it can be used to extend any kind of RL algorithm acting in any kind of MDP.

1 Introduction

Deep learning has proven very successful in RL ... The reason Deep Learning has been so successful is that it can learn the representations necessary for the task. Yet, the field of model based reinforcement learning (MBRL) has been mainly focused on learning a transition function of the state, and then using it to plan. But in cases where the state is high dimensional (e.g. an image), learning a transition function is a very challenging task to learn and carries a lot of computational complexity. Thus, the idea of learning the transition function of a representation especially learned to perform this task is very appealing.

Here I present a new architecture capable of learning the transition function of a inner, hidden layer and the current Q values.

I evaluate PDQN on the atari domain...

The reference implementation is released at ... with the aim of facilitating further development in this area.

The ability to "imagine" what would happen if an action is taken seems a fundamental property of intelligence. Considering different situations, and reasoning about them is a useful cognitive tool. Giving neural networks these tools is therefore a interesting research problem.

Here, I am only going to discuss a constrained form of imagination, model based planning.

In highly structured environments learning to plan could be much easier than directly learning a policy. Planning could then be used to take better actions or even update the policy.

The main component necessary for planning is the transition probability matrix or, if the MDP is deterministic, a transition function). In the case of a large, stochastic state space, it would probably be better to learn a stochastic transition function than the probability transition matrix.

Learning to plan will be easier than learning the value function because the world is constant, while the targets of the value function vary, (not totally true in PDQNs case). Therefore, the learning of the transition function will converge much quicker and need a lot less data. Once The model is learned we can use computation to learn, as opposed to using real world experience.

Also, learning a model of the (the right representation) world might simply be easier and generalize more quickly than learning a value function. It might be possible that even when having access to the optimal value function, an agent training on them would take longer and generalize worse than an agent trying to learn to predict future values.

Intuitively, at the beginning the inner representation is driven thowards a point where it represents all the necessary information to predict the future hidden state. Then, it will converge before the tar-

gets stabilize. It will become a quasi static problem early in the training process.

Searching deeper can be interpreted as reducing dependence on non static targets

The goal of this work is not to explore how planning can be used in the context of RL but rather to show that it is viable to efficiently to learn a model of the world. Since the model scores are greatly improved by the planning module, it follow by extension that the value prediction in improved.

It might be that the forcing the inner representations to be able to predict the task acts as a regularizer. Forcing the representations to contain information that will likely be key to predicting the current value. This is closely related to the work on auxiliary tasks and extra supervised targets that has recently shown significant advantage in RL. In the experiments I test this hypothesis.

2 Previous Work

MBRL is a tries to use MBRL can be viewed as a method to compute better value functions. In this work the value functions are only used for the action selection but other work like ... has shown that Q learning with better values can learn much better Q values.

Learning pixel models is video prediction Work on merging video prediction with RL

[3] [1] [2]

The predictron [4] Very similar, developed independently.

3 The model

X define standard RL stup X X define DQN X X define prediction extension to DQN X

3.1 Planning DQN

The principles are used to extend DQN with the ability to plan. X explain how, using watkins, it can be trained with very similar computational cost X

4 Experiments

I evaluate PDQN on the atari domain...

5 Discussion

The integration of planning into deep RL presents many new challenges and opportunities. Here I present some of them.

5.1 Using planning to learn the values

From the beginning of MBRL, one of the main motivations was to use the planning capabilities to improve the value functions. X use the predictive model to update the Q values, relation to enforcing consistency? X X Train with generated data. Related to enforcing consistency. Similar to training with the RM but since the data is unlimited it wont overfit. Therefore a lot of training could be done with little real world experience. X X Explain that planning can be seen as a kind of RM X

5.2 Planning algorithms

X explore different planning algorithms X This is equivalent to the adaptive computation in the predictrion.

5.3 Stochastic transition functions

X learn stochastic transition functions. They might learn more natural representations. Finding an abstraction of a stochastic world where the transition function is deterministic might be very hard or impossible. X

5.4 Continuous action spaces

Although learning to predict values in continuous spaces does not require any modification to the algorithm, it is unclear how planning would be done. Maybe similar methods to the ones used in continuous RL could be used to learn search policies.

6 Conclusions

I have presented a simple, yet powerful architecture capable of learning the transition function. It can be readily integrated with any planning algorithm.

References

[1] Justin Fu and Irving Hsu. Model-Based Reinforcement Learning for Playing Atari Games. Technical report, 2016.

[2] J. Oh, X. Guo, H. Lee, R. Lewis, and S. Singh. Action-Conditional Video Prediction using Deep Networks in Atari Games. *Nips*, page 9, 2015.

[3] J. Schmidhuber. On Learning to Think: Algorithmic Information Theory for Novel Combinations of Reinforcement Learning Controllers and Recurrent Neural World Models. *ArXiv e-prints*, 2015.

[4] D. Silver, H. V. Hasselt, M. Hessel, T. Schaul, A. Guez, T. Harley, G. Dulac-arnold, D. Reichert, N. Rabinowitz, A. Barreto, and T. Degris. THE PREDICTRON: END-TO-END LEARNING AND PLANNING. *submitted*, pages 1–11, 2017.