Sample-Efficiency in Complex Environments using Reinforcement Learning

Nimish S 312217104098
***** 312217104***
***** 312217104***
BE CSE, Semester 7

*****
Supervisor

Project Review: 0 (24 October 2020)
Department of Computer Science and Engineering
SSN College of Engineering

1 Abstract

Current Reinforcement Learning algorithms excel at performing in non-hierarchial environments. However, when it comes to complex, hierarchial & sparse environments they are impractical to use as sample-inefficiency becomes more prominent due to the curse of dimensionality. This makes reinforcement learning impractical to use in many real-world situations as many of these environments are complex, hierarchial & sparse in nature. Superior exploration techniques like Curiosity-driven Exploration have proven to improve sample efficiency in simple environments. We propose to use such techniques in combination with RL algorithms like DDPG and DQN to successfully navigate a complex environment.

2 Introduction

Reinforcement learning algorithms aim at learning policies for achieving target tasks by maximizing rewards provided by the environment. Reinforcement learning combined with neural networks has recently led to a wide range of successes in learning policies for sequential decision-making problems. From a practical standpoint, RL is currently used in simple environments where PID *(Proportional Integral Derivative)* controllers cannot be easily applied (e.g. simple robotic control, HVACs & autonomous vehicles). Currently, it is crucial for the environments to be designed in such a way as to encourage learning.

However, in most real-world applications, there are three important environmental characteristics most reinforcement learning algorithms struggle with. The rewards supplied to the agent are extremely sparse or sometimes missing altogether. The actions to be taken are hierarchial in nature (i.e. each goal state does not result in termination, rather opens the possibility to new goal states). The state space is extremely high dimensional leading to a lot of time spent on extracting useful features from higher-level representations. A combination of all these problems results in extremely sample-inefficient performance of existing algorithms.

In 2018, the team at OpenAI tried to attain human-level control in an extremely complex environment called Dota2. It is a strategic real-time video game with 170,000 possible actions in each time-step that took humans approximately 600 hours to learn. The rewards too were very sparse with the agent attaining significant reward once every 5-10 minutes. A *Proximal Policy Optimization* agent was trained over a period of 10 months. It cost 7.5 million dollars to train and was trained using 128,000 CPU cores and 256 GPUs. It read in 1,048,576 observations every second which is an unimaginably large number of samples. This entire setup is simply impractical for use on a large scale.

Sample-inefficiency is just the result of the explore-exploit dilemma in practice. In order to attain the maximum possible reward, sufficient exploration must be done to identify the best possible sequence of decisions in a given situation and then the agent must exploit its knowledge of the environment. There is a fine line between exploration and exploitation as more exploration would take a lot of time to experience states further down the hierarchy and pure exploitation will result in the agent getting stuck in a local maxima.

Currently for complex enironments, algorithms explore using a naive method based on adding random noise to the action probabilities of the agent. We propose to use intelligent exploration techniques introduced in much simpler environments in combination with powerful RL algorithms to significantly improve sample efficiency in complex environments. In order to test this theory, we plan to use a complex environment called MineRL which is based on a 3D sandbox survival video game called Minecraft. The game focuses on the player collecting resources and crafting materials in order to survive. We plan to test this theory first using the SuperMarioBros environment which is a 2D environment but presents its own set of complications due to its large non-repetitive state space and complex action space.

3 Literature survey

Some of the existing approaches use Inductive Logic Programming that do not use probabilistic graphical model to compute conditional probabilities for inferred facts. To avoid this, a statistical relational learning approach is adopted. SRL handles both uncertainty and structured data, integrating first-order logic and probabilistic graphical models. The alternative is to take SRL approaches such as Markov Logic Networks (MLN) framework for both learning first order rules and probabilistic inference of additional facts. But MLN may result in an intractably large graphical model for large datasets.

4 Proposed system

A Rule Learner is applied on the set of facts an IE system has extracted from the document. The Rule Learner first updates the frequency of occurrence of each relational predicate. Then it builds a Bayesian network whose nodes represent relation extractions. It then traverses the graph to know the first order rules. The learner traverses the resulting graph to construct rules. For each directed edge (x, y) in the graph, it constructs a rule in which the body contains x and the head is y head. System architecture for inferring implicit facts using BLPs is shown in Figure **??**.

References

[1] Sindhu Raghavan, Raymond J. Mooney, *Online Inference-Rule Learning from Natural-Language Extractions*, University of Texas at Austin, In Proceedings of the 3rd Statistical Relational AI (StaRAI-13) workshop at AAAI '13, July 2013.

[2] Sindhu Raghavan, Raymond J. Mooney, and Ku. H,*to read between the lines using Bayesian Logic Programs*, In Proceedings of ACL 2012.

[3] Cowie, J., and Lehnert, W. ,*Information extraction*, 1996 CACM.

[4] Kersting, K., and De Raedt, L., *Bayesian Logic Programming: Theory and tool*, 2007

[5] In Getoor, L., and Taskar, B, *Introduction to Statistical Relational Learning*, Cambridge, MA: MIT Press.