Udacity Capstone Project (Proposal)

Machine Learning Engineer Nanodegree

Kiyohito Kunii

March 2020

1 Domain Background

Reinforcement Learning (RL) has been applied and proven to be successful in many domains, and there has been a number of RL research using an simulation environment provided by OpenAI, an independent AI research institute. OpenAI provides a RL simulation environment called OpenAI Gym, a tool to help researchers develop RL algorithms.

One of the recent breakthroughs in the RL field is deep reinforcement learning (DRL), one of which applies a convolutional neural network to a variant of Q-learning to solve different Atari games [3]. Since then, there has been a number of innovation with DRL, including the famous AlgaGo [4]. My motivation is to explore the field of DRL and have a better understanding of some of the established algorithms through this capstone project.

2 Problem Statement

The problem is to solve a very simple CartPole environment (see Figure 1) by training a reinforcement learning agent.

In CartPole environment, a pole is attached by an un-actuated joint to a cart, which moves along a frictionless track. The pole starts upright, and the objective is to keep it upright to avoid failing over by moving the cart in either right or left direction.

While it is tempting to explore more complicated RL environments, such as Atari or MuJoCo Physics engine [1], because it is visually interesting, I chose CartPole for the following reasons:

1. CartPole is a very simple environment and therefore lets me iterate the experiments very quickly, allowing me to focus on learning fundamentals.

2. RL is completely new to me, so I decided to focus on building the basics which is crucial to explore more complex and realistic environments in the future.

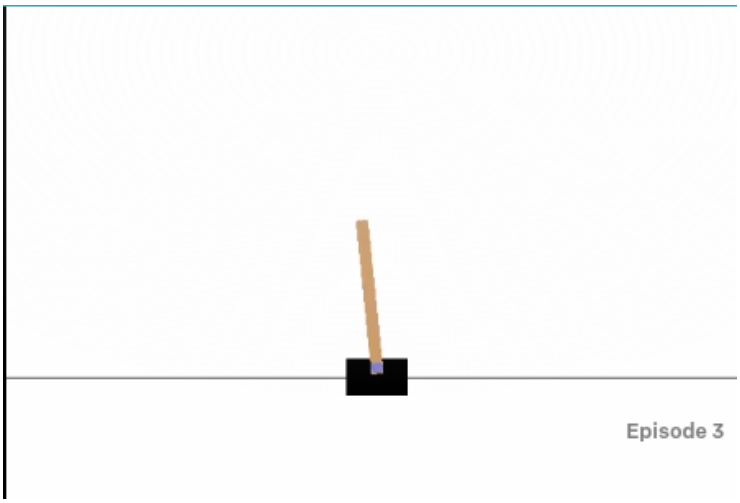

Figure 1: CartPole simulation example

Given the recent advancement in deep reinforcement learning (DRL), I chose to explore various foundations for DRL algorithms to solve this problem. While there are a number of algorithms to explore, I am planning to learn a vanilla policy gradient methods called REINFORCE (Monte-Carlo policy gradient), and Actor-Critic method, since these are good starting points for state-of-art DRL algorithms such as DQN (Deep Q-learning) [3], A3C (Asynchronous Advantage Actor Critic) [2], PPO (Proximal Policy Optimization Algorithms)[**?**], and DDPG (Deep Deterministic Policy Gradients) [1].

3 Datasets and Inputs

Unlike most of machine learning methods, reinforcement learnig does not have a starting dataset and the agent's experience, in the form of rewards, actions and state space, is the dataset to train the agent.

The inputs to the agent is the action to take in the environment. In the case of CartPole, the actions either 0 (push cart to the left) or 1 (push cart to the right).

4 Benchmark Model

As specified in the Github repository for OpenAi Gym[2], the CartPole is considered solved when the average reward is greater than or equal to 195.0 over 100 consecutive trials.

[1] http://www.mujoco.org/

[2] https://github.com/openai/gym/wiki/CartPole-v0

5 Evaluation Metrics

There are two evaluation metrics to use in this project. The first metrics is the convergence to the optimal policy measured by the average rewards over episodes. Another evaluation is to monitor the performance of the trained and I could upload the model to the OpenAI Gym leader board to get the ranking of the performance.

6 Project Design

The policy gradient methods are a way to learn a parametrized policy which select actions without relying on a value function. While vanilla policy gradient method does not rely on the value function, the value function could be used to evaluate the performance of the policy, and the method is called actor-critic method; Actor is the policy model and critic is value function.

I am planning to conduct the experiments using the following technology:

1. OpenAI gym library for simulation environment
2. PyTorch and Numpy for building non-linear function approximation
3. Matplotlib for visualising and reporting the experiment

References

[1] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[2] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[4] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.