ECE 276C Assignment 4

Jeff Wang, Michelle Sit, Akanimoh Adeleye

Problem Statement

A fundamental drawback in current Reinforcement Learning techniques is sampling efficiency. The famous deep Q-network (DQN) results by DeepMind, for example, require tens of millions of frames of training to acheive human-level perfomance [2]. You can imagine, however, that a human (who has never played the game before) would only require a few thousand frames to understand and play the game.

Accordingly, there has been a large amount of work focused on increasing the sample efficiency of DRL methods, of which Generalized Advantage Estimation (GAE) is one [3]. GAE reduces the variance and bias of policy estimates, thus improving the learning process.

Project Idea

In [3], the authors demonstrate that GAE produces improved performance on continuous control enviroments. In our baseline implementation, we extended the algorithm to work on discrete control environments, such as the Atari gym environments. For our project, we will further optimize our implementation in these environments.

As noted below, one of the limitations of GAE is that it still takes a considerable amount of time to converge and learn. Since 2016, when [3] was released, researchers have developed other novel policy gradient methods that claim they outperform other earlier algorithms. Our plan is to apply newer approaches such as PPO and ACKTR to GAE to try to improve the convergence and learning times.

Metrics for Evaluating Improvement

Our algorithms would often take upwards of 80,000 iterations to begin to show postive rewards. We tested 6 different environments to get a baseline of what the rewards patterns look like for GAE. As a result, we can compare the average number of iterations it takes to reach the breakeven reward (ie, the first time the reward hits 0).

Closest State-of-the-art Algorithm

The High-Dimensional Continuous Control Using Generalized Advantage Estimation (GAE) paper proposes a policy gradient method with two novel features. The first is a value function called generalized advantage estimation (GAE) that reduces the variance of policy estimates but introduces some bias by downweighting the rewards from delayed effects. The second are trust region optimizations for the policy and value functions. In essence, the algorithm to able to reduce variance and bias in the gradient and enables it to effectively learn policies for complex control problems that have high sample complexities.

One limitation of the algorithm is that it is not guarenteed to converge. Another is that it takes an extensive amount of time to learn and reach the reward values that it might converge to. The authors propose that a shared function approximation architecture for the policy and value function could potentially solve these issues.

Online Repo

https://github.com/higgsfield/RL-Adventure-2?files=1

References

[1] Fred G. Martin *Robotics Explorations: A Hands-On Introduction to Engineering.* New Jersey: Prentice Hall.

[2] Mnih et. al. *Human-level control through deep reinforcement learning.* Nature: Nature Publishing Group. 2015 http://dx.doi.org/10.1038/nature14236

[3] Schulman et. al. *High-dimensional continuous control using generalized advantage estimation.* arXiv preprint arXiv:1506.02438, 2015b.