Review1: Proximal Policy Optimization Algorithms

Vivswan Shitole

CS539 Embodied AI

1 Summary

This paper introduces a new family of algorithms called Proximal Policy Optimization (PPO) algorithms for policy optimization. PPO algorithms are inspired from the Trust Region Policy Optimization (TRPO) algorithm for policy-based reinforcement learning (RL). The authors claim that their modifications to the TRPO algorithm lead to a family of algorithms (PPO) that are much simpler, more general and have better sample complexity empirically. Experiments are conducted to test PPO on both continuous (MuJoCo) and discrete (ALE) domains, comparing its performance to other policy based RL algorithms.

The main approach is motivated from the fact that TRPO implementation uses the KL divergence constrain on the policy updates as a penalty on the surrogate objective rather than a constrain on the optimization problem because its hard to choose a single value for the KL penalty coefficient β on the constrain. Hence the TRPO algorithm can be improved by modifying the constrain. These modifications lead to two types of surrogate objectives: (1) clipped surrogate objective with a clipping hyperparameter ε and (2) surrogate objective with adaptive KL penalty coefficient β. The clipped surrogate objective is found to perform better (via experiments specified later). The final formulation of PPO uses a linear combination of the clipped loss function for the policy surrogate, a value function error term and an entropy bonus to encourage exploration. The updates are performed in actor-critic style, where the actor collects T timesteps of data using the current policy and the critic formulates the surrogate loss using the collected data. The surrogate objective is optimized with minibatch SGD for K epochs.

The conducted experiments and their results are divided into 4 categories: (1) Comparison between different surrogate objectives for policy optimization over 7 simulated robotics tasks on MuJoCo. These surrogate objectives include objectives with no clipping or penalty (vanilla policy gradient methods), objectives with fixed KL penalty (TRPO), objectives with adaptive KL penalty (PPO) and objectives with clipping (PPO). The algorithm with clipped surrogate objective (clipping hyperparameter $\varepsilon = 0.2$) yields the best average normalized score over the 7 tasks. (2) Comparison with other policy based RL algorithms over 7 simulated robotics tasks on MuJoCo. These policy based RL algorithms include TRPO, cross-entropy method (CEM), vanilla policy gradient with adaptive step size, A2C and A2C with trust region. PPO outperforms other policy based RL algorithms over 6 of the 7 tasks. (3) Showcasing performance of PPO on Humanoid running and steering (high-dimensional continuous domain). PPO converges to yield optimal policies within 50-100 million timesteps. (4) Comparison to tuned A2C and Acer algorithms over 49 Atari games (discrete domain). PPO achieves the highest average episode reward over all the training. But Acer outperforms PPO on the average episode reward over last 100 episodes.

2 Strengths

The PPO algorithms are a significant improvement over TRPO and most of the other existing policy based RL algorithms. The paper validates this improvement via a significant number of empirical experiments, comparing the performance of PPO with the other policy based methods over continuous and discrete domains. The improvement in performance is significant for the rather simple modification over TRPO (of clipping the surrogate objective). However, this simple modification is a result of a critical insight that much of the TRPO's complexity and sample inefficiency comes from the fixed KL divergence penalty coefficient. By clipping the surrogate objective, we achieve a simple first order optimization algorithm that emulates the monotonic improvement of TRPO and is better in performance and sample efficiency. PPO as an overall better policy optimization algorithm (in terms of simplicity, performance and sample efficiency) is a beneficial step for research in the RL community.

3 Weaknesses

PPO algorithms can be considered as a minor variation of the TRPO algorithm. The authors mention that PPO is a simpler algorithm than TRPO since PPO is a first order optimization algorithm. But the authors give no details on TRPO being a second order optimization method. The final formulation of PPO is stated to use a linear combination of the clipped loss function for the policy surrogate, a value function error term and an entropy bonus to encourage exploration. But for the conducted experiments, the value function error term and the entropy bonus is omitted from the loss function. In the experiments conducted over discrete domain (ALE), Acer outperforms PPO on the average episode reward over last 100 episodes. Moreover, there is no comparison with the A3C algorithm which is a major contender in policy based RL algorithms. The optimal value of the clipping hyperparameter $\varepsilon = 0.2$ is obtained by tuning over 7 tasks in continuous domain (MuJoCo). The authors do not state or validate if the same value of ε will work for other tasks and domains.

4 Reflections

The paper relates to the TRPO paper in particular and policy optimization methods in general. The experiments provide a whirlwind tour of the performance of all major policy based RL algorithms. For further research, this paper has given me a new direction of comparing the performance of the major policy based RL algorithms in the Embodied Agent setting. It will be interesting to see if PPO continues to be the winner in the embodied agent setting.

5 Most Interesting Thought

The most interesting thought from the paper for me was that a minor modification to an existing algorithm backed by a deep theoretical insight can lead to significant improvements. Moreover, it is one of those instances where an engineered modification (clipping the surrogate objective) turned out to yield better results than a learnt adaptive modification (adaptive KL penalty coefficient).