Proximal Policy Optimization Algorithms

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

This paper provides a brief review of [1].

1 Review

The authors introduce proximal policy optimization (PPO) which showcases the data efficiency and reliable performance of trust region policy optimization (TRPO) but is simpler to implement, robust against network stochasticity such as dropout, and more generalized.

In TRPO, a surrogate objective function maximizing the current policy parameters with respect to the old policy parameters is subjected to a KL divergence constraint, which limits the policy update step size. This is approximately solved using the conjugate gradient algorithm which doesn't implement a first-order optimization.

1.1 Clipped Surrogate Objective

Therefore, the Clipped Surrogate Objective modifies the surrogate objective function itself by limiting the probability ratio to an interval [1 - ϵ, 1 + ϵ], where ϵ is a hyperparameter (this is called clipping), thereby penalizing policy changes that move the probability ratio away from its maximum value, i.e., 1. Hence, it turns into an unconstrained satisfaction problem solved using only first-order optimization techniques.

Experiments show clipping outperforms previous tuned methods such as TRPO, A2C with Trust Region, etc on most continuous control environments. It also outperformed fine-tuned ACER and A2C in the Atari Domain in terms of fast policy learning.

1.2 Adaptive KL Penalty Coefficient

The surrogate objective can also be modified by penalizing a large KL divergence by a certain penalty coefficient β, thereby forming a pessimistic bound on policy performance. This penalty coefficient adapts after every policy update in order to achieve the required value of the KL divergence.

Experiments show adaptive KL performs worse than clipping.

References

[1] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv. https://doi.org/10.48550/arXiv.1707.06347

Preprint. Under review.