Thoughts on Implementation Matters in Deep RL: A Case Study on PPO and TRPO

Owen Lockwood

July 31, 2020

In short, this paper asserts that the advantage of Proximal Policy Optimization (PPO) over Trust Region Policy Optimization (TRPO) is largely due to Code Level Implementations (CLI). This is part of a growing body of work that questions the differences in algorithms for deep RL. This is something that I think is critically important. Very little in deep RL, or deep learning as a whole, is presented as a provably better or faster alternative. Rather we just see better results so we think the algorithm is better. However, the serial problem is that these algorithms are not compared evenly. Learning rates, layers, parameters, initializations, etc. all differ on the whim of the implementer. This causes even more problems in deep RL because of its extreme brittleness. I think this paper provides good ablation studies with decent mechanistic explanations. Hopefully this will lead to more development of shared standards and more equal comparisons; however, ease of access will be a challenge. You can make all the standards you want but if you make it difficult to work with or requiring computational power many lack then it will hinder the field more than helping.

I imagine this is also one of the big problems with applying deep RL to real world problems. The results based evaluations (rather than on solid theoretical improvements) leads to hesitations in applying deep RL to problems where failure could cost money/lifes. No one wants a brittle algorithm crashing their self driving car or algorithms for their energy grids that could fail for unknown theoretical reasons. Solving these problems will be integral to the success of deep RL.