Paper Review - "*CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING*"

Summary: The authors present a new model-free actor-critic algorithm that can operate over continuous action spaces. The motivation comes from the advances of deep reinforcement learning on discrete action spaces. With inspiration from DQN, they combine it with the DPG algorithm to get a robust enough model which performs well where the action space is continuous and high-dimensional. This expands the range of problems that can be tackled using reinforcement learning.

Main Contributions: The paper has four main contributions. First, they use a replay buffer like in DQn to make use of efficient hardware optimizations. This replay buffer allows the algorithm to benefit from learning across a set of uncorrelated transitions as DDPG is an off-policy algorithm. Second, they modified the actor-critic method for using 'soft' target updates rather than directly copying weights. This was achieved by having a copy of the actor and critic networks, which helped in improving the stability of learning. Third, they used batch normalization which helped the model generalize over many different tasks. Fourth, they overcome the problem of learning in continuous action spaces by adding noise to the actor policy

Strengths: The main strength of the DDPG algorithm is that it is able to work on both low-dimensional and high-dimensional state space with high-dimensional pixel space being the differentiator from the other algorithms. Also, it can be generalized well across many different tasks with keeping the same set of hyperparameters which makes it more robust.

Weakness: With the introduction of deep learning into DPG here, hyperparameter tuning becomes of utmost importance. We need to arrive at the best set of hyperparameters for optimal performance which is also highlighted in the paper. Another weakness is that although the network learns in a stable fashion, the soft target updates make the learning slow since the target network delays the propagation of the value estimates. Hence it requires a large number of training episodes.

Experiments: They evaluated the model on a very wide variety of tasks and levels of difficulty. These include classical problems like cartpole, high-dimensional tasks like gripper, and robotics tasks like cheetah and puck striking. All the tasks were evaluated both on low-dimensional state space (joint angles) and high-dimensional renderings of the environment. The problems were made fully observable and compared with DPG. They also tested the DDPG algorithm with and without noise, target network, or batch normalization. Results showed that with noise and target networks with BN always outperformed DPG in the low-dimensional state space and the learning with high-dimensional was at par in most of the applications. It also used fewer steps than DQN.

Extensions: Something similar to guided policy search of using spatial softmax to reduce the dimensionality of visual features and policy also receiving low-dimensional state information directly in the first fully connected layer may increase the power and data efficiency of the DDPG algorithm.