Paper Review - “*Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning, Nagabandi et al, 2017*”

Summary: The paper introduces a novel approach that combines the sample efficiency of model-based reinforcement learning with the high task-specific performance of model-free methods. To achieve this, a deep neural network is utilized to approximate the dynamics of the environment, and the resulting model is used to generate trajectories that are then used to train a model-free policy. The authors demonstrate the effectiveness of this approach on a range of challenging tasks, including locomotion tasks on the MuJoCo physics simulator, achieving good performance in these tasks.

Main Contributions: One of the main contributions of this work is to demonstrate the effectiveness of model-based reinforcement learning with neural network models for several contact-rich simulated locomotion tasks. The model-based trajectory optimizer generates trajectories by simulating future states and rewards using the neural network dynamics model. The optimizer can use different techniques, including model predictive control (MPC) or stochastic trajectory optimization, to generate the most promising trajectories based on the current state and the predicted future states. The generated trajectories are then used to train a model-free reinforcement learning algorithm using the Trust Region Policy Optimization (TRPO) algorithm. The model-free algorithm learns a policy that maximizes the expected reward based on the generated trajectories.

Strengths: One of the main strengths of the paper is that it combines both the Combines Model-Based and Model-Free approaches, i.e. best of both worlds. For many of the locomotion tasks shown in the paper, the MB-MF approach indicates a better sample complexity than the other models. Also, they use TRPO for model-free so as to show how their approach beats the SOTA, which was a smart move from the authors.

Weakness: One of the limitations of this approach is that for long horizons and high-dimensional action spaces, random sampling with MPC may not be sufficient and would require a large number of actions to be sampled. Due to this, it may not be feasible on real systems with limited computational power.

Experiments: Authors do multiple experiments. First, they evaluate their design decisions like the number of training steps, data aggregation, the controller used, and the number of initial random trajectories required for model-based RL. They evaluated it on swimmer and half-cheetah agents on the locomotion task of running forward quickly. Second, they evaluate the model-based RL approach on different tasks. They show that the model trained only once per agent was general enough to accommodate new tasks at test time, including more complex ones hence it is powerful across a variety of tasks. Third, they evaluate the mBMF approach. Authors first state that the rewards of the pure model-based approach were not sufficient to beat SOTA model-free approaches but the MBMF approach beats it on locomotive tasks in fewer steps.

Extensions: As a future direction one can integrate the MBMF model with different model free-learners like Q-learning and actor-critic methods for further gains on sample efficiency. Another interesting application of this approach is to test it out on real-world applications like robotics since the improved sample efficiency makes it practical to use it under the constraints of real-time sample collection.