Review-40

Neural Dynamic Policies for End-to-End Sensorimotor Learning

Imitation and Reinforcement Learning provision the learning of policies in raw action spaces which force the agent to make decisions at each timestep. This hinders scalability of agent to continuous and high-dimensional action spaces. The work aims to address this shortcoming of the learning paradigm by learning policies in the trajectory space. Towards this goal, the work introduces Neural Dynamic Policies (NDPs) which make use of a dynamical system-based framework within the deep neural network policy of the agent. NDPs reparameterize the action space with nonlinear differential equations which allows the policy to reason about actions only at sampling intervals. The dynamical structure of NDPs induces a smooth vector field in trajectory space, hence augmenting agents into sample-efficient learners.

NDPs aim to answer the pivotal question of *whether a dynamical system can be embedded into the robotic agent's policy to describe its behavior?*. To answer this central question, the work reparameterizes the action space in a deep policy network by virtues of nonlinear differential equations. The second-order differential equation structure of Dynamic Movement Primitives (DMPs) is utilized with a desired goal g and weights w a nonlinear forcing function f as its parameters. The NDP network predicts w and g based on the input state. Predicted parameters are further used to solve for dynamic system states $\{y,\dot{y},\ddot{y}\}$ which yield the final action using an inverse controller. In the imitation learning setting, NDPs are trained using behavior cloning between the demonstrated action sequence and the NDP policy. In the Reinforcement Learning setting, NDPs may be trained using any underlying RL algorithm by utilizing the NDP procedure for a finite number of integration steps. The NDP policy is used once every finite k steps which require k-times fewer forward passes for action sampling.

NDPs, when combined with PPO, demonstrate sample-efficient learning in comparison to strong baselines including the multi-action critic architecture for PPO. Corresponidng to both imitation and reinforcement learning, NDPs present higher task success rates and improved quality of trajectories depicted on the digital writing task. Additionally, the work provides a rigorous ablation study of NDP's components and its variation with different parameter values which demonstrate suitability of the proposed method. However, the experimental setup demonstrates a few caveats which may be further improved. Firstly, the work does not compare NDPs to strong baselines such as TD3 and SAC which present state-of-the-art performance. This leaves the adaptation of dynamical systems in policies an open question from the perspective of improved performance. Secondly, NDPs do not sufficiently present scalability to high-dimensional action spaces, the motivating problem for this work. A better experiment design would consist of evaluating NDPs on tasks such as Humanoid which present a larger number of action configurations.