Offline vs. Online Learning for Deep Driving from Demonstrations

Michael Laskey (PhD Candidate) and Ken Goldberg, UC Berkeley

Keywords: Deep Learning from Demonstration, Imitation Learning, Robotics

**DART: Reduce Covariate Shift by Sampling from Current Robot’s Policy**

1.) Initial Traj. 2.) Rollout Robot
3.) Get Feedback 4.) Update Robot

We are studying alternative learning methods for Driving from Demonstrations. The focus on offline learning dating back almost 30 years has been largely supplanted by online learning methods such as DAgger, which focus on difficult-to-learn areas of state space and require tedious and error-prone corrective feedback from human supervisors. Our initial results suggest that offline learning may be preferable for highly-expressive policy classes such as Deep Learning.

One approach to Imitation Learning is Behavior Cloning, in which a robot observes a supervisor and infers a control policy. A known problem with this “off-policy” (Offline) approach is that the robot’s errors compound when drifting away from the supervisor’s demonstrations. On-policy (Online) Imitation Learning alleviate this by iteratively collecting corrective actions for the current robot policy. However, these techniques can be tedious for human supervisors, add significant computation burden, and may visit dangerous states during training. We propose a new off-policy (Offline) approach that injects noise into the supervisor’s policy while demonstrating. This forces the supervisor to demonstrate how to recover from errors. We propose a new
algorithm, DART (Disturbances for Augmenting Robot Trajectories), that collects demonstrations with injected noise, and optimizes the noise level to approximate the error of the robot’s trained policy during data collection. In a paper published in the 1st Conference on Robot Learning, we compared DART with DAgger and Behavior Cloning in two domains: in simulation with an algorithmic supervisor on the MuJoCo tasks (Walker, Humanoid, Hopper, Half-Cheetah) and in physical experiments with human supervisors training a Toyota HSR robot to perform grasping in clutter. For high dimensional tasks like Humanoid, we found that DART can be up to 3x faster in computation time and only decreases the supervisor’s cumulative reward by 5% during training, whereas DAgger executes policies that have 80% less cumulative reward than the supervisor. On a grasping in clutter task, DART obtains on average a 62% performance increase over Behavior Cloning. We are now studying how DART performs on driving tasks and developing a new driving simulator to facilitate experiments.