Video-Conditioned Multi-Task Imitation Learning for Robotic Systems: Enhancing Robustness Through Object-Centric Reasoning

This thesis explores advancements in robotic control for complex, multi-task environments, focusing on a technique called Learning from Demonstration (LfD). The main challenge addressed is enabling robots to autonomously perform a wide range of tasks based on video demonstrations. In this approach, the robot receives a demonstration performed by another agent in a different setting and must interpret the input video to complete the task in its own environment, which may have different configurations.

The control policy is trained to handle various variations of a specific task (e.g., picking an object from different possible locations) or even entirely different tasks (e.g., a single control policy that solves both picking and placing tasks as well as assembly tasks). The goal is to generalize not only with respect to the objects being manipulated and the initial conditions but also with respect to the tasks themselves. In this context, a key issue identified in the current state-of-the-art methods is related to target misindentification, where the learned control policy generates valid trajectories, enabling the robot to reach, pick, and place objects, but frequently manipulates the wrong object.

To solve this problem, two main considerations were made: (1) Architectures proposed in the current literature are predominantly end-to-end, translating high-dimensional inputs, such as images, into corresponding low-dimensional actions. As a result, the model must learn an implicit representation that encodes both the task objective and the current state of the environment, including the location of the target object. (2) The learning procedure optimizes an **action-centric** metric, meaning that it is not directly linked to task success but instead focuses on mimicking the expert's actions on average. This action-focused optimization can lead to poor encoding of critical information, such as object positions.

To address these limitations, the thesis proposes a modular system that divides manipulation tasks into two subcomponents: (1) a cognitive task focused on identifying target locations, and (2) a control task that generates the required actions.

To solve the cognitive task, a Conditioned Object Detector (COD) has been developed. This module, given the video demonstration and the current agent observation as input, predicts the category-agnostic bounding box related to the target object and the final placing location. This low-level positional information is then provided to the control module, which predicts the actions to perform.

To solve the control task the Object Conditioned Control Policy (OCCP) has been proposed, which focuses on solving the control problem using low-level positional information that can be easily mapped into the corresponding actions.

The final system has been extensively tested in simulation environments, then it was also validated on a real-world robotic platform.

The proposed methods showed promising performance, with significant improvements over baseline methods that lack object-oriented reasoning. This modular, object-oriented approach proved effective for LfD tasks, enhancing both interpretability and reliability, as the identified object locations (bounding boxes) clarify the robot’s intended movements for the end user.

In real-world tests, the proposed method demonstrated robustness even with limited, noisy data collected through teleoperation, underscoring the feasibility of using object-focused approaches in real environments. These results confirm that incorporating object-awareness can support reliable and adaptable robotic systems, even under challenging conditions with sparse data.