Third-Person Visual Imitation Learning via Decoupled Hierarchical Control

Pratyusha Sharma, Deepak Pathak, Abhinav Gupta

Carnegie Mellon University
The Robotics Institute
Problem / Goal

Can our robot manipulate a new object given a single human video alone?
Why is it hard?

- Inferring useful information in the video
- Handling domain shift
- Every major part of the sequence needs to be executed correctly - Ex: For pouring, it needs to reach the cup before twisting its hand
- The manipulation is challenging. (6D, novel objects and positioning, no force feedback)
Scenario 1:

Issue: Not closed loop. No understanding of how the positions of the objects placed in front of the robot change with time!
Issue

Scenario 2:

Input: Human Demo + Robot visual state

Output

Sequentially predict the states of the robot arm

How do we force it to use task information from Human demonstration alone but condition its action on current observable state?
We want to build a model that can *infer the intent from the Human Demonstration* of a task and *act in the Robot’s current environment* to then accomplish the task.
We decouple the task of **Goal Inference** from **Local Control**.
Training and Test Scenarios - Data Availability

Training

- Human demo video
- Robot demo video
- Robot joint angles

Test (deployment)

- Human demo video
- Current visible image of the table
Approach - Training

**Goal Generator:** Given human demo and present visual state of the robot we hallucinate the next step.
Approach - Training

**Goal Generator**: Given human demo and present visual state of the robot we hallucinate the next step.

**Inverse Model**: Use the hallucinated prediction with the current visual state to predict the action!
Train Time: The **Goal Generator** and **Inverse Model** are trained separately

Test Time: The **Goal Generator** and **Inverse Model** are executed alternatingly
Approach - Test
Experiments and Results

We evaluate the models trained as follows:

- Goal generation model with a perfect inverse model
- Inverse model with a perfect goal generation model
- Goal generation model and inverse model in tandem
Results: Goal generation model with perfect inverse model
Results: Inverse model with perfect goal generator

GT trajectory

Predicted trajectory from GT-images
Results: Final experiment runs
Results: Final Experimental Runs: Placing in a box
Shortcomings:

1. Robot trajectory is shaky: The robot trajectory looks shaky because of the absence of any temporal knowledge. Though trajectories predicted by inverse models with memory units (LSTM) look far less shaky but the models then over fit to the task.
Thank you!