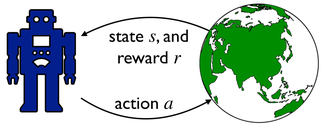
Scaling up Robot Learning by Understanding Videos


Matthew Chang Arjun Gupta Aditya Prakash Mohit Goyal Sahil Modi Rishabh Goyal

Saurabh Gupta

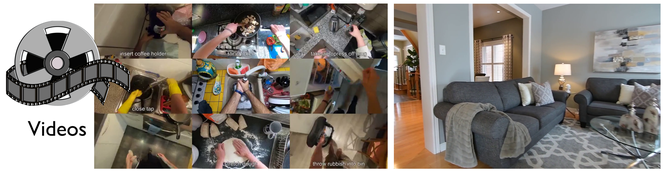
Overview

Policy learning from interaction is challenging



- Challenging to specify reward functions
- Impractically large sample complexity
- Learning signal derived solely from interaction
- Poor generalization due to lack of visual diversity in training, sim2real transfer

Videos can aid

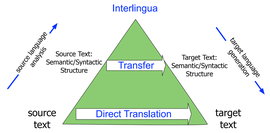


- Large diversity may provide good generalization.
- Demonstrations may directly show how to solve long horizon tasks.
- Depict what the world is like, and how it works.

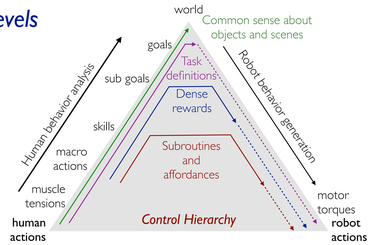
However,

- Videos don't come with action labels
- Goals and intents are not known
- Depicted trajectories may be sub-optimal
- Embodiment gap (sensors / actions / capabilities)
- Only showcase positive data

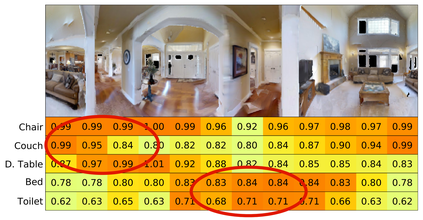
Learning at different abstraction levels



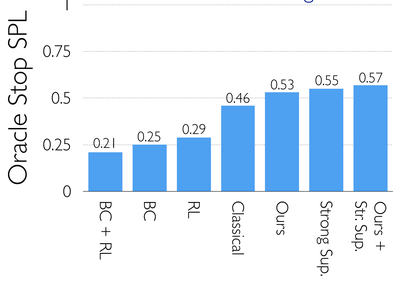
The Vauquois triangle for machine translation



A. Value Learning from Videos

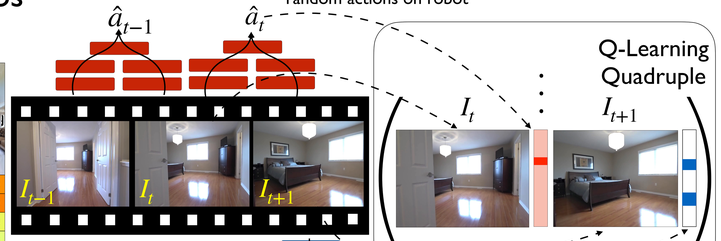


Predictions for nearness to goal



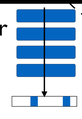
a) Action Grounding

Inverse model built by executing random actions on robot



b) Goal Labeling

Object Detector trained on COCO



r^c_{t+1}

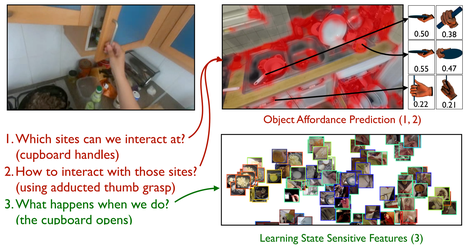
c) Q-Learning

Value function that predicts nearness to goal: $f(I, c) = \max_a Q^*(I, a, c)$

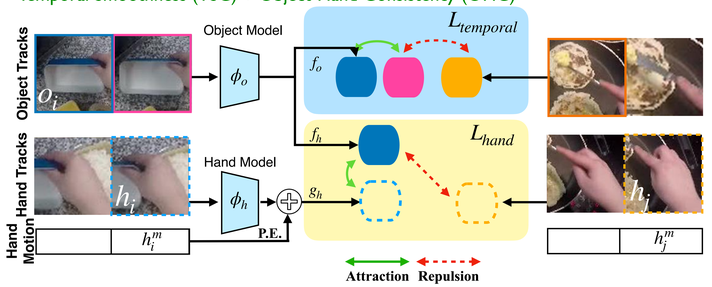
- Stronger than behavior cloning on videos and BC + RL
- Stronger than even RL methods trained with dense rewards with 250x more interaction samples and 6x more environments with direct interaction access
- Better than strong exploration baselines
- Improves performance when combined with strongly supervised model

B. Human Hands as Probes for Interactive Object Understanding

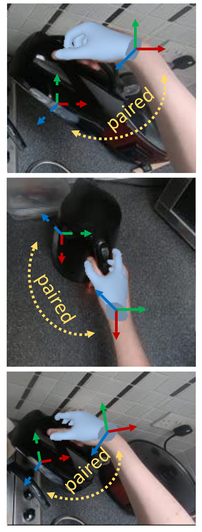
We develop *interactive object understanding* from the *natural ways* in which people interact with objects in *egocentric videos*.



Learning features using:
Temporal Smoothness (TSC) + Object-Hand Consistency (OHC)



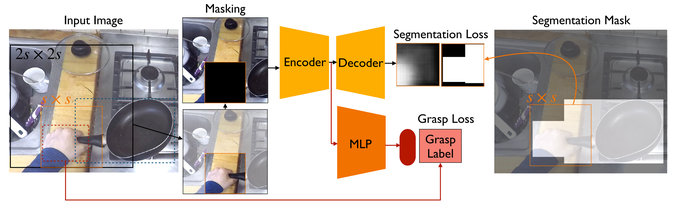
Also: Hand pose as a proxy for object pose



Human hands provide cues for interactive object understanding

- Attending to hands localizes and stabilizes active objects.
- Hands show where all we can interact in the scene.
- Analyzing hands reveals information about objects: state and how to interact.
- Hand pose may provide hints for object pose.

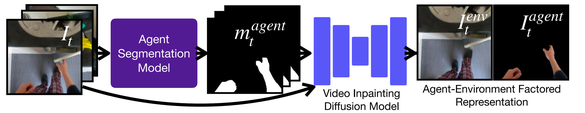
Affordances via Context Prediction (ACP):
Predicting Region of Interaction (ROI) and Grasps afforded by Objects (GAO)

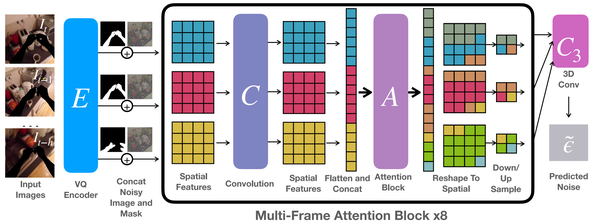


C. Look Ma, No Hands! Agent-Environment Factorization of Egocentric Videos

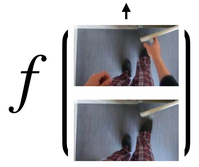
Hands cause occlusion and differ in appearance from robotic end-effectors. But, they also provide useful signal.

We extract a factored representation of the scene that separates the agent (human hand) and the environment.





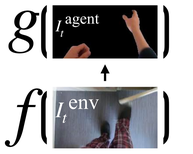
a) Visual Perception



I^{env}_t used in addition to the original image.

mAR : 0.26 → 0.38

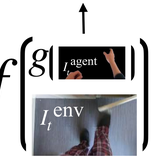
b) Affordances



I^{env}_t used to predict relevant aspects of I^{agent}

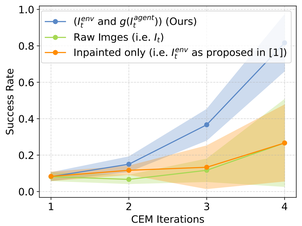
Acc : 0.35 → 0.41

c) Reward Functions



The agent representation transformed into agent-agnostic format

ρ : 0.56 → 0.61



Drawer opening policy trained in the in the real world. We find that using the agent-environment factored representation improves learning speed.