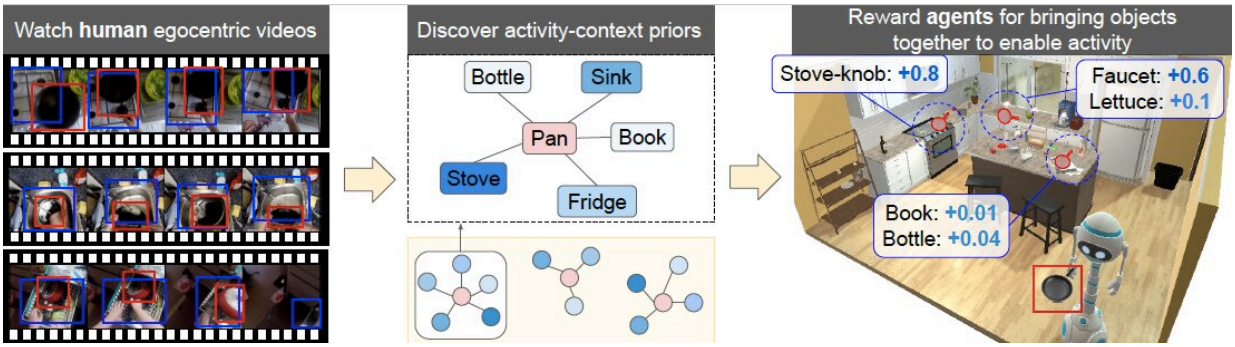
A model that translates everyday human activities into skills for an embodied artificial agent

November 3 2021, by Ingrid Fadelli



The main idea behind the researchers' paper. Left and middle panel: The team discovered activity-contexts for objects directly from egocentric video of human activity. A given object's activity-context goes beyond "what objects are found together" to capture the likelihood that each other object in the environment participates in activities involving it (i.e., "what objects together enable action"). Right panel: The team's approach guides agents to bring compatible objects—objects with high likelihood—together to enable activities. For example, bringing a pan to the sink increases the value of faucet interactions, but bringing it to the table has little effect on interactions with a book. Credit: Nagarajan & Grauman.

Over the past decade or so, many roboticists and computer scientists have been trying to develop robots that can complete tasks in spaces populated by humans; for instance, helping users to cook, clean and tidy

up. To tackle household chores and other manual tasks, robots should be able to solve complex planning tasks that involve navigating environments and interacting with objects following specific sequences.

While some techniques for solving these complex planning tasks have achieved promising results, most of them are not fully equipped to tackle them. As a result, robots cannot yet complete these tasks as well as human agents.

Researchers at UT Austin and Facebook AI Research have recently developed a new framework that could shape the behavior of embodied agents more effectively, using ego-centric videos of humans completing everyday tasks. Their paper, pre-published on arXiv and set to be presented at the Neural Information Processing Systems (NeurIPS) Conference in December, introduces a more efficient approach for training robots to complete household chores and other interaction-heavy tasks.

"The overreaching goal of this project was to build embodied robotic agents that can learn by watching people interact with their surroundings," Tushar Nagarajan, one of the researchers who carried out the study, told TechXplore. "Reinforcement learning (RL) approaches require millions of attempts to learn intelligent behavior as agents begin by randomly attempting actions, while imitation learning (IL) approaches require experts to control and demonstrate ideal agent behavior, which is costly to collect and requires extra hardware."

In contrast with robotic systems, when entering a new environment, humans can effortlessly complete tasks that involve different objects. Nagarajan and his colleague Kristen Grauman thus set out to investigate whether embodied agents could learn to complete tasks in similar environments simply by observing how humans behave.

Rather than training agents using video demonstrations labeled by humans, which are often expensive to collect, the researchers wanted to leverage egocentric (first-person) video footage showing people performing everyday activities, such as cooking a meal or washing dishes. These videos are easier to collect and more readily accessible than annotated demonstrations.

"Our work is the first to use free-form human-generated video captured in the real world to learn priors for object interactions," Nagarajan said. "Our approach converts egocentric video of humans interacting with their surroundings into 'activity-context' priors, which capture what objects, when brought together, enable activities. For example, watching humans do the dishes suggests that utensils, dish soap and a sponge are good objects to have before turning on the faucet at the sink."

To acquire these 'priors' (e.g., useful information about what objects to gather before completing a task), the model created by Nagarajan and Grauman accumulates statistics about pairs of objects that humans tend to use during specific activities. Their model directly detected these objects in ego-centric videos from the large dataset used by the researchers.

Subsequently, the model encoded the priors it acquired as a reward in a reinforcement learning framework. Essentially, this means that an agent is rewarded based on what objects it selected for completing a given task .

"For example, turning-on the faucet is given a high reward when a pan is brought near the sink (and a low reward if, say, a book is brought near it)," Nagarajan explained. "As a consequence, an agent must intelligently bring the right set of objects to the right locations before attempting interactions with objects, in order to maximize their reward. This helps them reach states that lead to activities, which speeds up learning."

Previous studies have tried to accelerate robot policy learning using similar reward functions. However, typically these are exploration rewards that encourage agents to explore new locations or perform new interactions, without specifically considering the human tasks they are learning to complete.

"Our formulation improves on these previous approaches by aligning the rewards with human activities, helping agents explore more relevant object interactions," Nagarajan said. "Our work is also unique in that it learns priors about object interactions from free-form video, rather than video tied to specific goals (as in behavior cloning). The result is a general-purpose auxiliary reward to encourage efficient RL."

In contrast with priors considered by previously developed approaches, the priors considered by the researchers' model also capture how objects are related in the context of actions that the robot is learning to perform, rather than merely their physical co-occurrence (e.g., spoons can be found near knives) or semantic similarity (e.g., potatoes and tomatoes are similar objects).

The researchers evaluated their model using a dataset of ego-centric videos showing humans as they complete everyday chores and tasks in the kitchen. Their results were promising, suggesting that their model could be used to train household robots more effectively than other previously developed techniques.

"Our work is the first to demonstrate that passive video of humans performing daily activities can be used to learn embodied interaction policies," Nagarajan said. "This is a significant achievement, as egocentric video is readily available in large amounts from recent datasets. Our work is a first step towards enabling applications that can learn about how humans perform activities (without the need for costly demonstrations) and then offer assistance in the home-robotics setting."

In the future, the new framework developed by this team of researchers could be used to train a variety of physical robots to complete a variety of simple everyday tasks. In addition, it could be used to train and augmented reality (AR) assistants, which could, for instance, observe how a human cooks a specific dish and then teach new users to prepare it.

"Our research is an important step towards learning by watching humans, as it captures simple, yet powerful priors about objects involved in activities," Nagarajan added. "However, there are other meaningful things to learn such as: What parts of the environment support activities (scene affordances)? How should objects be manipulated or grasped to use them? Are there important sequences of actions (routines) that can be learned and leveraged by embodied agents? Finally, an important future research direction to pursue is how to take policies learned in simulated environments and deploy them onto mobile robot platforms or AR glasses, in order to build agents that can cooperate with humans in the real world."

More information: Tushar Nagarajan, Kristen Grauman, Shaping embodied agent behavior with activity-context priors from egocentric video. arXiv:2110.07692v1 [cs.CV], arxiv.org/abs/2110.07692

vision.cs.utexas.edu/projects/ego-rewards/