Interactive Visual Task Learning for Robots

Weiwei Gu, Anant Sah, Nakul Gopalan

Arizona State University
weiweigu@asu.edu, asah4@asu.edu, ngopala6@asu.edu

Abstract

We present a demonstrable framework for robots to learn novel visual concepts and visual tasks via in-situ linguistic interactions with human users. Previous approaches in computer vision have either used large pre-trained visual models to infer novel objects zero-shot, or added novel concepts along with their attributes and representations to a concept hierarchy. We extend the approaches that focus on learning visual concept hierarchies and take this ability one step further to demonstrate novel task solving on robots along with the learned visual concepts. To enable a visual concept learner to solve robotics tasks one-shot, we developed two distinct techniques. Firstly, we propose a novel approach, Hi-Viscont(HIerarchical VISual CONcept learner for Task), which augments information of a novel concept, that is being taught, to its parent nodes within a concept hierarchy. This information propagation allows all concepts in a hierarchy to update as novel concepts are taught in a continual learning setting. Secondly, we represent a visual task as a scene graph with language annotations. The scene graph allows us to create novel permutations of a demonstrated task zero-shot in-situ. Combining the two techniques, we present a demonstration on a real robot that learns visual task and concepts in one-shot from in-situ interactions with human users, and generalize to perform a novel visual task of the same type in zero-shot. As shown by the studies in the main conference paper, our system achieves a success rate of 50% on solving the whole task correctly with generalization where the baseline performs at 17% without any ability to generalize to novel tasks and concepts. We will demonstrate our working interactive learning pipeline at AAAI 2024 in person with our robot and other required hardware.

Introduction

Robots in a household will encounter novel objects and tasks all the time. For example, a robot might need to use a novel vegetable peeler to peel potatoes even though it has never seen, let alone used such a peeler before. Our work focuses on teaching robots novel concepts and tasks one-shot via human-robot interactions, which include demonstrations and linguistic explanations. We then want the robot to generalize to a similar but unseen visual task. A robotic system that can learn generalizable tasks and concepts from few natural interactions with a human-user would represent a large leap for robotics applications in everyday settings. In this work we aim to take a step in the direction of generalizable interactive learning as demonstrated Fig. 1.

Previously, large image and language models have been extended to robotics to manipulate novel objects, and create visual scenes (Shridhar, Manuelli, and Fox 2021; Brohan et al. 2023). These methods recognize novel objects by using their underlying large language and visual models to extract task-relevant knowledge. However, they are not capable of learning to create a novel visual scene from an in-situ interaction with a human user. There is also significant work in few-shot learning of visual concepts in computer vision (Mei et al. 2022; Snell, Swersky, and Zemel 2017; Vinyals et al. 2017; Sung et al. 2018; Wang, Ye, and Gupta 2018; Tian et al. 2020), albeit without extensions to robotics domains. These approaches focus on learning novel concepts for image classification, but ignore the fact that the novel concepts also bring new information to update our understanding of concepts already known to the robot. The reverse path of knowledge propagation, that is, from novel concepts to previously known concepts is equivalently important in performing tasks in the real-life scenarios, especially when the agent has little knowledge of the world and needs to continually add information to known concepts.

In this work, we propose a novel framework, Hi-Viscont, that enables robots to learn visual tasks and visual concepts from natural interactions with a human user. We learn the task type and concepts from users one-shot, and then generalize to tasks within the task type zero-shot. We do this by connecting our insights on *one-shot visual concept learning* and the use of *scene graphs*. The robot learns the structure of a visual task by converting linguistic interactions with a human user into a contextualized scene graph with language annotations. Moreover, Hi-Viscont updates parental concepts of the novel concept being taught. Such updates allow us to generalize the use of the novel concepts in to solve novel tasks.

Demonstration Overview

We will show that our pipeline's capability of learning novel visual tasks and novel visual concepts from in-situ interactions with human users in the demonstration. The domain of the demonstration is a house construction domain with building blocks from children's toys. In our demonstration,

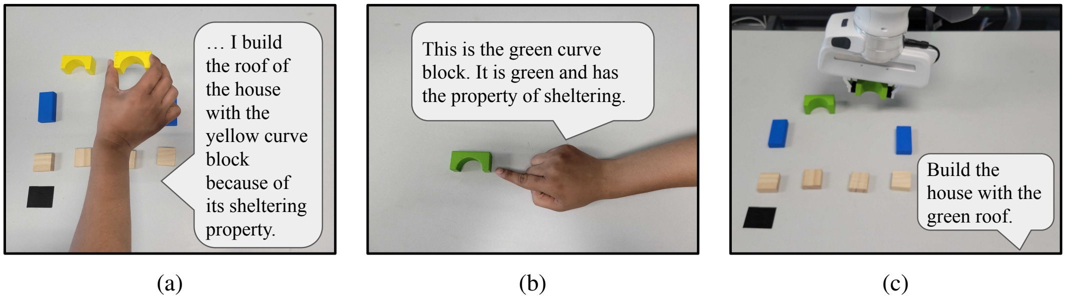
… I build the roof of the house with the yellow curve block because of its sheltering property.
This is the green curve block. It is green and has the property of sheltering.
Build the house with the green roof.
(a)
(b)
(c)

Figure 1: This figure demonstrates how Hi-Viscont learns from users interactively. (a) First the user demonstrates a structure, say a "house," with its sub-components such as its "roof" and the concepts used to make the "roof" such as a "yellow curve block". (b) The user then teaches a novel concept such as a "green curve block" and describes its properties. (c) The user can now ask the robot to create a new structure ("house with green roof") zero-shot with the taught component without explicitly asking for the object of interest.

human users will interact with the robot in three phases, including a task teaching phase, a concept teaching phase, and a request phase, sequentially as described below.

Task Teaching Phase - In the task teaching phase, the users teach the robot a visual task by demonstrating the scene with its constituent structures one by one. The users also describe the structures and the objects used to build the structure with natural language. For example, a user might build a house, with floors, pillars and a roof. While building the roof, the user might say "This a roof which I build with the curved blue tile because of it's sheltering capability." The users demonstrate each of the structures with their chosen language commands one after another to build a house. We record all descriptions in audio and convert them into text using audio to text tools.

Concept Teaching Phase - In this phase, the users teach a novel concept of their choice to both systems by showing the object to the camera, and describing the concept's properties such as the color of the object and the functional characteristics of the object, in natural language. The description to the novel object concept will be converted to neuro-symbolic programs which are given to both models for updates as described in the Methods section.

Request Phase - In the request phase, the users provide a request in natural language for a novel scene that they did not demonstrate in the task teaching phase. They are asked to use the object taught in the concept teaching phase in the request. The task requested still needs to be a house which is the same task type as they demonstrated previously. Albeit it is a house that both models have not seen previously.

After the three interactive phases, the robot will then complete the requested visual task based on the request of the human users in real time. The process of the demonstration is similar to the procedure of our human subject study, in which our system achieved a success rate of 50%. The full video of demonstration with explanation can be found in the associated webpage [1]. More details of the study can be found in the anonymized full paper in the supplementary materials. Additionally, we can understand the corner cases of the robot learning process better by demonstrating our system at the conference, which introduce users from a more complicated demographics than our human subject study.

Setup

We integrate our visual task learning and concept learning model with a Franka Emika Resarch 3 arm(FR3). To set this demonstration up we use a Franka Emika Research 3 arm (FR3), two calibrated realsense D435 depth cameras, and a mono-colored table to allow for background subtraction. *The authors will carry all the required hardware mentioned above to the conference location to demonstrate learning on the real robot at the AAAI 2024 venue.*

Methods

Our pipeline has two major components: Hi-Viscont, a hierarchical visual concept learner, and a scene graph with linguistic annotation.

Hi-Viscont. Hi-Viscont is a visual concept learner that learns a novel visual concept one-shot, and actively updates all related known concepts when a novel concept is introduced. We adopted multiple modules from FALCON(Mei et al. 2022), a meta-learning framework for one-shot concept learning, including the visual feature extractor, the neuro-symbolic program executor, the box embedding space, and the novel concept learner. To improve FALCON's generalizability, we introduce an additional module in Hi-Viscont to update the related concepts when a novel concept is introduced. The additional module enables the robot leverage its knowledge and generalize to an unseen visual task in zero-shot.

Scene Graph. To learn a visual task from a single in-situ interaction with human user, we first convert the user's demonstration (Fig. 1.a) into an initial scene graph. Each node of the initial scene graph corresponds to an object that the user placed, and it contains the bounding box information of the object and the user's linguistic description of the object. Based on the initial scene graph and the user's linguistic request for the desired variant of the visual scene, we infer a goal scene graph using a BERT based classifier.

[1] https://sites.google.com/view/ivtl

References

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; Florence, P.; Fu, C.; Arenas, M. G.; Gopalakrishnan, K.; Han, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ichter, B.; Irpan, A.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, L.; Lee, T.-W. E.; Levine, S.; Lu, Y.; Michalewski, H.; Mordatch, I.; Pertsch, K.; Rao, K.; Reymann, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sermanet, P.; Singh, J.; Singh, A.; Soricut, R.; Tran, H.; Vanhoucke, V.; Vuong, Q.; Wahid, A.; Welker, S.; Wohlhart, P.; Wu, J.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.

Mei, L.; Mao, J.; Wang, Z.; Gan, C.; and Tenenbaum, J. B. 2022. FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic descriptions, and Conceptual Relations. In *International Conference on Learning Representations*.

Shridhar, M.; Manuelli, L.; and Fox, D. 2021. CLIPort: What and Where Pathways for Robotic Manipulation. arXiv:2109.12098.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. arXiv:1703.05175.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. arXiv:1711.06025.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need? arXiv:2003.11539.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2017. Matching Networks for One Shot Learning. arXiv:1606.04080.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs. arXiv:1803.08035.