Bringing the RT-1-X Foundation Model to a SCARA robot

Jonathan Salzer and Arnoud Visser

Intelligent Robotics Lab, Universiteit van Amsterdam, NL
www.intelligentroboticslab.nl

Abstract. Traditional robotic systems require specific training data for each task, environment, and robot form. While recent advancements in machine learning have enabled models to generalize across new tasks and environments, the challenge of adapting these models to entirely new settings remains largely unexplored. This study addresses this by investigating the generalization capabilities of the RT-1-X robotic foundation model to a type of robot unseen during its training: a SCARA robot from UMI-RTX.
Initial experiments reveal that RT-1-X does not generalize zero-shot to the unseen type of robot. However, fine-tuning of the RT-1-X model by demonstration allows the robot to learn a pickup task which was part of the foundation model (but learned for another type of robot). When the robot is presented with an object that is included in the foundation model but not in the fine-tuning dataset, it demonstrates that only the skill, but not the object-specific knowledge, has been transferred.

Keywords: Imitation learning · Tokenization · Conditioning.

1 Introduction

Recent breakthroughs in machine learning and artificial intelligence suggest that training on large, diverse datasets can lead to highly adaptable models, which often exceed the performance of models developed for specific tasks using smaller datasets [1]. Recent advancements like transformer architectures and robotic foundation models have led to models like Google's RT-1 [2], which demonstrate the potential for robots to generalize across various tasks and environments. However, one critical area remains underexplored: the ability of these models to generalize across entirely new robotic embodiments.

Google's RT-1 model is an impressive work, tested on a collection of real-world robotic experiences, where in different institutes a fleet of robots were performing 700 tasks [2]. The robots in the training set, such as the Franka, Kuka iiwa, UR5 and the EveryDay robot, can move their end-effector in a spherical working-space. None of the robots in the dataset is of the SCARA type. As example of such SCARA robot is the classic UMI RTX robot, a robot which is still functional at the University of Amsterdam's Intelligent Robotics Lab, despite its age of nearly 40 years (see Fig. 1).

More details on this particular robot and the experiment are given in [3] and [4]. The focus of this study is to see if generalization capabilities of the Google's RT-1 model can be extended to an unseen robot, of a complete different SCARA type.

1.1 RT-1

The RT-1 model [2] was presented as a joint effort between Robotics at Google, Everyday Robots, and Google Research, at the end of 2022. The motivation is the following: end-to-end robotic learning generally relies on task-specific datasets that are narrowly tailored towards the robots intended tasks. In recent years however, there has been a shift in those areas, moving away from isolated, small-scale models and datasets towards large, general foundation models that are pre-trained on extensive, diverse datasets.

These models are able to absorb experience from large datasets to learn general patterns across tasks, allowing highly improved generalisation to unseen tasks compared to the traditional approach. While removing the need for task specific datasets is generally appealing in many domains, it is of very high significance in robotics, where the collection of datasets is typically very costly due to the need of either engineering-heavy autonomous operation or expensive human demonstrations.

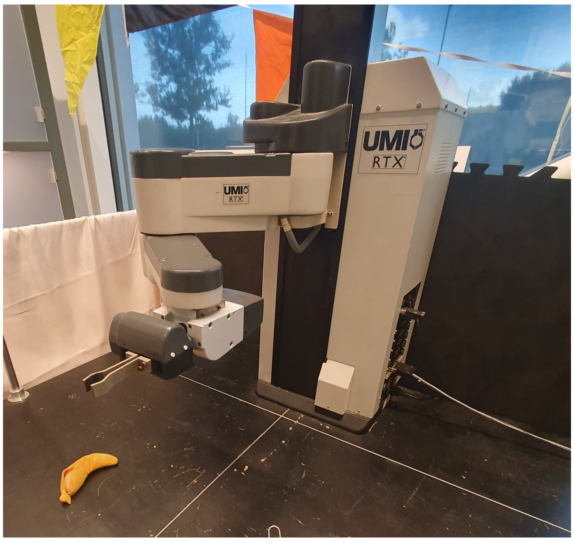

Fig. 1: The UMI-RTX robot with an object in its working space.

2 Zero-shot task execution

The Open X-Embodiment dataset [1] is such large dataset, consisting of more than 1M robot trajectories, collected from 22 different robots from 34 robotic research institutes around the world. All this experience was used to train the the RT-1 model.

The RT-1 model takes two kinds of input: images from the environment and a natural language instruction (e.g. "Pick up the banana"). This is enough to start moving the UMI-RTX robot around [4]. During the zero-shot experiment 48 different environments are tested (e.g. other objects, other camera positions). In some environments the robot was nearly stationary, but in some environments action can be seen on all robot axes, although not in a sequence needed to complete the instruction.

3 Fine-Tuning

When zero-shot execution of the task was not successful [4], we investigate if the performance could be improved by fine-tuning the model on demonstrations from the UMI robot. The RT-1-X model was fine-tuned with a dataset of 100 manual demonstrations of the UMI executing a simple task (picking up a banana from the workspace).

This experiment evaluated the performance of the fine-tuned model on the demonstrated task. In this experiment, a success rate of 23% could be achieved. In most of the non-successful evaluation runs, the robot executed the correct movements, and was off from the target object by less than five centimeters. Experiment runs where the robot executes the correct sequence of actions and is off by less than five centimeters are classified as "near miss". All evaluation runs that resulted either in success or a near miss combined make up 80% of all evaluation runs.

4 Conclusion

This study explored the generalization capabilities of the RT-1-X model, particularly its ability to adapt to a robot type not seen before. A dataset of demonstrations was collected using the UMI robot, and a fine-tuning pipeline for RT-1-X was developed. The fine-tuning process effectively enhanced the model's performance on the new embodiment and the learned task, although the performance did not reach the levels achieved on the embodiments seen during pre-training. Additionally (see [3]), it was found that no concrete knowledge about objects to be manipulated was transferred from the pre-training dataset.

References

[1] A. O'Neill *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6892–6903. DOI: 10.1109/ICRA57147.2024.10611477.

[2] A. Brohan *et al.*, *RT-1: Robotics Transformer for Real-World Control at Scale*, preprint arXiv 2212.06817, Aug. 2022. DOI: 10.48550/ARXIV.2212.06817.

[3] J. Salzer and A. Visser, *Bringing the rt-1-x foundation model to a scara robot*, 2024. arXiv: 2409.03299 [cs.RO].

[4] J. Salzer, "Bringing the rt-1-x foundation model for robotic control to new embodiments," M.S. thesis, Universiteit van Amsterdam, Aug. 2024.