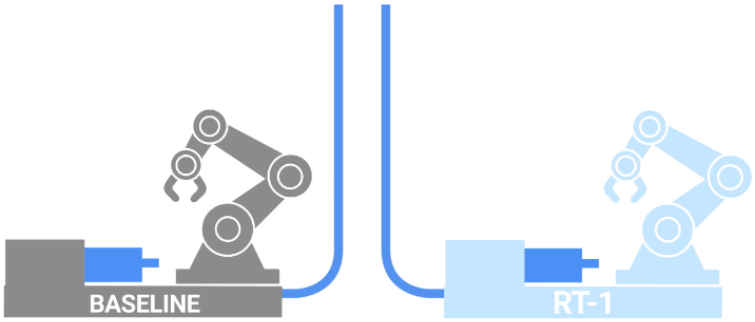
RT-1: Robotics Transformer for real-world control at scale

December 13, 2022 · Posted Keerthana Gopalakrishnan and Kanishka Rao, Google Research, Robotics at Google

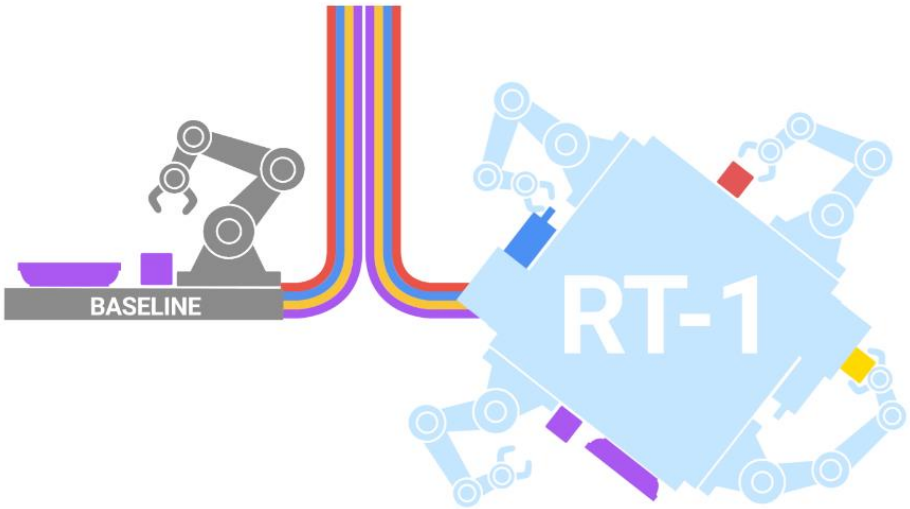
Major recent advances in multiple subfields of machine learning (ML) research, such as computer vision and natural language processing, have been enabled by a shared common approach that leverages large, diverse datasets and expressive models that can absorb all of the data effectively. Although there have been various attempts to apply this approach to robotics, robots have not yet leveraged highly-capable models as well as other subfields.

Several factors contribute to this challenge. First, there's the lack of large-scale and diverse robotic data, which limits a model's ability to absorb a broad set of robotic experiences. Data collection is particularly expensive and challenging for robotics because dataset curation requires engineering-heavy autonomous operation, or demonstrations collected using human teleoperations. A second factor is the lack of expressive, scalable, and fast-enough-for-real-time-inference models that can learn from such datasets and generalize effectively.

To address these challenges, we propose the Robotics Transformer 1 (RT-1), a multi-task model that tokenizes robot inputs and outputs actions (e.g., camera images, task instructions, and motor commands) to enable efficient inference at runtime, which makes real-time control feasible. This model is trained on a large-scale, real-world robotics dataset of 130k episodes that cover 700+ tasks, collected using a fleet of 13 robots from Everyday Robots (EDR) over 17 months. We demonstrate that RT-1 can exhibit significantly improved zero-shot generalization to new tasks, environments and objects compared to prior techniques. Moreover, we carefully evaluate and ablate many of the design choices in the model and training set, analyzing the effects of tokenization, action representation, and dataset composition. Finally, we're open-sourcing the RT-1 code, and hope it will provide a valuable resource for future research on scaling up robot learning.



To address these challenges, we propose the Robotics Transformer 1 (RT-1), a multi-task model that tokenizes robot inputs and outputs actions (e.g., camera images, task instructions, and motor commands) to enable efficient inference at runtime, which makes real-time control feasible. This model is trained on a large-scale, real-world robotics dataset of 130k episodes that cover 700+ tasks, collected using a fleet of 13 robots from Everyday Robots (EDR) over 17 months. We demonstrate that RT-1 can exhibit significantly improved zero-shot generalization to new tasks, environments and objects compared to prior techniques. Moreover, we carefully evaluate and ablate many of the design choices in the model and training set, analyzing the effects of tokenization, action representation, and dataset composition. Finally, we're open-sourcing the RT-1 code, and hope it will provide a valuable resource for future research on scaling up robot learning.



https://research.google/blog/rt-1-robotics-transformer-for-real-world-control-at-scale/

RT-1 absorbs large amounts of data, including robot trajectories with multiple tasks, objects and environments, resulting in better performance and generalization.

Robotics Transformer (RT-1)

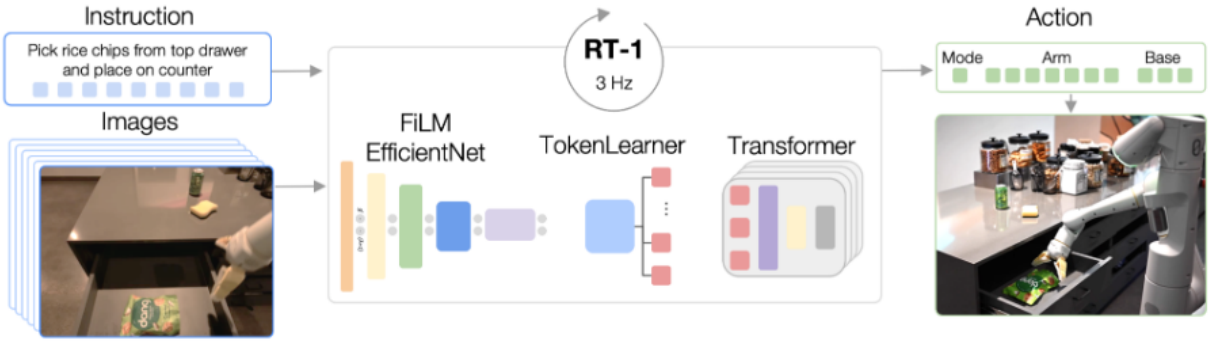
RT-1 is built on a transformer architecture that takes a short history of images from a robot's camera along with task descriptions expressed in natural language as inputs and directly outputs tokenized actions.

RT-1's architecture is similar to that of a contemporary decoder-only sequence model trained against a standard categorical cross-entropy objective with causal masking. Its key features include: image tokenization, action tokenization, and token compression, described below.

Image tokenization: We pass images through an EfficientNet-B3 model that is pre-trained on ImageNet, and then flatten the resulting 9×9×512 spatial feature map to 81 tokens. The image tokenizer is conditioned on natural language task instructions, and uses FiLM layers initialized to identity to extract task-relevant image features early on.

Action tokenization: The robot's action dimensions are 7 variables for arm movement (x, y, z, roll, pitch, yaw, gripper opening), 3 variables for base movement (x, y, yaw), and an extra discrete variable to switch between three modes: controlling arm, controlling base, or terminating the episode. Each action dimension is discretized into 256 bins.

Token Compression: The model adaptively selects soft combinations of image tokens that can be compressed based on their impact towards learning with the element-wise attention module TokenLearner, resulting in over 2.4x inference speed-up.



RT-1's architecture: The model takes a text instruction and set of images as inputs, encodes them as tokens via a pre-trained FiLM EfficientNet model and compresses them via TokenLearner. These are then fed into the Transformer, which outputs action tokens.

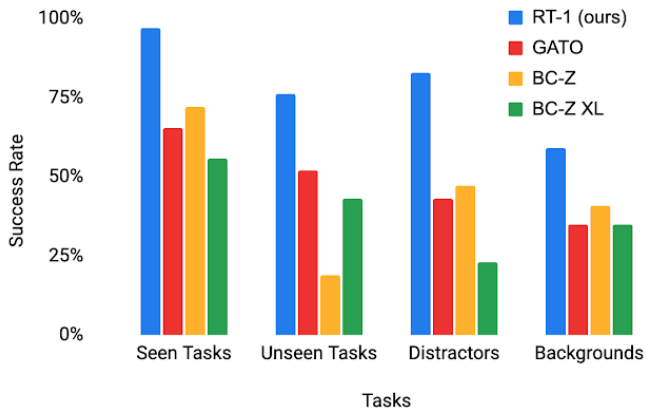
To build a system that could generalize to new tasks and show robustness to different distractors and backgrounds, we collected a large, diverse dataset of robot trajectories. We used 13 EDR robot manipulators, each with a 7-degree-of-freedom arm, a 2-fingered gripper, and a mobile base, to collect 130k episodes over 17 months. We used demonstrations provided by humans through remote teleoperation, and annotated each episode with a textual description of the instruction that the robot just performed. The set of high-level skills represented in the dataset includes picking and placing items, opening and closing drawers, getting items in and out drawers, placing elongated items up-right, knocking objects over, pulling napkins and opening jars. The resulting dataset includes 130k+ episodes that cover 700+ tasks using many different objects.

Experiments and Results

To better understand RT-1's generalization abilities, we study its performance against three baselines: Gato, BC-Z and BC-Z XL (i.e., BC-Z with same number of parameters as RT-1), across four categories:

1. **Seen tasks performance**: performance on tasks seen during training
2. **Unseen tasks performance**: performance on unseen tasks where the skill and object(s) were seen separately in the training set, but combined in novel ways
3. **Robustness (distractors and backgrounds)**: performance with distractors (up to 9 distractors and occlusion) and performance with background changes (new kitchen, lighting, background scenes)
4. **Long-horizon scenarios**: execution of SayCan-type natural language instructions in a real kitchen

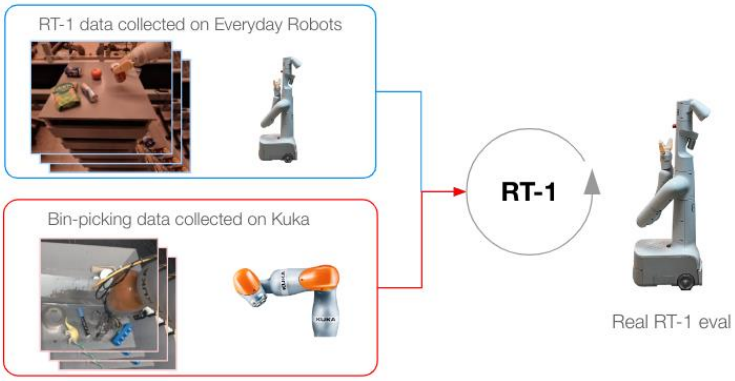
RT-1 outperforms baselines by large margins in all four categories, exhibiting impressive degrees of generalization and robustness.



Performance of RT-1 vs. baselines on evaluation scenarios.

Incorporating Heterogeneous Data Sources

To push RT-1 further, we train it on data gathered from another robot to test if (1) the model retains its performance on the original tasks when a new data source is presented and (2) if the model sees a boost in generalization with new and different data, both of which are desirable for a general robot learning model. Specifically, we use 209k episodes of indiscriminate grasping that were autonomously collected on a fixed-base Kuka arm for the QT-Opt project. We transform the data collected to match the action specs and bounds of our original dataset collected with EDR, and label every episode with the task instruction "pick anything" (the Kuka dataset doesn't have object labels). Kuka data is then mixed with EDR data in a 1:2 ratio in every training batch to control for regression in original EDR skills.



Training methodology when data has been collected from multiple robots.

Our results indicate that RT-1 is able to acquire new skills by observing other robots' experiences. In particular, the 22% accuracy seen when training with EDR data alone jumps by almost 2x to 39% when RT-1 is trained on both bin-picking data from Kuka and existing EDR data from robot classrooms, where we collected most of RT-1 data. When training RT-1 on bin-picking data from Kuka alone, and then evaluating it on bin-picking from the EDR robot, we see 0% accuracy. Mixing data from both robots, on the other hand, allows RT-1 to infer the actions of the EDR robot when faced with the states observed by Kuka, without explicit demonstrations of bin-picking on the EDR robot, and by taking advantage of experiences collected by Kuka. This presents an opportunity for future work to combine more multi-robot datasets to enhance robot capabilities.

Training Data	Classroom Eval	Bin-picking Eval
Kuka bin-picking data + EDR data	90%	**39%**
EDR only data	92%	22%
Kuka bin-picking only data	0	0

RT-1 accuracy evaluation using various training data.

Long-Horizon SayCan Tasks

RT-1's high performance and generalization abilities can enable long-horizon, mobile manipulation tasks through SayCan. SayCan works by grounding language models in robotic affordances, and leveraging few-shot prompting to break down a long-horizon task expressed in natural language into a sequence of low-level skills.

SayCan tasks present an ideal evaluation setting to test various features:

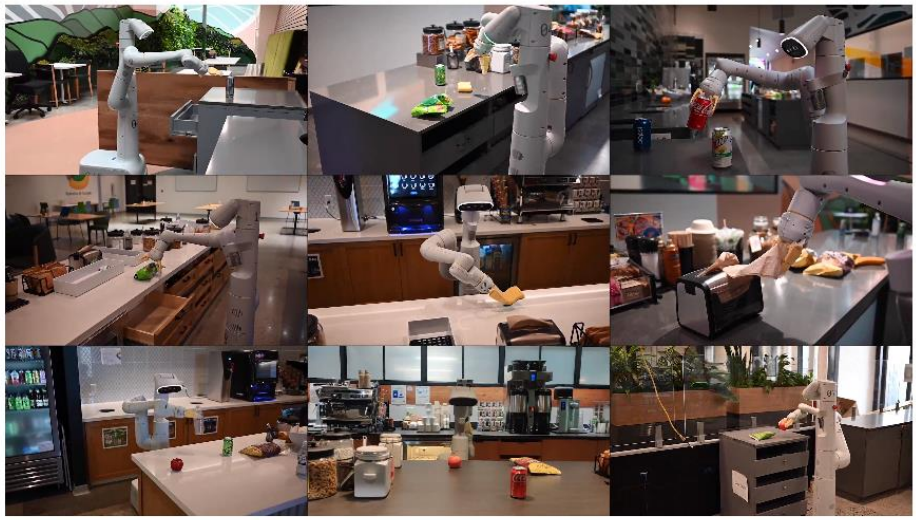
1. Long-horizon task success falls exponentially with task length, so high manipulation success is important.
2. Mobile manipulation tasks require multiple handoffs between navigation and manipulation, so the robustness to variations in initial policy conditions (e.g., base position) is essential.
3. The number of possible high-level instructions increases combinatorially with skill-breadth of the manipulation primitive.

We evaluate SayCan with RT-1 and two other baselines (SayCan with Gato and SayCan with BC-Z) in two real kitchens. Below, "Kitchen2" constitutes a much more challenging generalization scene than "Kitchen1". The mock kitchen used to gather most of the training data was modeled after Kitchen1.

SayCan with RT-1 achieves a 67% execution success rate in Kitchen1, outperforming other baselines. Due to the generalization difficulty presented by the new unseen kitchen, the performance of SayCan with Gato and SayCan with BCZ shapely falls, while RT-1 does not show a visible drop.

	SayCan tasks in Kitchen1		SayCan tasks in Kitchen2	
	Planning	Execution	Planning	Execution
Original Saycan	73	47	-	-
SayCan w/ Gato	87	33	87	0
SayCan w/ BC-Z	87	53	87	13
SayCan w/ RT-1	87	**67**	87	**67**

The following video shows a few example PaLM-SayCan-RT1 executions of long-horizon tasks in multiple real kitchens.



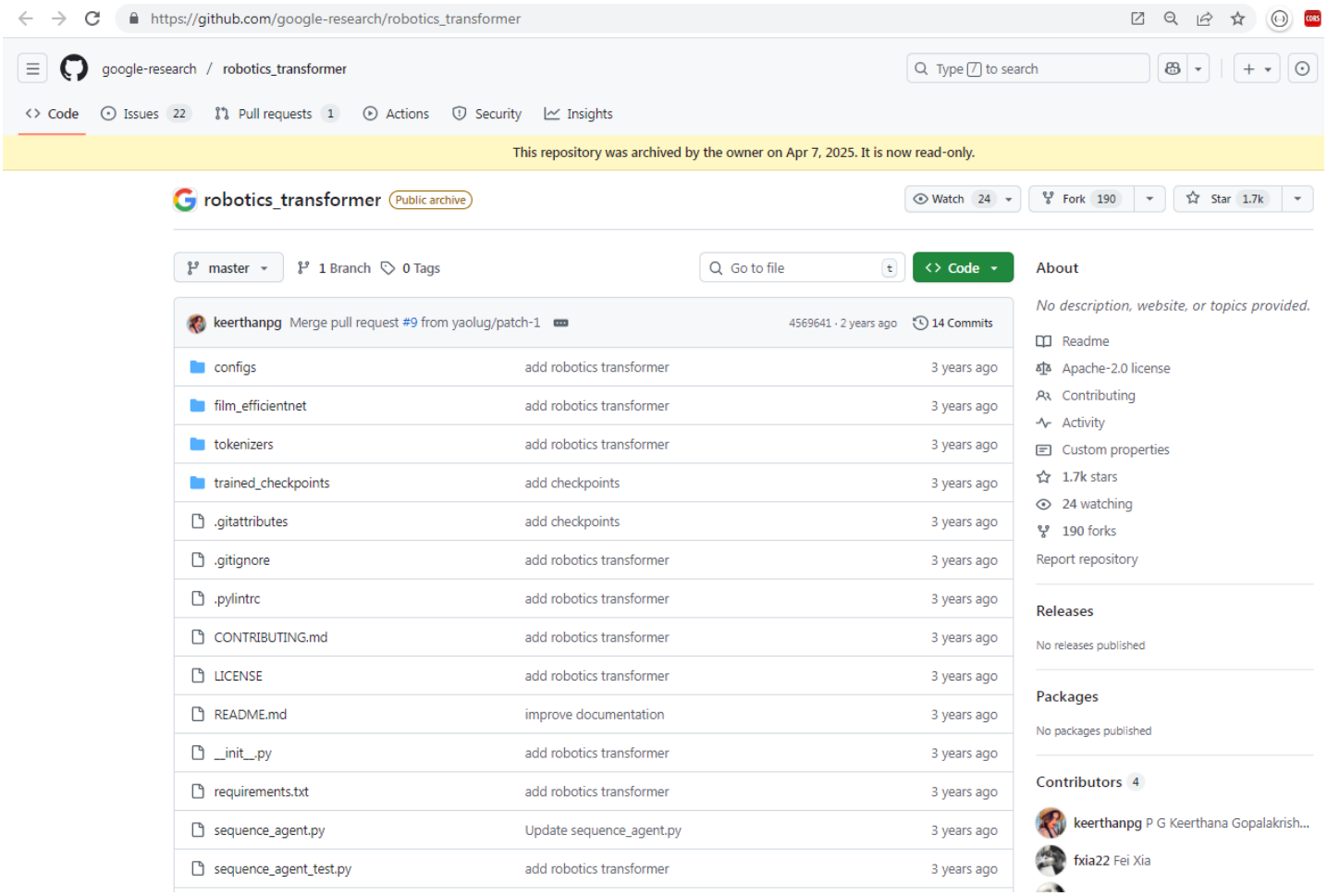
Conclusion

The RT-1 Robotics Transformer is a simple and scalable action-generation model for real-world robotics tasks. It tokenizes all inputs and outputs, and uses a pre-trained EfficientNet model with early language fusion, and a token learner for compression. RT-1 shows strong performance across hundreds of tasks, and extensive generalization abilities and robustness in real-world settings.

As we explore future directions for this work, we hope to scale the number of robot skills faster by developing methods that allow non-experts to train the robot with directed data collection and model prompting. We also look forward to improving robotics transformers' reaction speeds and context retention with scalable attention and memory. To learn more, check out the paper, open-sourced RT-1 code, and the project website.

Acknowledgements

This work was done in collaboration with Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich.


https://github.com/google-research/robotics_transformer
google-research / robotics_transformer
Type / to search
Code
Issues 22
Pull requests 1
Actions
Security
Insights
This repository was archived by the owner on Apr 7, 2025. It is now read-only.
robotics_transformer
Public archive
Watch 24
Fork 190
Star 1.7k
master
1 Branch
0 Tags
Go to file
t
Code
keerthanpg Merge pull request #9 from yaolug/patch-1
4569641 · 2 years ago
14 Commits
configs add robotics transformer 3 years ago
film_efficientnet add robotics transformer 3 years ago
tokenizers add robotics transformer 3 years ago
trained_checkpoints add checkpoints 3 years ago
.gitattributes add checkpoints 3 years ago
.gitignore add robotics transformer 3 years ago
.pylintrc add robotics transformer 3 years ago
CONTRIBUTING.md add robotics transformer 3 years ago
LICENSE add robotics transformer 3 years ago
README.md improve documentation 3 years ago
__init__.py add robotics transformer 3 years ago
requirements.txt add robotics transformer 3 years ago
sequence_agent.py Update sequence_agent.py 3 years ago
sequence_agent_test.py add robotics transformer 3 years ago
About
No description, website, or topics provided.
Readme
Apache-2.0 license
Contributing
Activity
Custom properties
1.7k stars
24 watching
190 forks
Report repository
Releases
No releases published
Packages
No packages published
Contributors 4
keerthanpg P G Keerthana Gopalakrish...
fxia22 Fei Xia