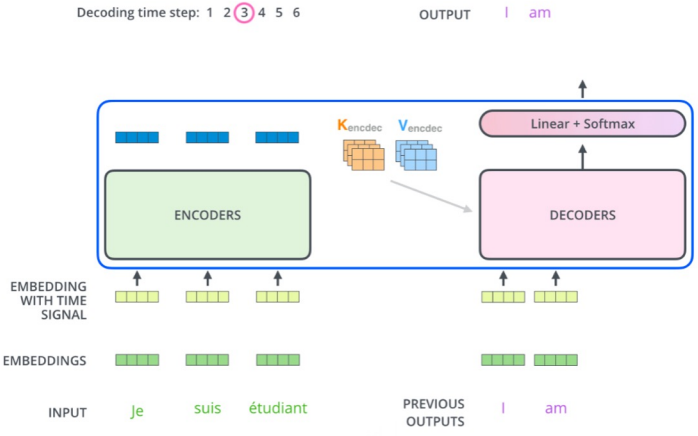
PERCEIVER-ACTOR:A Multi-Task Transformer for Robotic Manipulation

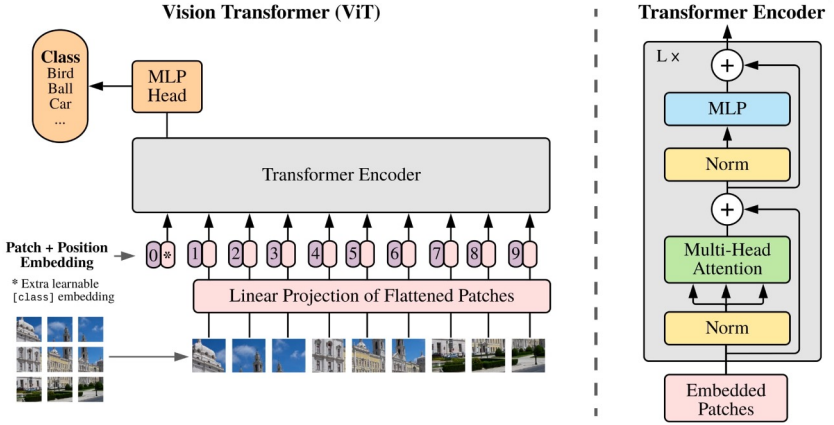
Presenter: Gengcong (Dimi) Yang

10/19/2023

Main Problem

Transformers have revolutionized natural language processing and computer vision with their ability to scale with large datasets. Even in domains that do not conventionally involve sequence modeling, Transformers have been adopted as a general architecture.

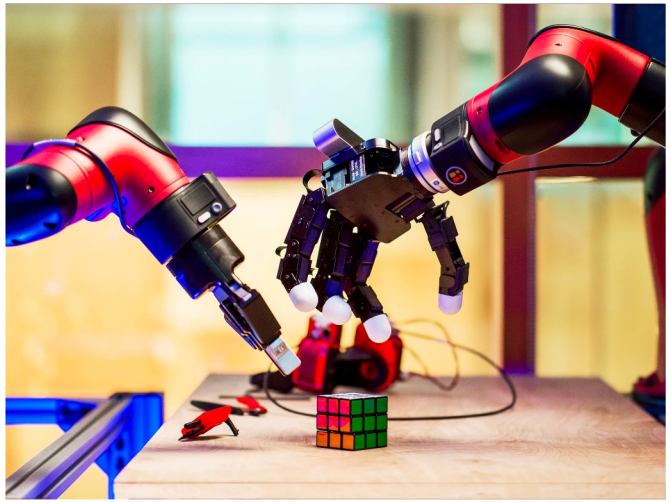


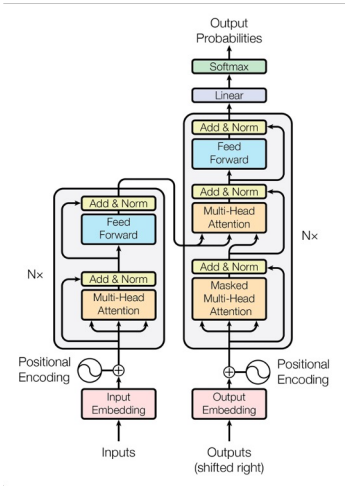


Main Problem

But in robotic manipulation, data is both *limited* and *expensive*.

Can robotic manipulation still benefit from Transformers with the right problem formulation?





Motivation

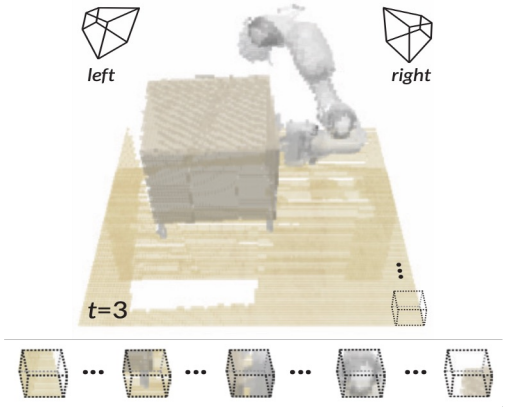
Transformers require the right problem formulation to be *data efficient*.

- Prior works that directly **map 2D images to 6-DoF actions** have shown impressive multi-task capabilities, but they require several weeks or even months of data collection.

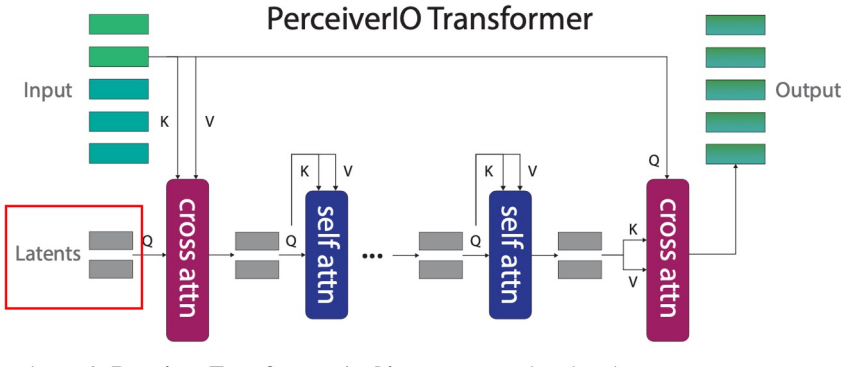
- Recent works in reinforcement-learning like C2FARM construct a **voxelized observation and action space** to efficiently learn visual representations of 3D actions with 3D ConvNets.

Key Insights

- Exploit the 3D structure of **voxel patches**



- Encode high-dimensional input with **small latent vectors**



Problem Setting

Aim to train a **behavior-cloning** agent **grounded on language** with **supervised learning** to **detect actions**.

- Observation voxels **v**
- Language goal **l**

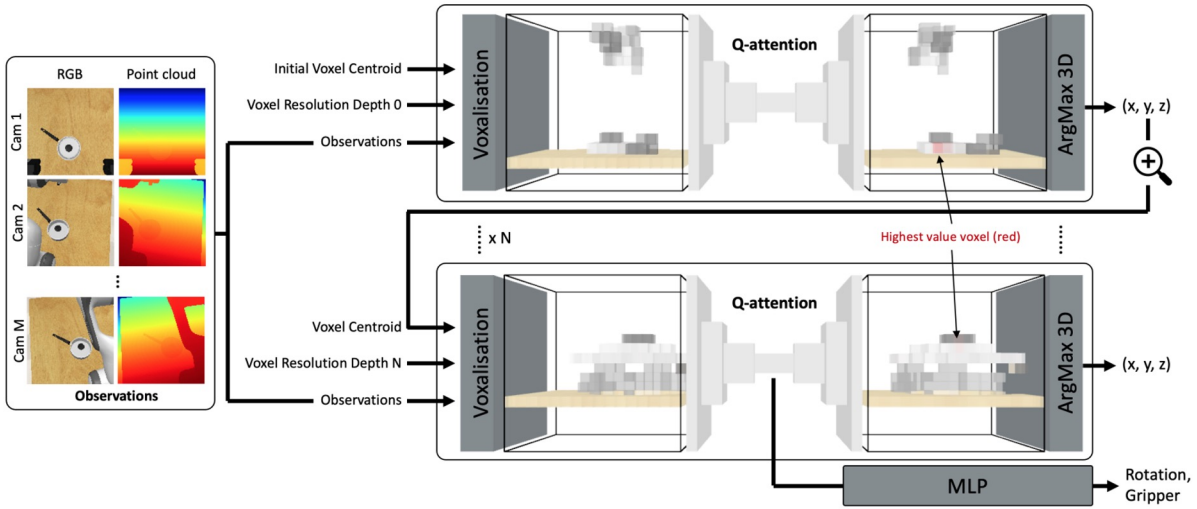
input ⟹ Transformer-based agent ⟹ output

- Voxel location in the grid **(x, y, z)**
- Discrete rotations in Euler angles **(ψ, θ, φ)**
- Gripper open state **ω**
- Collide variable **κ**

The agent is trained with **cross-entropy loss** like a classifier.

Prior Works

- 3D ConvNets

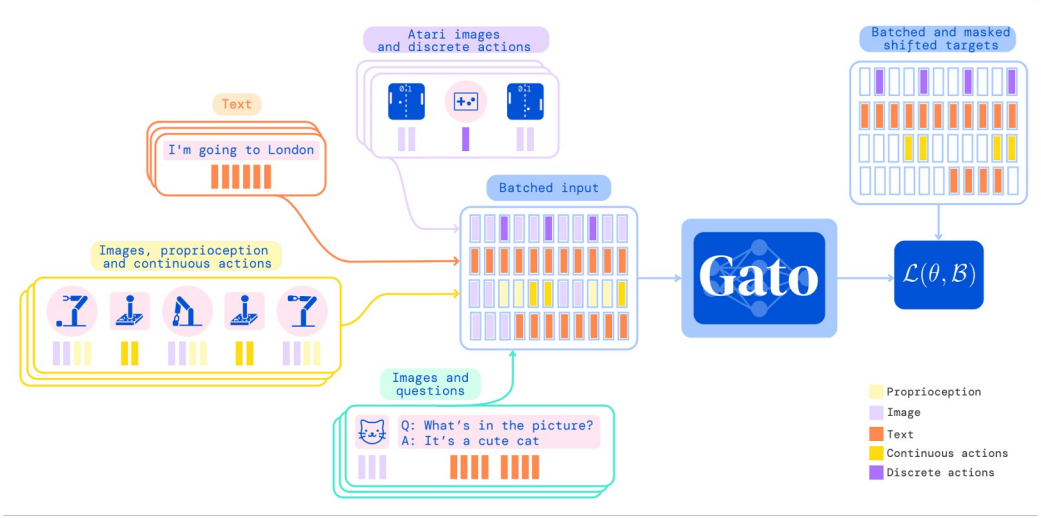


C2FARM, an action-centric reinforcement learning (RL) agent with a coarse-to-fine-grain 3D-UNet backbone, which has a limited receptive field that cannot look at the entire scene at the finest level.

S. James, K. Wada, T. Laidlow, and A. J. Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In Computer Vision and Pattern Recognition (CVPR), 2022.

Prior Works

- Transformer



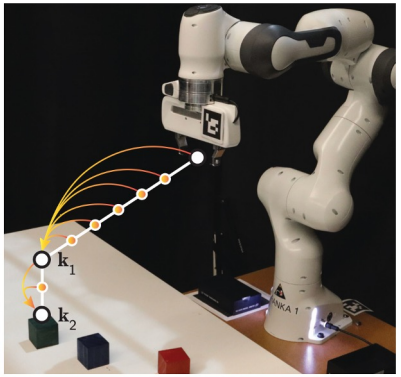
Gato, a Transformer trained in a multi-domain setting. For the robotic control task, it uses 2D formulation and relies on extremely large datasets.

S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, et al. A generalist agent. arXiv preprint arXiv:2205.06175, 2022.

Proposed Approach

Data Preprocessing

- Keyframe Extraction



Each expert demonstration ζ is a sequence of continuous actions A paired with observations O.

For each expert demonstration, we extract a set of keyframe actions $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_m\} \subset \mathcal{A}$.
An action is a keyframe if

1. the joint-velocities are near zero
2. the gripper open state has not changed

Each datapoint in the demonstration ζ can then be cast as a “predict the next keyframe action” task.

Proposed Approach

Data Preprocessing

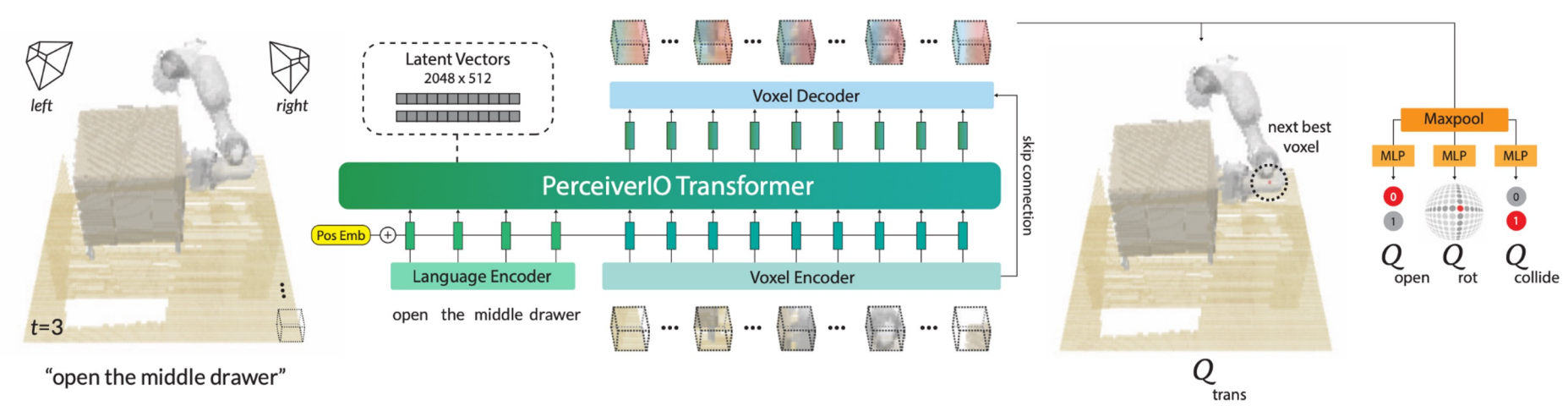
- Voxelization

  Use a voxel grid to represent both the observation and action space.

  1. The observation voxels are reconstructed from RGB-D observations.
  2. The keyframe actions are discretized in the voxel grid.

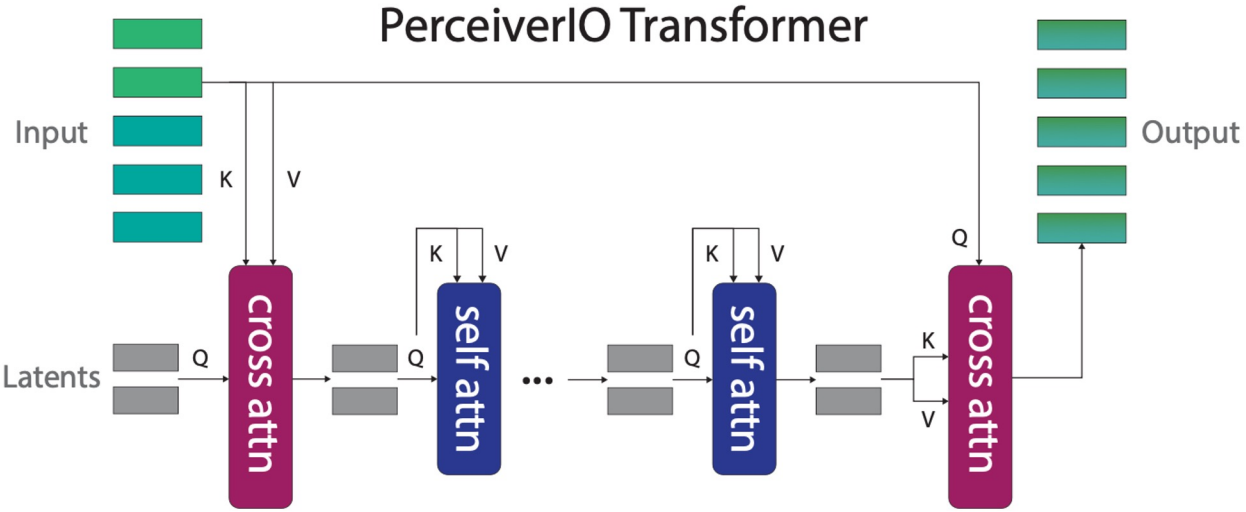
Proposed Approach

Perceiver-Actor (PerAct)



Proposed Approach

PerceiverIO Transformer (Perceiver)

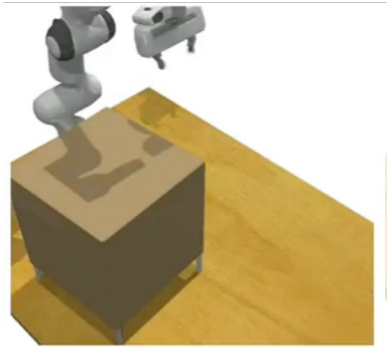


Instead of attending to the entire input, it first computes cross-attention between the input and a much smaller set of latent vectors (which are randomly initialized and trained).

A.Jaegle,S.Borgeaud,J.-B.Alayrac,C.Doersch,C.Ionescu,D.Ding,S.Koppula,D.Zoran, A. Brock, E. Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795, 2021.

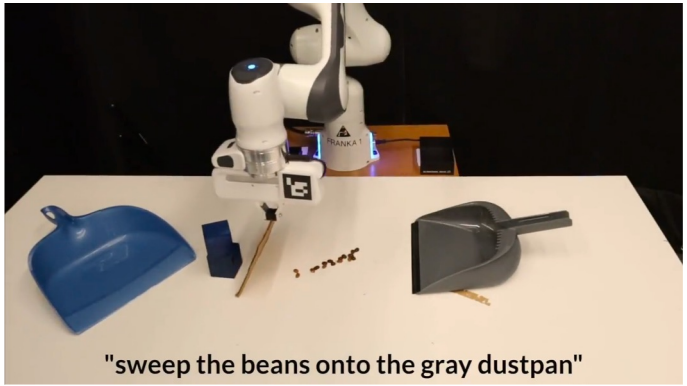
Experimental Setup

- Tasks
  - ➔ 18 RLBench tasks in simulation setup.



"open the middle drawer"

  - ➔ 7 tasks in real-robot setup.



"sweep the beans onto the gray dustpan"

Experimental Setup

- Evaluation Metric
  - Each multi-task agent is evaluated independently on all tasks.
  - Evaluations are scored either 0 for failures or 100 for complete successes.
  - Average success rates on 25 evaluation episodes per task are reported.

Experimental Setup

- Baselines
  - ❖ Image-BC (CNN): an image-to-action agent using a CNN encoder.
  - ❖ Image-BC (ViT): an image-to-action agent using a ViT encoder.
  - ❖ C2FARM-BC: prior state-of-the-art 3D ConvNet, but trained with cross-entropy loss instead of RL and conditioned with language features.

Experimental Results

- Multi-Task Performance

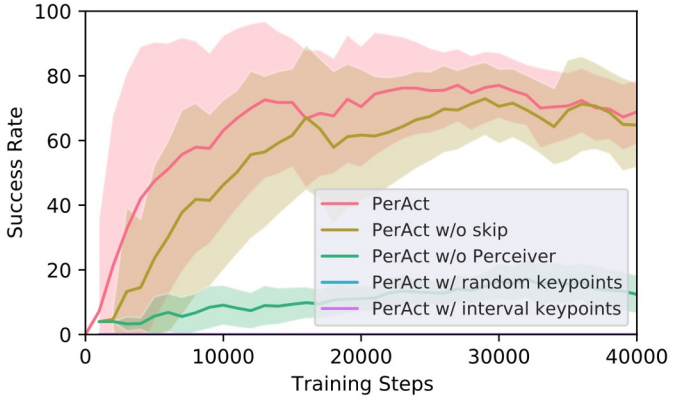
	open drawer		slide block		sweep to dustpan		meat off grill		turn tap		put in drawer		close jar		drag stick		stack blocks	
Method	10	100	10	100	10	100	10	100	10	100	10	100	10	100	10	100	10	100
Image-BC (CNN)	4	4	4	0	0	0	0	0	20	8	0	8	0	0	0	0	0	0
Image-BC (ViT)	16	0	8	0	8	0	0	0	24	16	0	0	0	0	0	0	0	0
C2FARM-BC	28	20	12	16	4	0	40	20	60	68	12	4	28	24	**72**	24	4	0
PERACT (w/o Lang)	20	28	8	12	20	16	40	48	36	60	16	16	16	12	48	60	0	0
PERACT	**68**	**80**	**32**	**72**	**72**	**56**	**68**	**84**	**72**	**80**	**16**	**68**	**32**	**60**	36	**68**	**12**	**36**

	screw bulb		put in safe		place wine		put in cupboard		sort shape		push buttons		insert peg		stack cups		place cups	
	10	100	10	100	10	100	10	100	10	100	10	100	10	100	10	100	10	100
Image-BC (CNN)	0	0	0	4	0	0	0	0	0	0	4	0	0	0	0	0	0	0
Image-BC (ViT)	0	0	0	0	4	0	4	0	0	0	16	0	0	0	0	0	0	0
C2FARM-BC	12	8	0	12	**36**	8	**4**	0	8	8	**88**	**72**	0	**4**	0	0	0	0
PERACT (w/o Lang)	0	**24**	8	20	8	**20**	0	0	0	0	60	68	4	0	0	0	0	0
PERACT	**28**	**24**	**16**	**44**	20	12	0	**16**	**16**	**20**	56	48	**4**	0	0	0	0	0

- PerAct significantly outperforms the other baselines.
- Tasks with more variations like *stack blocks* need substantially more data.
- All agents achieve near zero success for the high-precision tasks.
- Without a language goal, the agent performs poorly.

Experimental Results

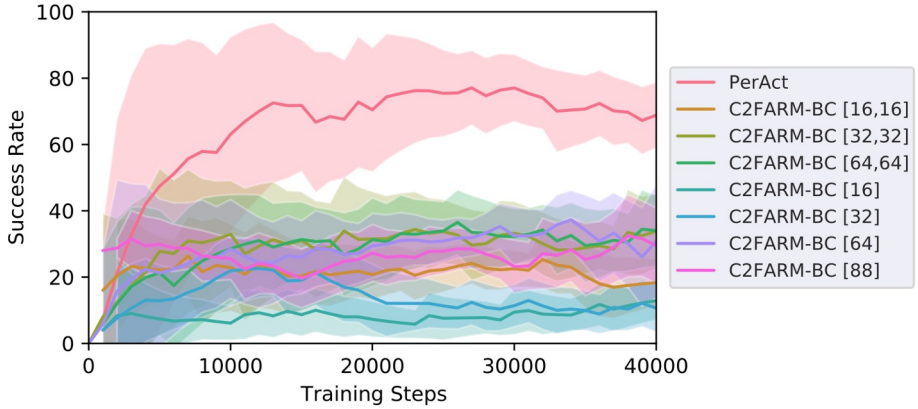
- Ablations



- ❖ The skip connection helps train the agent slightly faster.
- ❖ The Perceiver Transformer is crucial for achieving good performance with the global receptive field.
- ❖ Extracting good keyframes actions is essential for supervised training as randomly chosen or fixed-interval keyframes lead to zero-performance.

Experimental Results

- Global vs. Local Receptive Field Experiments

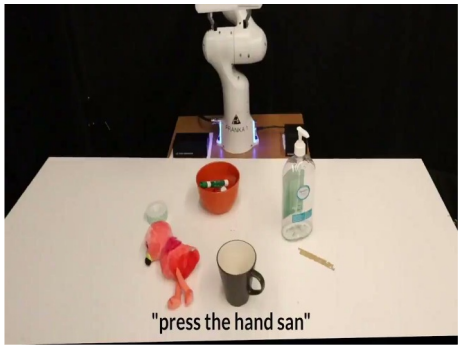


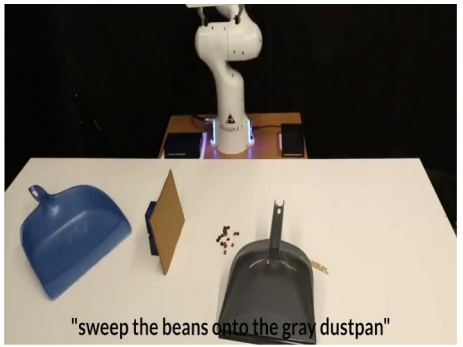
- ❖ PerAct is the only agent that achieves > 70% success.
- ❖ All C2FARM-BC versions perform at chance with ~ 33%.
- ❖ The result indicates that the global receptive field of the Transformer is crucial for solving the task.

Experimental Results

- Real-Robot Results

Task	# Train	# Test	Succ. %
Press Handsan	5	10	90
Put Marker	8	10	70
Place Food	8	10	60
Put in Drawer	8	10	40
Hit Ball	8	10	60
Stack Blocks	10	10	40
Sweep Beans	8	5	20





- ❖ Train one multi-task Transformer from scratch on 7 real-world tasks with just 53 demos in total.
- ❖ PerAct is able to achieve > 65% success on simple short-horizon tasks like pressing hand-sanitizers from just a handful number of demonstrations.
- ❖ But still perform poorly on high-precision tasks like sweeping beans.

Result Insights

- The proposed Transformer model significantly outperforms the existing works in the low-data setting, which demonstrates data efficiency.
- Language grounding is important.
- Global receptive field and latent space encoding play crucial roles.
- Extracting good keyframes actions in the data preprocessing stage is essential for supervised training.

Limitations

- High-precision tasks remain challenging under the multi-task setting.
- The approach relies on a sampling-based motion planner to execute discretized actions.
- The experimental results don’t show generalization to unseen objects.
- The approach relies purely on the current observation to predict the next action.
- The results don’t strongly demonstrate the superiority of voxels over images and Transformer over ConvNets.

Future Work

- Use pre-trained vision features that are aligned with the language for better generalization.
- Use dynamic task-weighting methods for better multi-task optimization.
- Add a memory mechanism to utilize historical information.

Extended Readings

- **Website**: https://peract.github.io/
- **C2FARM**: S. James, et al. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation.
- **Perceiver**: A.Jaegle, et al. Perceiver io: A general architecture for structured inputs & outputs.

Summary

- **Problem**: Design a Transformer-based architecture for 6-DoF manipulation under the requirement of data efficiency
- **Importance**: Enable robotic manipulation learning from limited data
- **Challenge**: Exploit the 3D structure of voxel patches for efficient 6-DoF behavior-cloning with Transformers
- **Limitation of prior work**: Ineffective in low-data setting or lacking global receptive field
- **Key insight**: Achieving data efficiency by processing voxels with Transformer
- **Demonstrated by**: State-of-the-art behavior-cloning success across 18 simulation tasks with very little data