DeepMind RT-2

A Multi-Modal Vision + Language Model for Robotic Control

By Ammar Siddiqui

UW DATA SCIENCE CLUB.

WAT.ai

Agenda

- **Context and Motivation**
  - Why should you care?
- **Related + Foundational Work**
  - Vision-Language Models
  - Robotic Control
- **How does RT-2 work?**
  - Model Architecture
  - Fine-tuning and Datasets
- **Experiments and Results**
  - Tasks and Datasets
- **Limitations**
  - Inference Latency
  - Challenges in Embedded ML
- **References**

UW DSC.

Context and Motivation

- Using text and vision prompts to control a robot
  - Vision-Language-Action
- Massive knowledge base of an LLM helps agents understand real-world context
- VLMs (Vision-Language Models) can be fine-tuned to become general task learners through “action tokens”
- **Embodied AI**

UW DSC.

Foundational Work

- **Transformers and Large Language Models**
  - Attention is all you need, PaLM, GPT, Mistral, etc
- **Representation Learning**
  - Multi-Modal Embeddings -> Images + Text
- **Vision Transformers (ViT)**
  - Training transformers to learn image tasks
  - Large scale image understanding -> Object Detection Action Recognition, Segmentation, etc

UW DSC.

Related Work

- **Vision-Language Models (VLMs)**
  - Visual Question-Answering {text, image} -> {text}
  - Image Captioning {image} -> {text}
  - Image Generation {text} -> {image}
  - Model examples:
    - TimesFormer, LWM, GIT (Generative Image2Text)
- **Robot Learning and Embodied AI**
  - Learning to represent robotic tasks as sequential data that models can understand and generate
  - Simulators -> Habitat, DeepMind Lab, AI2-THOR
  - Model examples:
    - RT-1, Octo, MOO, VC-1, R3M

UW DSC.

How does RT-2 work?

- LLM + ViT -> VLM -> VLA
  - Action Tokens enable the model to output sequences of decisions that the robot's control system can utilize
- Better Generalization of Tasks
  - Understanding tasks beyond the robotic data it was exposed to
- Less training required to learn new tasks
  - Compared to RT-1 + other Embodied AI models

UW DSC.

Training and Fine-tuning

Internet-Scale VQA + Robot Action Data

Q: What is happening in the image?

A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets?

Faire cuire un gâteau.

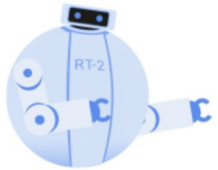
Q: What should the robot do to <task>?

Δ Translation = [0.1, -0.2, 0]
Δ Rotation = [10°, 25°, -7°]

Co-Fine-Tune

Vision-Language-Action Models for Robot Control

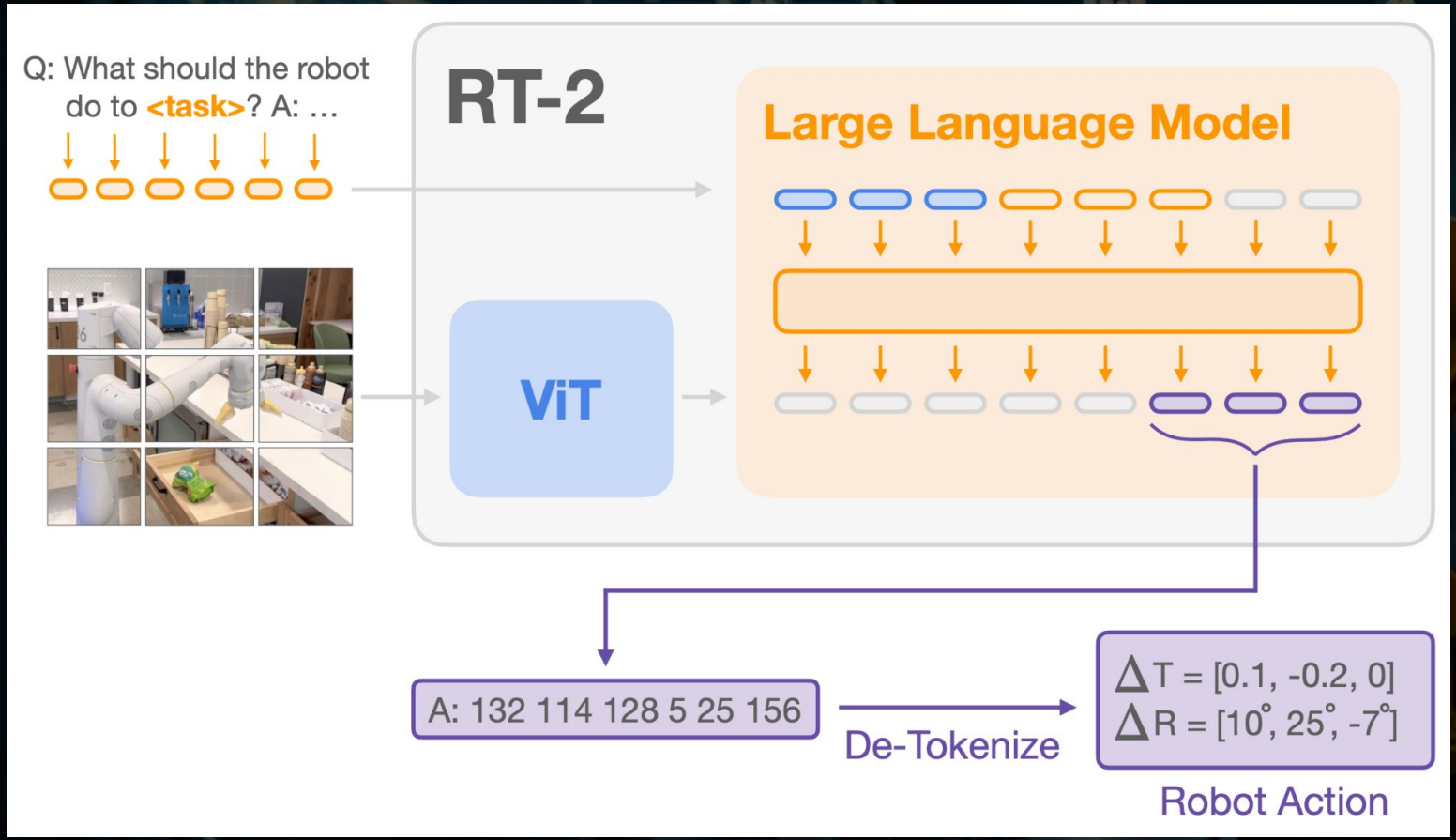
RT-2



Deploy

UW DSC.

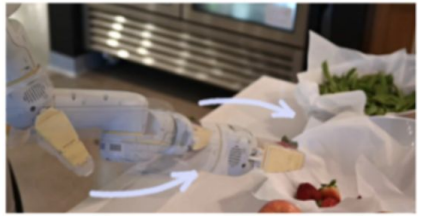
Model Architecture



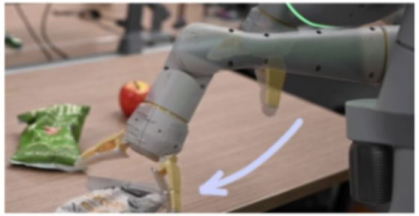
UW DSC.

Inference –> Responding to prompts

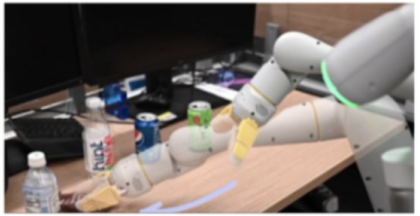
Closed-Loop Robot Control



Put the strawberry into the correct bowl

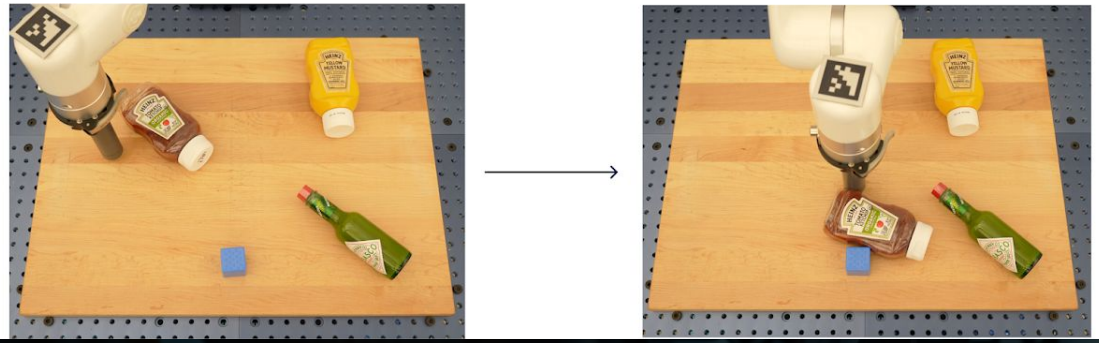


Pick the nearly falling bag



Pick object that is different

Push the *ketchup to the blue cube*



UW DSC.

Datasets and Metrics

- DeepMind Open-X Embodiment Datasets
  - 22 Robot types
  - 311 scenes each with 1M+ episodes of tasks each
  - 527 skills -> pour, grab, stack, connect wires, push object, twist, etc
  - 60 Datasets with ~1800 attributes, 5k+ objects, and 23k+ spatial relations

UW DSC.

Task Generalization Examples

Unseen objects

Unseen backgrounds

Unseen environments

UW DSC.

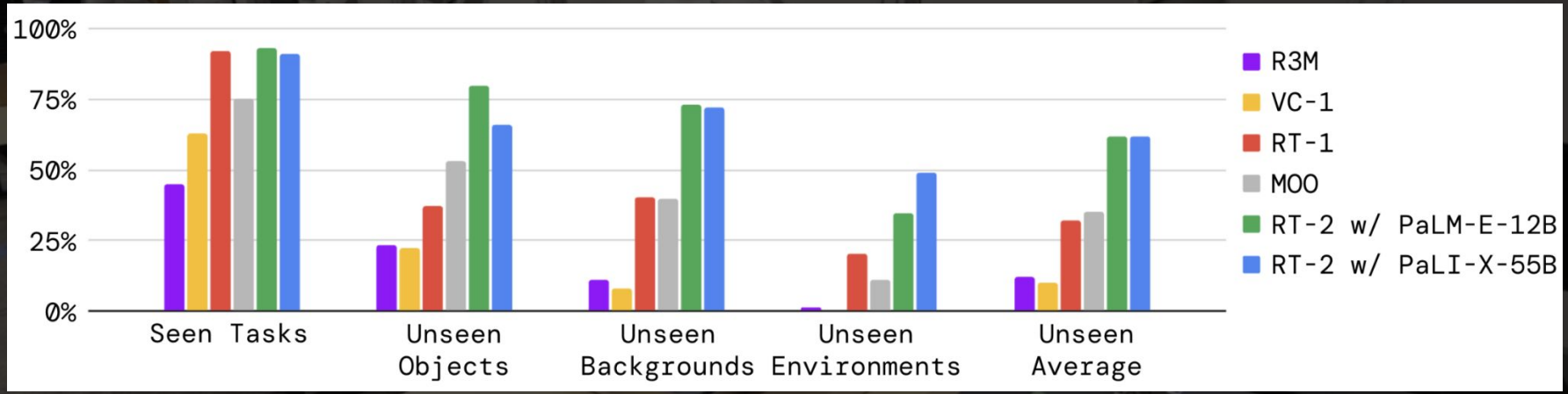
Task Generalization Examples

“Move the apple to Denver Nuggets”

UW DSC.

Experiments

How does RT-2 perform on seen tasks and more importantly, generalize over new objects, backgrounds, and environments?



UW DSC.

Results

put strawberry into the correct bowl

pick up the bag about to fall off the table

move apple to Denver Nuggets

pick robot

place orange in matching bowl

move Red Bull can to H

move soccer ball to basketball

move banana to Germany

move cup to the wine bottle

pick animal with different colour

move coke can to Taylor Swift

move coke can to X

move bag to Google

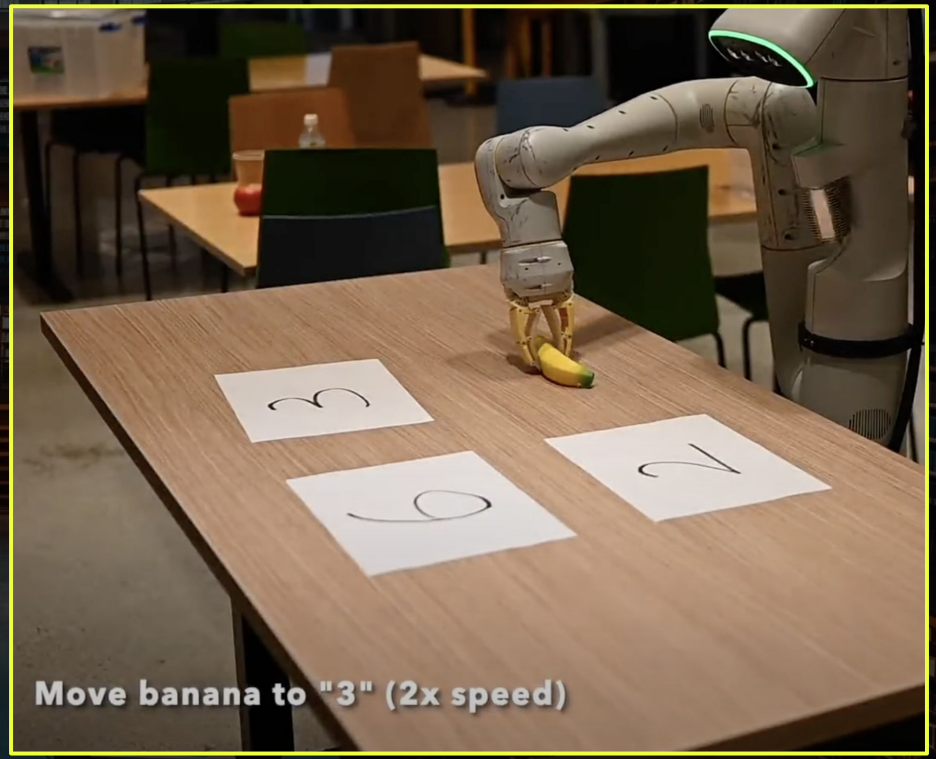
move banana to the sum of two plus one

pick land animal

UW
DSC.

Limitations

- **Inference Cost and Latency**
  - Currently very slow,
  - Video is sped up when showcasing examples
- **Challenges in Embedded Machine Learning**
  - Impossible (for now) to run massive models on the robot's onboard GPU
  - RT-2 ran on a TPU cluster -> 1-3 actions per second, 5 at best



UW DSC.

Thank You!

References

RT-2 Paper
https://arxiv.org/pdf/2307.15818.pdf

Video Demonstration
https://www.youtube.com/watch?v=F3xCTq15mQM

Open X-Embodiment Datasets
https://robotics-transformer-x.github.io/

UW DSC.