Embodied Reasoning Through Planning with Language and Vision Foundation Models

Georgia Tech CS 7643/4644: Deep Learning
Fei Xia, Google DeepMind
11/7/2023
From “Internet AI” to “Embodied AI”

Datasets

Tasks
- Classification
- Segmentation
- Generation
- Captioning

Internet AI

Embodied AI
- Visual Navigation
- Manipulation
- Rearrangement
- Embodied-QA
- Mobile Manipulation
- Instruction Following
Do as I Can, Not as I Say (SayCan): Grounding Language In Robotic Affordances

Say-Can.github.io
I spilled my drink, can you help?

I just worked out, can you bring me a snack and a drink to recover?

I'm feeling tired, can you make me a latte?

How do we make robot learning more useful?
How do we make robot learning more useful?

I spilled my drink, can you help?

I just worked out, can you bring me a snack and a drink to recover?

I'm feeling tired, can you make me a latte?
Language Conditioned Robot Behavior

Naive language conditioned imitation learning works on short horizon tasks but struggles with long-horizon tasks and complex instructions.

“I spilled my drink, can you help with that?”

“I just worked out, can you bring me a snack and a drink to recover?”

[BC-Z, E. Jang et al, 2021]
Mixing language and robotics

Large Language Models (LLMs)

Lots of recent breakthroughs, contain a wealth of knowledge and can handle sequences, memory, and more

PaLM,
Chowdhery et al, 2022
LLMs for robotics

Challenges:
1. Robot Language: Our robots can only do a fixed number of commands and need the problem broken down in actionable steps. This is not what LLMs have seen.

2. Grounding: LLMs have not directly “experienced” the physical world.

3. Safety, alignment, interpretability…
LLMs for robotics

Problem: Our robots can only do a fixed number of commands and need the problem broken down in actionable steps. This is not what LLMs have seen.

We need to get LLMs to speak “robot language”!
Problem: LLMs aren’t grounded in the real-world. They don’t know what’s actually possible from a state with a given embodiment.

We need to ground LLMs in robotic affordances!
Robotic affordances

Reinforcement learning already provides task-based affordances.

They are encoded in the value function!

[Value Function Spaces, Shah, Xu, Lu, Xiao, Toshev, Levine, Ichter, ICLR 2022]

Q-Transformer, 2023.
How would you put an apple on the table?

I would: 1. Find an apple

Language Grounding: Instruction Relevance with LLMs

World Grounding: Task Affordances with Value Functions

I would: 1. Find an apple, 2. ___
Experiment Overview

- **70% planning rate**
- **61% execution rate**
- **101 long-horizon instructions**
- **10+ navigation and manipulation skills in a row**
- **Without grounding nearly halves performance**

<table>
<thead>
<tr>
<th>Instruction Family</th>
<th>Num</th>
<th>Plan</th>
<th>Execution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Language Single Primitive</td>
<td>15</td>
<td>67%</td>
<td>67%</td>
</tr>
<tr>
<td>Natural Language Nouns</td>
<td>15</td>
<td>60%</td>
<td>53%</td>
</tr>
<tr>
<td>Natural Language Verbs</td>
<td>15</td>
<td>80%</td>
<td>67%</td>
</tr>
<tr>
<td>Structured Language</td>
<td>15</td>
<td>100%</td>
<td>87%</td>
</tr>
<tr>
<td>Embodiment</td>
<td>11</td>
<td>64%</td>
<td>55%</td>
</tr>
<tr>
<td>Crowd Sourced</td>
<td>15</td>
<td>73%</td>
<td>67%</td>
</tr>
<tr>
<td>Long-Horizon</td>
<td>15</td>
<td>47%</td>
<td>33%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>101</strong></td>
<td><strong>70%</strong></td>
<td><strong>61%</strong></td>
</tr>
</tbody>
</table>

User: I left out a coke, apple, and water, can you throw them away and then bring me a sponge to wipe the table?
## PaLM-SayCan vs FLAN-SayCan

<table>
<thead>
<tr>
<th></th>
<th>plan</th>
<th>execute</th>
<th>plan</th>
<th>execute</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PaLM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NL Single</td>
<td>15</td>
<td>100%</td>
<td>100%</td>
<td>67%</td>
</tr>
<tr>
<td>NL Nouns</td>
<td>15</td>
<td>67%</td>
<td>47%</td>
<td>60%</td>
</tr>
<tr>
<td>NL Verbs</td>
<td>15</td>
<td>100%</td>
<td>93%</td>
<td>80%</td>
</tr>
<tr>
<td>Structured</td>
<td>15</td>
<td>93%</td>
<td>87%</td>
<td>100%</td>
</tr>
<tr>
<td>Embodiment</td>
<td>11</td>
<td>64%</td>
<td>55%</td>
<td>64%</td>
</tr>
<tr>
<td>Crowd Sourced</td>
<td>15</td>
<td>87%</td>
<td>87%</td>
<td>73%</td>
</tr>
<tr>
<td>Long-Horizon</td>
<td>15</td>
<td>73%</td>
<td>47%</td>
<td>47%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>101</td>
<td>84%</td>
<td>74%</td>
<td>70%</td>
</tr>
</tbody>
</table>

+14% Planning success rate overall
+26% Planning success rate on long-horizon tasks
SayCan: Grounding Language in Robotic Affordances
SayCan: Grounding Language in Robotic Affordances

10x speed
SayCan: Takeaways

- LLMs can provide task grounding
- (Robotic) value functions provide real-world grounding
- This is compatible with any policy as long as there is an affordance

Challenge:

- One bottleneck is still on the skills
- Language-conditioned affordance model
RT-1: Robotics Transformer for Real-World Control at Scale
ROSIE: Scaling Robot Learning with Semantically Imagined Experience
Discussions
PaLM-E: An Embodied Multimodal Language Model

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, Pete Florence

Google Research
Closed-loop end-to-end planning
("Given <img>... Bring me the rice chips from the drawer ")

Long-horizon tasks
("Given <img>.... Sort the blocks by colors into corners")

Vision-language generalist

One model”
- Embodied robotics tasks
- Vision-language
- Language
- … across multiple robot embodiments
- … across multiple modalities (vision, states, neural scenes)

Positive transfer

Emergent visual-language capabilities
Zero-shot multimodal CoT, multi-image reasoning

Zero-shot generalization
(unseen object pairings, or objects)
Multimodal Language Models

- “Frozen”, Tsimpoukelli et al.
- Flamingo, Alayrac et al.
- PaLI, Chen et al.
- BLIP-2, Kosmos-1, GPT-4, …

Language + Robotics

- LLMs + robots for high-level planning
  - SayCan
  - Socratic Models
  - InnerMonologue
  - PaLM-E

Language conditioned policies

- Interactive Language, Lynch et al.
- RT-1, Brohan et al.
PaLI (Google 2022)
PaLI (Google 2022)

<table>
<thead>
<tr>
<th>Method</th>
<th>VQAv2</th>
<th>OKVQA</th>
<th>TextVQA</th>
<th>VizWiz-QA</th>
<th>ST-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test-dev</td>
<td>test-std</td>
<td>val</td>
<td>val</td>
<td>test</td>
</tr>
<tr>
<td>SimVLM</td>
<td>80.03</td>
<td>80.34</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CoCa (2.1B)</td>
<td>82.3</td>
<td>82.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GIT (0.7B)</td>
<td>78.56</td>
<td>78.81</td>
<td>-</td>
<td>59.93</td>
<td>59.75</td>
</tr>
<tr>
<td>GIT2 (5.1B)</td>
<td>81.74</td>
<td>81.92</td>
<td>-</td>
<td>68.38</td>
<td>67.27</td>
</tr>
<tr>
<td>OFA (0.9B)</td>
<td>82.0</td>
<td>82.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flamingo (80B)</td>
<td>82.0</td>
<td>82.1</td>
<td>57.8*</td>
<td>57.1</td>
<td>54.1</td>
</tr>
<tr>
<td>BEiT-3 (1.9B)</td>
<td>84.2</td>
<td>84.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KAT</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mia</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLI-3B</td>
<td>81.4</td>
<td>-</td>
<td>52.4</td>
<td>60.12</td>
<td>-</td>
</tr>
<tr>
<td>PaLI-15B</td>
<td>82.9</td>
<td>-</td>
<td>56.5</td>
<td>65.49</td>
<td>-</td>
</tr>
<tr>
<td>PaLI-17B</td>
<td>84.3</td>
<td>84.3</td>
<td>64.5</td>
<td>71.81</td>
<td>73.06</td>
</tr>
</tbody>
</table>

**COCO**

<table>
<thead>
<tr>
<th>Model</th>
<th>Karpathy-test</th>
<th>val</th>
<th>test</th>
<th>Karpathy-test</th>
<th>val</th>
<th>test</th>
<th>Karpathy-test</th>
<th>val</th>
<th>test</th>
<th>Karpathy-test</th>
<th>val</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEMON (0.7B)</td>
<td>139.1</td>
<td>117.3</td>
<td>114.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SimVLM</td>
<td>143.3</td>
<td>112.2</td>
<td>110.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CoCa (2.1B)</td>
<td>143.6</td>
<td>122.4</td>
<td>120.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GIT (0.7B)</td>
<td>144.8</td>
<td>125.5</td>
<td>123.4</td>
<td>143.7</td>
<td>138.2</td>
<td>113.1</td>
<td>114.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GIT2 (5.1B)</td>
<td>145.0</td>
<td>126.9</td>
<td>124.8</td>
<td>146.6</td>
<td>145.0</td>
<td>119.4</td>
<td>120.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OFA (0.9B)</td>
<td>145.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flamingo (80B)</td>
<td>138.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BEiT-3 (1.9B)</td>
<td>147.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLI-3B</td>
<td>145.4</td>
<td>121.1</td>
<td>-</td>
<td>143.6</td>
<td>-</td>
<td>-</td>
<td>117.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLI-15B</td>
<td>146.2</td>
<td>121.2</td>
<td>-</td>
<td>150.1</td>
<td>-</td>
<td>-</td>
<td>121.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLI-17B</td>
<td>149.1</td>
<td>127.0</td>
<td>124.4</td>
<td>160.0</td>
<td>160.4</td>
<td>123.0</td>
<td>124.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Diagram**

The diagram shows the absolute score difference for various models across different datasets. The bars indicate the performance improvements of PaLI-3B, PaLI-15B, and PaLI-17B compared to baseline models (VGG and ViT). The models are represented by different colors: PaLI-3B (VGG), PaLI-15B (XXL, ViT-G), PaLI-17B (XXL, ViT-G), and PaLI-17B (w/ high-res phase).
PaLI-X (Google 2023)

<table>
<thead>
<tr>
<th>Model</th>
<th>LVIS AP</th>
<th>LVIS AP_Rare</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLD [74] (tuned on non-rare LVIS)</td>
<td>29.3</td>
<td>26.3</td>
</tr>
<tr>
<td>Region-CLIP [75] (tuned on non-rare LVIS)</td>
<td>32.3</td>
<td>22.0</td>
</tr>
<tr>
<td>OwLViT-L/16 [28] (tuned on non-rare LVIS)</td>
<td>34.7</td>
<td>25.6</td>
</tr>
<tr>
<td>OwLViT-L/16 [28] (with Object365 and VG datasets)</td>
<td>34.6</td>
<td>31.2</td>
</tr>
<tr>
<td>PaLI-X (Zeroshot)</td>
<td>12.36</td>
<td>12.16</td>
</tr>
<tr>
<td>PaLI-X (Detection-tuned)</td>
<td>30.64</td>
<td>31.42</td>
</tr>
</tbody>
</table>

Figure 2: Examples demonstrating multilingual, OCR and other capabilities transferred to detection.
# PaLI-3: Smaller, Faster, Stronger

Contrastive or classification pretraining for ViT?

<table>
<thead>
<tr>
<th>Model</th>
<th>COCO Karp-test</th>
<th>VQA v2 test-dev</th>
<th>OKVQA test-std</th>
<th>TallyQA Simple</th>
<th>TallyQA Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimVLM</td>
<td>143.3</td>
<td>80.03</td>
<td>80.34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CoCa (2.1B)</td>
<td>143.6</td>
<td>82.3</td>
<td>82.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GIT (0.7B)</td>
<td>144.8</td>
<td>78.56</td>
<td>78.81</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GIT2 (5.1B)</td>
<td>145.0</td>
<td>81.74</td>
<td>81.92</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OFA (0.9B)</td>
<td>145.3</td>
<td>82.0</td>
<td>82.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flamingo (80B)</td>
<td>138.1</td>
<td>82.0</td>
<td>82.1</td>
<td>57.8*</td>
<td>-</td>
</tr>
<tr>
<td>BEiT-3 (1.9B)</td>
<td>147.6</td>
<td>84.2</td>
<td>84.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLM-E (562B)</td>
<td>138.7</td>
<td>80.0</td>
<td>-</td>
<td><strong>66.1</strong></td>
<td>-</td>
</tr>
<tr>
<td>MoVie</td>
<td>-</td>
<td>69.26</td>
<td>-</td>
<td>-</td>
<td>74.9</td>
</tr>
<tr>
<td>PaLI-17B</td>
<td>149.1</td>
<td>84.3</td>
<td>84.3</td>
<td>64.5</td>
<td>81.7</td>
</tr>
<tr>
<td>PaLI-X (55B)</td>
<td><strong>149.2</strong></td>
<td><strong>86.0</strong></td>
<td><strong>86.1</strong></td>
<td><strong>66.1</strong></td>
<td><strong>86.0</strong></td>
</tr>
<tr>
<td>PaLI-3 (5B)</td>
<td>145.9</td>
<td>85.0</td>
<td>85.2</td>
<td>60.1</td>
<td>83.3</td>
</tr>
</tbody>
</table>

### Contrastive or classification pretraining for ViT?

<table>
<thead>
<tr>
<th>Probe</th>
<th>Captioning</th>
<th>VQA</th>
<th>RefCOCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 tasks</td>
<td>COCO</td>
<td>XM3600</td>
</tr>
<tr>
<td>Classif</td>
<td>88.1</td>
<td>139.9</td>
<td>94.5</td>
</tr>
<tr>
<td>SigLIP</td>
<td><strong>-2.5</strong></td>
<td>+0.4</td>
<td>+1.6</td>
</tr>
<tr>
<td>Classif</td>
<td>86.2</td>
<td>132.6</td>
<td>93.0</td>
</tr>
<tr>
<td>SigLIP</td>
<td><strong>-2.8</strong></td>
<td>+3.2</td>
<td>+1.4</td>
</tr>
<tr>
<td>Classif</td>
<td>83.7</td>
<td>127.7</td>
<td>91.7</td>
</tr>
<tr>
<td>SigLIP</td>
<td><strong>-2.6</strong></td>
<td>+3.6</td>
<td><strong>-2.0</strong></td>
</tr>
</tbody>
</table>
Method & detailed experiments

“Main” model: PaLM-E-562B
- Generalist visual-language model
- PaLM-540B and ViT-22B!
- Trained on: robot data, Internet-scale VQA, captioning

Also explored with PaLM-E:
- Neural 3D scene, and robot state encoders into the LLM
- Object-centric reasoning
- Arbitrary interleaving of text + multimodal modalities

Experimentation
- Several different domains/categories of robot tasks
- Standard vision-language tasks
- Standard language-only tasks
Simple Architecture of PaLM-E

PaLM-E: An Embodied Multimodal Language Model

Given \texttt{<emb> ... <img>}

Q: How to grasp blue block? A: First, grasp yellow block

\text{ViT}

Large Language Model (PaLM)

\text{Control}

A: First, grasp yellow block and...

\text{ViT}

Q: What happened between \texttt{<img> and <img>}

A: The...

Arbitrary interleaving
Comparison to Flamingo
Scene Representation: Object Scene Representation Transformer

![Diagram](image)

- **Novel Scene Input Views** (one or more)
- **Set-Latent Scene Representation**
- **Slot Scene Representation**
- **Slot Mixer**
- **Object-Decomposed Novel Views**

<table>
<thead>
<tr>
<th>Model</th>
<th>Object-centric</th>
<th>LLM pre-train</th>
<th>Embodied VQA</th>
<th>Planning</th>
</tr>
</thead>
<tbody>
<tr>
<td>SayCan (oracle afford.) (Ahn et al., 2022)</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLi (zero-shot) (Chen et al., 2022)</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PaLiC (ours) w/ input enc.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>State</td>
<td>✓ (CT)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>State</td>
<td>✓ (CT)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ViT + TL</td>
<td>✓ (CT)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ViT+IB single robot</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ViT+IB full mixture</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>OBRP (no VQA)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
PaLM-E on TAMP-like tasks

Given image. Q: How to stack the white object on top of the red object?
A: First grasp the red object and place it on the table, then grasp the white object and place it on the red object.
Highlight: Zero-shot Multimodal Chain-of-Thought

Given <img>.
Q: In this photo, are there more cats, or more dogs?
A: Let’s think step by step.
There are 2 dogs and 2 cats.
2 and 2 are equal.
Therefore there are equal number of cats and dogs.

Given <img>. Q: Can I go down this street on a bicycle, yes or no? A: Let’s think step by step.
1. do not enter.
2. except bicycles.
3. do not entry except bicycles.
4. yes.

I see (image omitted). Description: A cow sitting on a rug. Joke: Why is the cow sitting on a rug? Because it wants to be a cow-ch! ... I see <img>. Description: a donkey is carrying a dog, cat, and rooster.
Joke: what do you call a donkey with a rooster on his back. A rooster booster.

Given <img>.
Q: I am getting just two custom pizza for me and my friend. How much should I pay in total? A: Let’s think step by step.
1st step: 2 custom pizzas.
2nd step: 2 x $9.99.
3rd step: $19.98.
Highlight: Complex Multi-image Reasoning

Photo 1: <img1>. Photo 2: <img2>. Q: What is in Photo 1 but NOT in Photo 2? A: Let’s think step by step and explain. Photo 1 has sunglasses on top of folded clothes. Photo 2 does not have sunglasses on top of folded clothes. Therefore, sunglasses are in Photo 1 but not in Photo 2.

Photo 1: <img1>. Photo 2: <img2>. Q: Which t-shirt from Photo 2 matches the sweatshirt from Photo 1? A: Let’s think step by step and explain.
1. photo 1 has sweatshirt.
2. photo 2 has t shirts.
3. photo 1 has blue sweatshirt.
4. photo 2 has blue t shirt.
5. blue t shirt matches blue sweatshirt.

Photo 1, at 10:30 am: <img1>. Photo 2, at 12:45 pm: <img2>. Photo 3, at 3:45 pm: <img3>. Q: I forget, what did I have for lunch, and what time was it? A: Let’s think step by step.
1. you had a sandwich for lunch.
2. it was 12:45 pm.
PaLM-E: Positive Transfer
Real Robot Results

Same exact model checkpoint!
(PaLM-E can be a multi-embodiment robot brain)

Q: How to sort the blocks by colors into the corners? A:
Sample-efficient learning

PaLM-E guiding a real robot through long horizon tasks

50 demonstrations

1-shot learning

zero-shot learning (new object pair)
Sample-efficient learning

zero-shot learning (unseen object)
Language catastrophic forgetting reduced with scale
For more on PaLM-E

- See paper + videos + inference examples at [palm-e.github.io]
Hands-on VLMs

**LLaVA: Large Language and Vision Assistant**

**Visual Instruction Tuning**

*NeurIPS 2023 (Oral)*

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee

▸ University of Wisconsin-Madison ▸ Microsoft Research ▸ Columbia University

*Equal Contribution*
Hands-on VLMs, **Fuyu-8b** and open source PaLM-E

- A good programming exercise:
- Fix the bug in
Discussions
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, Brianna Zitkovich
Let’s dive into RT-2!
Vision-Language Models

- VLMs encompass both **visual** and **semantic** understanding of the world
- In Robotics we have to deal a lot with **both** of these
- How do we leverage all of this knowledge?

VLMs as Robot Policies

- **RT-1**: image + text $\rightarrow$ **discretized actions**
- Similar to a Visual-Language Model (VLM) with different **output tokens**
- Use large pre-trained VLMs directly as the **policy**!
- How do we deal with **actions** when using pre-trained VLMs?

---

Robot actions:
- Moving the robot arm and gripper
- Discretized into 256 bins

Actions in VLMs
- Convert to a string of numbers
- Example: “1 127 115 218 101 56 90 255”
- Alternatives:
  - *Float numbers* - more tokens needed
  - *Human language (left, right etc.*) - can’t be directly executed on a robot

→ Vision-Language-Action (VLA) model!
Training data and underlying models

Models
- PaLI-X (5B, 55B)
- PaLM-E (12B)

Data
- Pretraining: Web-data
- Robot data
  - RT-1 data
  - Human demos
  - 13 robots
  - 17 months
Inference

Q: What should the robot do to \texttt{task}\? A: ...

\[ A: = 132 \ 114 \ 128 \ 5 \ 25 \ 156 \]

\[ \Delta T = [0.1, -0.2, 0] \]
\[ \Delta R = [10^\circ, 25^\circ, -7^\circ] \]

De-tokenize

Robot action

Closed-loop robot control

Pick the nearly falling bag

Pick object that is different
Results: Emergent skills

- 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 the matching bowl
- move redbull can to H
- move soccer ball to basketball
- move banana to Germany
- move cup to the wine bottle
- pick animal with different color
- 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
Results: Emergent skills
Results: Emergent skills

RT-2 generalization and emergent semantic reasoning
Results: Quantitative evals
**Results: Quantitative evals**

RT2 w/ PaLI-X-55B ablations

- Co-Fine-Tuning with VQA data
- Fine-Tuning on robot data only
- Training on robot data from scratch
Results: Language Table

Language Table Benchmark

- Trained on pushing cubes only
- Generalizing to new objects

<table>
<thead>
<tr>
<th>Model</th>
<th>Language-Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC-Zero (Jang et al., 2021)</td>
<td>72 ± 3</td>
</tr>
<tr>
<td>RT-1 (Brohan et al., 2022)</td>
<td>74 ± 13</td>
</tr>
<tr>
<td>LAVA (Lynch et al., 2022)</td>
<td>77 ± 4</td>
</tr>
<tr>
<td>RT-2-PaLI-3B (ours)</td>
<td>90 ± 10</td>
</tr>
</tbody>
</table>
Results: Chain-of-Thought with RT-2-PaLM-E

Prompt: Given <img> Instruction: Bring me a drink.
Prediction:
Plan: pick 7up can.
Action: 1 143 129 123 145 114 115 127

Prompt: Given <img> Instruction: Move all the objects together.
Prediction:
Plan: move green can near green rice chip bag.
Action: 1 128 126 127 135 123 119 127

Prompt: Given <img> Instruction: Pick the object that is different from all other objects
Prediction:
Plan: pick rxbar chocolate.
Action: 1 128 129 125 131 125 128 127

Prompt: Given <img> Instruction: Move the green objects together.
Prediction:
Plan: move green can near green rice chip bag.
Action: 1 130 129 121 131 127 128 127

Prompt: Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127
Results: Chain-of-Thought with RT-2-PaLM-E
Project idea: Open-source VLMs on RT-X data
Embodied Reasoning Through Planning with Language and Vision Foundation Models

Georgia Tech CS 7643/4644: Deep Learning
Fei Xia, Google DeepMind
11/7/2023