Google Research

# Spotlight: Mobile UI Understanding Using Vision-language Models with a Focus

Gang Li, Yang Li {leebird, liyang}@google.com

## Introduction

**Mobile UI understanding is an important task**

- Enable UI automation
- Support accessibility use cases
- Facilitate UI design

**Various UI tasks have been proposed**

- Widget captioning, screen summary, command grounding, tappability prediction, etc.

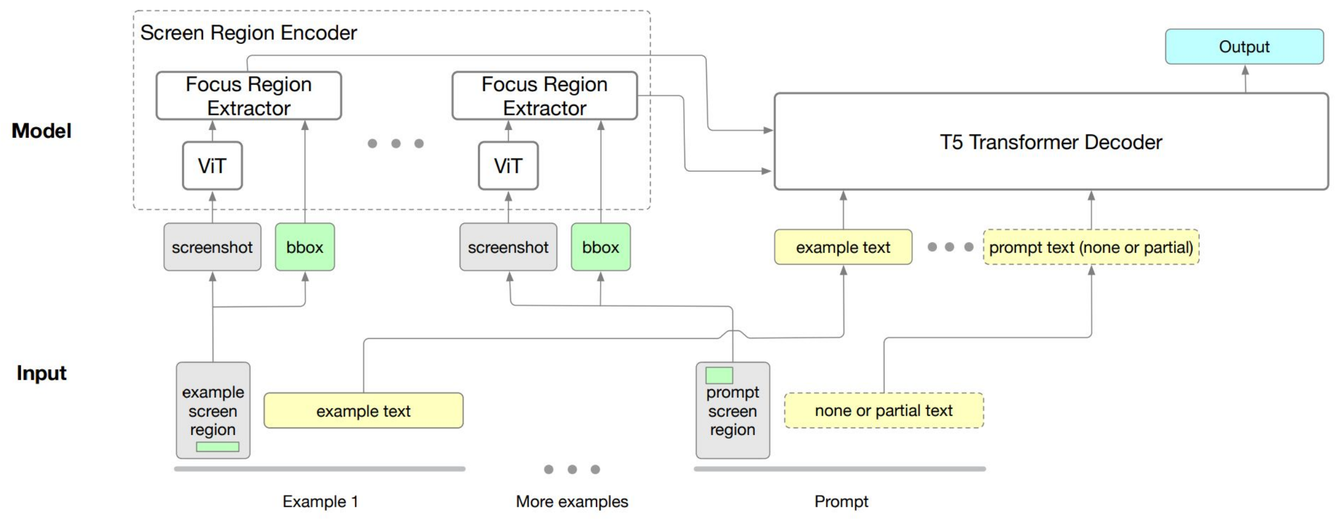
**Problems**

- Most models use screenshot and view hierarchy
- View hierarchy is noisy
  - Li et al. (2022): 37.4% of view hierarchies contain objects with invalid bounding boxes
  - Ross et al. (2018): 92.0% of Floating Action Buttons had missing text labels

## Methods

**A vision-only model with a focus**

- Based on pretrained ViT and T5
- Pretrained on C4 and mobile screenshots
- Pretrained on text decoding task for a screen region




## Results

| | Model | Captioning | Summarization | Grounding | Tappability |
|---|---|---|---|---|---|
| Baselines | Widget Caption | 97.0 | - | - | - |
| | Screen2Words | - | 61.3 | - | - |
| | VUT | 99.3 | 65.6 | 82.1 | - |
| | Taperception | - | - | - | 85.5 |
| | Swearngin & Li (2019) | - | - | - | 87.9* |
| Spotlight | B/16 | 136.6 | 103.5 | 95.7 | 86.9 |
| | L/16 | **141.8** | **106.7** | **95.8** | **88.4** |

### Single Task Fintuning

| T5 | ViT | ViT Enc Layers | Dec Layers | # Parameters |
|---|---|---|---|---|
| base | B/16 | 12 | 12 | 619M |
| base | L/16 | 24 | 12 | 843M |

| Model | Captioning | Summarization | Grounding | Tappability |
|---|---|---|---|---|
| VUT multi-task | 99.3 | 65.1 | 80.8 | - |
| Spotlight B/16 | 140.0 | **102.7** | 90.8 | 89.4 |
| Spotlight L/16 | **141.3** | 99.2 | **94.2** | **89.5** |

### Multitask Fintuning

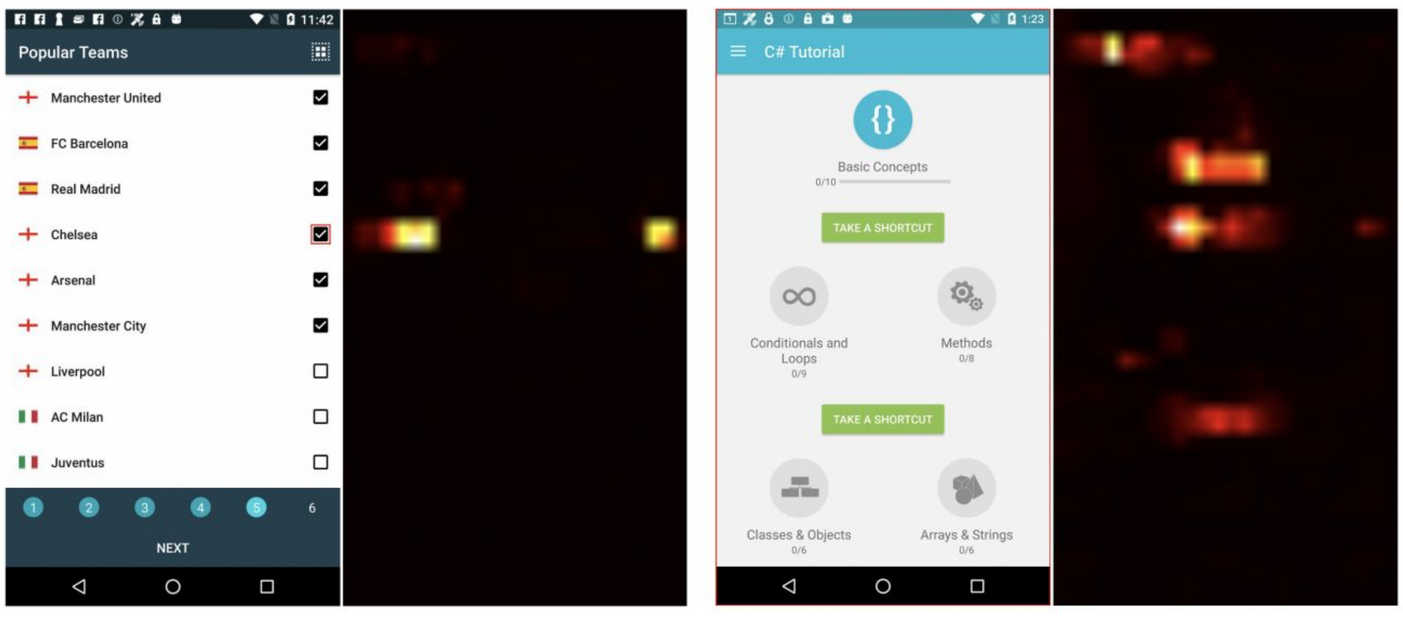
| ViT | 0 | 4 | 8 | 16 | 32 |
|---|---|---|---|---|---|
| B/16 | 57.1 | 56.7 | 55.6 | 55.5 | 54.9 |
| L/16 | 61.6 | 61.9 | 62.0 | 61.9 | 62.1 |

### Few-shot for Widget Captioning

## Analysis & Conclusion

**Analysis**

- The region summarizer focuses on relevant parts even out of the given region




**Conclusion**

- Vision-only models obtain SToA for various UI tasks
- Scaling up model and data sizes is promising direction