When Parameter-efficient Tuning Meets General-purpose Vision-language Models

By: Yihang Zhai et al.

Presenter:
Hari Umesh
Example of Mutimodal Q/A

**Instruction:**
- What objects are in the picture?
- What is the relationship between the objects in the picture?
- What is the relationship between the objects in the picture?

**Ground Truth**
A man wearing a red t-shirt sweeps the sidewalk in front of a brick building

**Full finetune:**
A man in a red shirt is sweeping the sidewalk

**LoRA:**
A man in a red shirt is sweeping the sidewalk

**PETAL:**
A man in a red shirt is sweeping the sidewalk in front of a brick building
Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary
Outline

● Introduction
● PETAL Architecture
● Evaluation
● Summary
Larger Models require larger compute

- As model size and accuracy increases, the demand for the amount of compute also increases
- Training Large models from scratch are expensive

Figure 3. Compute required for training transformer models.
Instruction Tuning

- Training LLMs based on instructions
- Allows models adaptable to a wide-range of tasks without task-specific training

Instruction:
What fruit is typically added to the top of cereal?

Answer:
Overview of Existing PET Multimodal Tasks

Figure 1. Overview of existing parameter-efficient tuning methods applied in multimodal tasks.
InstructBLIP

Choose the correct option to the following question: which picture shows the pizza inside the oven? Options: (a) left one (b) right one. Answer: left one
Multimodal Instruction Challenges

1. Finetuning full models is expensive

2. Lack in semantic information in instructions, which hinders multimodal alignment
Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary
PETAL Main Contributions

1. Novel **Dynamic Mode Approximation** for efficient tuning

2. Enhanced Instruction Semantics
   - Adaptive Instruction MOEs module
   - Score-based information bottleneck
Model Overview Diagram

### Instruction:
What are the characteristics of the objects in the picture?

Adaptive Instruction MOEs

Enhanced Instruction

Learnable Query

Large-scale Language Model

Response:
A cute dog is running
Dynamic Mode Approximation for Efficient Tuning

- Approximates the attention weights in Transformer architecture based on CP decomposition with a dynamic weighting scheme

\[ H^m = \Gamma W_0 X^m + \left( \sum_{r=1}^{R} \lambda_r^m (u_r \circ v_r \circ p_r) \right) X^m \]
Adaptive Instruction MOEs Module

- Extracts information from multiple perspectives by setting multiple instructions for each image with a different focus then stack them in a text paragraph
- Then features are extracted and merged then
Score-based Information bottleneck loss

- Mutual information (MI) loss that enhances semantics of instructions
- Calculate loss based on the product of normalized features and instructions
- Maximizes the mutual information between the representation and the target and minimizes between the representation and the input

\[
\max \text{MI} = \max I(Z; Y) - \eta I(Z; X)
\]
\[
\mathcal{L}_{IB} = \text{MI}(\hat{z}; y)
\]
Outline

● Introduction
● PETAL Architecture
● Evaluation
● Summary
Tasks, Datasets, and Baselines

**Tasks:** Image Captioning and Question Answering

**Datasets**
- Captioning: Flickr30K, TextCaps
- Q/A: OKVQA, A-OKVQA, TextVQA

**Baselines:** InstructBLIP and LLaVA on 4 PEFT METHODS
- HEAD TUNING
- MAPLE
- LLAMA-ADAPTER
- LORA
Implementation?

- Apply PET exclusively to the Q-Former enhanced by approximation techniques
- LLMs Used: FlanT5 and Vicuna-7B
- GPU: 5 epochs 8x A100 (80GB) GPU
### Results: Image Captioning

Table 1. We select two captioning datasets, Flickr30K and TextCaps, for performance comparison on the famous image captioning benchmark. We report CIDEr score, ROUGE-1 F1 score, ROUGE-1 recall, ROUGE-L F1 score, and ROUGE-1 recall for both of them, where ↑ and ↓ respectively indicates how much our method has improved or declined compared to the best parameter-efficient baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tunable</th>
<th>Flickr30K</th>
<th>TextCaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CIDEr</td>
<td>ROUGE-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F1</td>
<td>Recall</td>
</tr>
<tr>
<td>Fine-tune InstructBLIP (FlanT5_64x)</td>
<td>188 M</td>
<td>63.5</td>
<td>34.9</td>
</tr>
<tr>
<td>Head (FlanT5_64x)</td>
<td>11.8 M</td>
<td>60.8</td>
<td>34.5</td>
</tr>
<tr>
<td>MAPLE (FlanT5_64x)</td>
<td>2.9 M</td>
<td>59.4</td>
<td>34.2</td>
</tr>
<tr>
<td>LLaMA-Adapter (FlanT5_64x,R=128)</td>
<td>4.8 M</td>
<td>60.5</td>
<td>33.7</td>
</tr>
<tr>
<td>LoRA (FlanT5_64x,R=64)</td>
<td>5.0 M</td>
<td>59.8</td>
<td>34.0</td>
</tr>
<tr>
<td>PETAL (FlanT5_64x,R=64)</td>
<td>1 M</td>
<td>63.3</td>
<td>35.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.5↑</td>
<td>0.7↑</td>
</tr>
<tr>
<td>Fine-tune InstructBLIP (Vicuna-7B)</td>
<td>188 M</td>
<td>65.8</td>
<td>35.5</td>
</tr>
<tr>
<td>LLaVA (Vicuna-7B)</td>
<td>7B</td>
<td>64.6</td>
<td>35.9</td>
</tr>
<tr>
<td>Head (Vicuna-7B)</td>
<td>11.8 M</td>
<td>61.9</td>
<td>35.1</td>
</tr>
<tr>
<td>MAPLE (Vicuna-7B)</td>
<td>2.9 M</td>
<td>61.3</td>
<td>35.6</td>
</tr>
<tr>
<td>LLaMA-Adapter (Vicuna-7,B,R=128)</td>
<td>4.8 M</td>
<td>62.4</td>
<td>36.6</td>
</tr>
<tr>
<td>LoRA (Vicuna-7B,R=64)</td>
<td>5.0 M</td>
<td>62.5</td>
<td>35.0</td>
</tr>
<tr>
<td>PETAL (Vicuna-7B,R=64)</td>
<td>1 M</td>
<td>63.6</td>
<td>35.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.1↑</td>
<td>0.4↓</td>
</tr>
</tbody>
</table>
Results: Question Answering

Table 2. Performance comparison on VQA benchmarks. We conduct experiments on three datasets: A-OKVQA, OKVQA, and TextVQA, using accuracy as the evaluation metric. ↑ and ↓ respectively indicates our improvement compared to the best baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tunable</th>
<th>A-OKVQA</th>
<th>TextVQA</th>
<th>OKVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Based on FlanT5-xl</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine-tune</td>
<td>188 M</td>
<td>56.7</td>
<td>24.1</td>
<td>55.2</td>
</tr>
<tr>
<td>Head</td>
<td>11.8 M</td>
<td>54.3</td>
<td>21.0</td>
<td>52.1</td>
</tr>
<tr>
<td>MAPLE</td>
<td>2.9 M</td>
<td>54.1</td>
<td>21.2</td>
<td>52.4</td>
</tr>
<tr>
<td>LLaMA-Adapter</td>
<td>4.8 M</td>
<td>53.2</td>
<td>20.9</td>
<td>52.8</td>
</tr>
<tr>
<td>LoRA</td>
<td>5.0 M</td>
<td>54.5</td>
<td>21.4</td>
<td>53.4</td>
</tr>
<tr>
<td>PETAL</td>
<td>1.0 M</td>
<td>55.8</td>
<td>21.5</td>
<td>53.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>↑1.3</td>
<td>↑0.1</td>
<td>↑0.2</td>
</tr>
<tr>
<td><strong>Based on Vicuna-7B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine-tune</td>
<td>188 M</td>
<td>63.6</td>
<td>62.4</td>
<td>27.7</td>
</tr>
<tr>
<td>LLaVA</td>
<td>7B</td>
<td>52.2</td>
<td>52.7</td>
<td>21.3</td>
</tr>
<tr>
<td>Head</td>
<td>11.8 M</td>
<td>63.2</td>
<td>60.1</td>
<td>25.8</td>
</tr>
<tr>
<td>MAPLE</td>
<td>2.9 M</td>
<td>62.8</td>
<td>60.3</td>
<td>24.5</td>
</tr>
<tr>
<td>LLaMA-Adapter</td>
<td>4.8 M</td>
<td>63.0</td>
<td>60.5</td>
<td>26.4</td>
</tr>
<tr>
<td>LoRA</td>
<td>5.0 M</td>
<td>63.5</td>
<td>60.8</td>
<td>26.3</td>
</tr>
<tr>
<td>PETAL</td>
<td>1.0 M</td>
<td><strong>63.8</strong></td>
<td>61.8</td>
<td><strong>27.7</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>↑0.3</td>
<td>↑1.0</td>
<td>↑1.3</td>
</tr>
</tbody>
</table>
PETAL captures greater object information

**Instruction:**
- What objects are in the picture?
- What is the relationship between the objects in the picture?
- What is the relationship between the objects in the picture?

**Ground Truth**
A man wearing a red t-shirt sweeps the sidewalk in front of a brick building.

**Full finetune:**
A man in a red shirt is sweeping the sidewalk.

**LoRA:**
A man in a red shirt is sweeping the sidewalk.

**PETAL:**
A man in a red shirt is sweeping the sidewalk in front of a brick building.
PETAL captures the relationship between objects

**Instruction:**
- What objects are in the picture?
- What is the relationship between the objects in the picture?
- What is the relationship between the objects in the picture?

**Ground Truth:**
On a sunny, dry day, wearing full football gear, a Texas A&M football player tries to reach an Iowa State football player, for the football during the game.

**Full finetune:**
The football player is wearing a red jersey and a yellow helmet

**LoRA:**
A football player is running with the ball

**PETAL:**
A football player is being tackled by another player
Table 3. Results of few-shot instruction tuning. We have two configurations: we extract 50/150 data items from the training set for training. We conducted tests on AOKVQA, OKVQA, FLickr30K, and TextCaps. For AOKVQA and OKVQA, we calculate the accuracy, while for Flickr30K and TextCaps, we compute the CIDEr Score and ROGUE-1 F1 Score.

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>A-OKVQA</th>
<th>OKVQA</th>
<th>Flickr30K</th>
<th>TextCaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>Accuracy</td>
<td>CIDEr</td>
<td>ROGUE-1 F1</td>
</tr>
<tr>
<td>50-shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>188 M</td>
<td>53.2</td>
<td>52.0</td>
<td>52.6</td>
<td>53.5</td>
</tr>
<tr>
<td>LoRA</td>
<td>5 M</td>
<td>53.1</td>
<td>52.1</td>
<td>52.6</td>
<td>52.7</td>
</tr>
<tr>
<td>Ours</td>
<td>1 M</td>
<td><strong>53.3</strong></td>
<td><strong>52.6</strong></td>
<td><strong>53.0</strong></td>
<td><strong>52.9</strong></td>
</tr>
<tr>
<td>150-shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>188 M</td>
<td>53.0</td>
<td>52.2</td>
<td>52.6</td>
<td>53.5</td>
</tr>
<tr>
<td>LoRA</td>
<td>5 M</td>
<td>53.1</td>
<td>52.1</td>
<td>52.6</td>
<td>52.7</td>
</tr>
<tr>
<td>Ours</td>
<td>1 M</td>
<td><strong>53.1</strong></td>
<td><strong>52.6</strong></td>
<td><strong>52.9</strong></td>
<td><strong>53.1</strong></td>
</tr>
</tbody>
</table>
Visualization Comparison of Instruction Enhancement

Caption: Dog with orange ball at feet, stands on shore shaking off water.

Caption: Two construction workers are sitting up on the side of a building.

Figure 4. Cross-attention visualization results on Flickr30k dataset.
## Table 5. Training time and parameter size comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Tunable</th>
<th>Flickr30K</th>
<th>TextCaps</th>
<th>AOKVQA</th>
<th>OKVQA</th>
<th>TextVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tune</td>
<td>188M</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>MAPLE</td>
<td>4.8M</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
<td>0.90</td>
<td>0.91</td>
</tr>
<tr>
<td>LoRA</td>
<td>5.0M</td>
<td>0.91</td>
<td>0.92</td>
<td>0.86</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td>PETAL</td>
<td>1M</td>
<td>0.85</td>
<td>0.88</td>
<td>0.86</td>
<td>0.85</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- PETAL Architecture
- Evaluation
- Summary
Summary of Key Contributions

- PETAL: novel approach for parameter efficient tuning in vision-language models
- Dynamic mode approximation increases efficiency
- Enhanced Instructions through adaptive instruction MOEs and mutual information loss
Discussion and Future Contributions

**Strengths**
- Multimodal paper with specific instruction tuning optimizations and dynamic mode approximation

**Future Work**
- Quantization
- Inference level optimizations for Multimodal inputs
- Switch Transformer optimization w/ MOEs
Table 4. Results of ablation studies that remove the important components. ↑ and ↓ respectively indicates how much the variant has improved or declined compared to our PETAL. DMA stands for Dynamic Mode Approximation, AIM represents Adaptive Instruction MOEs, and SIB is Score-based Information Bottleneck loss.

<table>
<thead>
<tr>
<th>Method</th>
<th>DMA</th>
<th>AIM</th>
<th>SIB</th>
<th>A-OKVQA</th>
<th>Flickr30K</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Accuracy</td>
<td>CIDEr</td>
<td>ROGUE-L</td>
</tr>
<tr>
<td>PETAL V1</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>55.2</td>
<td>61.1</td>
<td>34.5</td>
</tr>
<tr>
<td>PETAL V2</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>55.6</td>
<td>62.6</td>
<td>34.8</td>
</tr>
<tr>
<td>PETAL V3</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>54.8</td>
<td>60.7</td>
<td>31.0</td>
</tr>
<tr>
<td>PETAL V4</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>55.4</td>
<td>61.2</td>
<td>33.8</td>
</tr>
<tr>
<td>PETAL w. random instruction</td>
<td></td>
<td></td>
<td></td>
<td>55.1</td>
<td>60.8</td>
<td>32.0</td>
</tr>
<tr>
<td>PETAL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>55.8</td>
<td>63.4</td>
<td>35.1</td>
</tr>
</tbody>
</table>
Appendix 1: Ablation Study PETAL Architecture

Figure 3. Results of different number of experts.