CRCV REU Week 3

Misaki Matsuura
Dr. Ser Nam Lim
Overview

Accomplished:

● Read papers on CLIP and LLaVA
● Ran CLIP and LLaVA with CIFAR-100 and scikit-image

Working on:

● Formulating accurate means of comparison between CLIP and LLaVA
● Running full LLaVA model on cluster
Motivation

Problems with standard deep learning approach to computer vision:

● Costly datasets
  ○ Manual labeling
● Narrow
  ○ Struggles with novel classes
● Poor real-world performance
Model 1: CLIP (Contrastive Language-Image Pre-Training)

- **Vision transformer**
  - ~150M parameters (ViT-B/32)
- Trained with 400M (image, text) pairs from internet
- Predicts most relevant text snippet given image
  - Encodes image and text to calculate similarity
- Matches performance of ResNet-50 on ImageNet without using labeled data

CLIP examples

Top predictions:
- snake: 65.31%
- turtle: 12.29%
- sweet_pepper: 3.83%
- lizard: 1.88%
- crocodile: 1.75%

CIFAR-100
Model 2: LLaVA (Large Language-and-Vision Assistant)

- Connects vision encoder and LLM
  - CLIP + LLaMA
- Visual instruction-tuning
  - ~80K images
- Autoregressive
  - Takes in previous word and predicts next word
- Chatbot demo

The image features a large, green, patterned snake sitting on a brown background. The snake appears to be camouflaged, blending in with its surroundings. It is positioned in the center of the scene, covering a significant portion of the image. The close-up view of the snake emphasizes its intricate pattern and texture, making it an interesting and visually striking creature.

The image features a small cup of coffee with a spoon in it, placed on a wooden dining table. The cup is filled with coffee, and the spoon is resting inside the cup, likely used for stirring or sipping the beverage. The table appears to be well-maintained and serves as a suitable surface for enjoying a hot cup of coffee.

The image features a red motorcycle parked inside a garage. The motorcycle is positioned near the center of the scene, occupying a significant portion of the space. The garage also contains various items, such as shelves with multiple bottles placed on them, a suitcase, and a bowl. The shelves are located on the left and right sides of the motorcycle, while the suitcase and bowl can be seen towards the right side of the scene. The overall setting suggests that the motorcycle is being stored or maintained in the garage.
**Change in project approach**

**Original plan:** Build multimodal model using CLIP + LLaMA

- CLIP “converts” image to text
- LLaMA generates textual description

**However...**

- LLaVA already performs well, so it may be redundant to build another model

**New plan:** Evaluate LLaVA against CLIP on zero-shot classification

- Run experiments on LLaVA similar to CLIP
- Compare performance
- Datasets:
  - ImageNet
  - SUN
- Run 1 word label tests on LLaVA → use output to train CLIP
Preliminary experiment results: mini CIFAR-100 dataset

**CLIP**
- **Data:** first 20 images from CIFAR-100 test set
- **Accuracy:** 65%

**LLaVA**
- **Data:** first 20 images from CIFAR-100 test set
- **Accuracy:** 35%

*CLIP was given 100 classes to choose from, while LLaVA was not → working on formulating a fair method of comparison*
Next steps

1. Set up LLaVA 13B
   ○ Ran into many issues due to large size
   ○ Attempted:
     ■ Newton cluster
     ■ Shared group checkpoints
     ■ Google Cloud TPU
   ○ Currently attempting:
     ■ Convert GPU model to TPU
     ■ CRCV cluster

2. Formulate accurate and fair evaluation method

3. Run experiments on LLaVA with ImageNet data

<table>
<thead>
<tr>
<th></th>
<th>ImageNet</th>
<th>SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>76.2</td>
<td>58.5</td>
</tr>
<tr>
<td>LLaVA</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Thank you! Questions?