Report On:

**Fine Tuning LLaVA on Custom Dataset**

horizontal line

# Placeholder image

# Introduction

LLaVA or the Large Language and Visual Assistant is a type of Vision Language Model that combines both the capabilities of vision models and large language models, such models are suited to tasks such as **Image Captioning**, **Chatting with pictures**, **Visual Question-Answering(VQA)** etc.

My task was to fine tune any VLM model to a custom dataset and provide a report to further enhance what was learned during the process.

## Methodology

The model that I selected was *“bczhou/tiny-llava-v1-hf”*(ref. HuggingFace), since this model was a much smaller variant of the original LLaVA model, tiny-llava contains about 3-billion parameters, whereas the original one contains about 7 billion and upwards. Fine Tuning the large LLaVA was an impractical solution which I discovered while experimenting, *llava-v1.5-7b-hf* and *llava-v1.6-mistral-7b* models would approximately take 54-60 hours for fine tuning for 1 epoch on low-end GPUs like T4 and P100(GPUs freely available on Kaggle and Google Colab). *Tiny-llava-v1-hf* takes around 10 hours for full fine tuning for 1 epoch on the same GPUs.

The dataset selected was <https://huggingface.co/datasets/HuggingFaceH4/llava-instruct-mix-vsft>, this data contains texts and images both [Size of dataset: 11.4GB, Training samples: 259,000].

When fine tuning the model the settings that I used were:

1. Model was quantized to load in 4bit, with Low CPU memory Usage.
2. LoRA configuration was implemented, with r=64, alpha value=16.
3. Training Arguments were as follows:
4. Learning Rate: 2e-5 (as specified by the authors)
5. Per Device training batch size: 8,
6. Gradient Accumulation steps: 1,
7. Logging Steps: 5,
8. Epochs: 1,
9. Max Global Steps: 500 (Since GPU resources & time was limited)

After applying all these settings the model was fine tuned up to 500 steps, with the final training loss being 0.6735.

*Absence of Demo:*

Since the model training was a computationally exhaustive task, I ran out of GPU memory allocation a lot of times, after further experimentation, I stopped the training at 500 steps. Thus the model was not fully fine-tuned for an epoch (which required around 32659 steps, for 10 hours which was practically not feasible on free resources over cloud like kaggle and colab).

I hope you will consider my work for the assignment and I look forward to hearing back from you soon. Excited to be a part of AI Planet.

Notebook link: [Fine Tuning LLaVA on Custom Data](https://www.kaggle.com/code/paradoxicalvariable/finetuning-llava-on-custom-data/edit)

Thank You.