Report fine tuning multi modal LLAVA

Sri Harsha P.
AI Grad Student
Northeastern University
patallapalli.s@northeastern.edu,sriharsha.py@gmail.com

Abstract

This study fine-tunes a 7B LLaVA model on a custom dataset integrating images and text. The model excels in multimodal tasks like image captioning, QA based on the given image. Results demonstrate fast inference due to the 4-bit and also underscoring the LLAVA architecture's efficacy in specialized domain adaptation.

1 Methodology and Obervations

1.1 Base Model

The base model used is from huggingface llava-hf/llava-1.5-7b-hf (Link here). The model has 7B parameters and its of Tensortype FP16. This model is made by fine-tuning LLaMA/Vicuna with GPT-generated multimodal instruction-following data. It is an auto-regressive language model, based on the transformer architecture. More information availabe here link.

1.2 Dataset used for further fine tuning

The dataset used for further fine tuning is HuggingFaceH4/llava-instruct-mix-vsft (link here). This dataset has over 272k examples which contain a conversational intruction-response style. It containes both images and text, making it ideal for multi modal fine tuning. For the purpose of our current fine tuning task only 10000 examples were considered.

1.3 Compute platform

The fine tuning and development was done on google colab. The gpu Nvidia A100 with about 20 compute units were used. The training time was approximately 40 mins. Subsequently T4 gpu was used for testing and developing the gradio UI for inference.

1.4 Training stats

The training was done with the powerful A100 with was able to complete 600 steps of training under 40 minutes. The same estimated training time with T4 is 6 hours.

The following are the stats collected during training.

Submitted to AI Planet as part of assignment. Do not distribute.

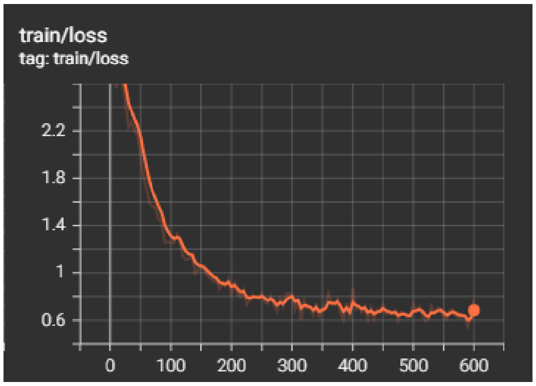


Figure 1: Training loss

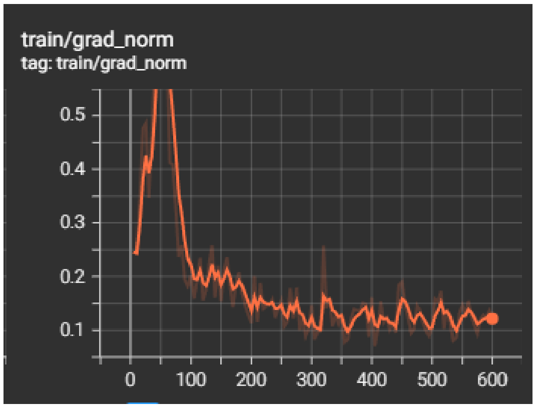


Figure 2: Grad Norm

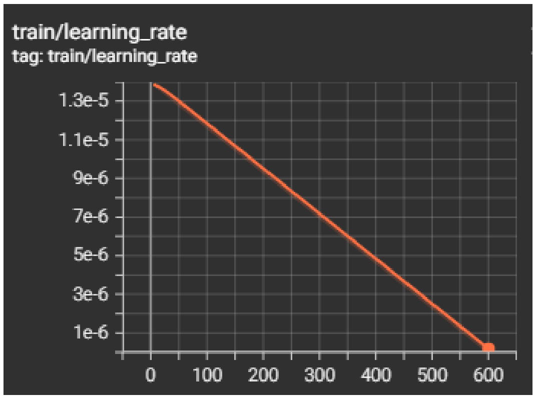


Figure 3: Learning rate

1.5 Chat template

The following is the chat template used for training.

```
A chat between a curious user and an artificial intelligence assistant. \
The assistant gives helpful, detailed, and polite answers to the user's questions. \
{% for message in messages %}{% if message['role'] == 'user' %}\
USER: {% else %}ASSISTANT: {% endif %}{% for item in message['content'] %}{% if item
{% if message['role'] == 'user' %} {% else %}{{eos_token}}{% endif %}{% endfor %}
```

The chat template is used to emulate the following structure.

```
A chat between a curious user and an artificial intelligence assistant. \
The assistant gives helpful, detailed, and polite answers to the user's questions. \
USER: Explain this image. <image>
ASSISTANT: The image shows....... <EOS>
```

1.6 Target modules

The target modules for the fine tuning process where the LORA adapters are added are the following.

- q_proj
- v_proj
- k_proj
- o_proj
- feed_forward

These layers make it about 0.13% of the total parameters.

1.7 Lora Config

the Lora config used is as follows:

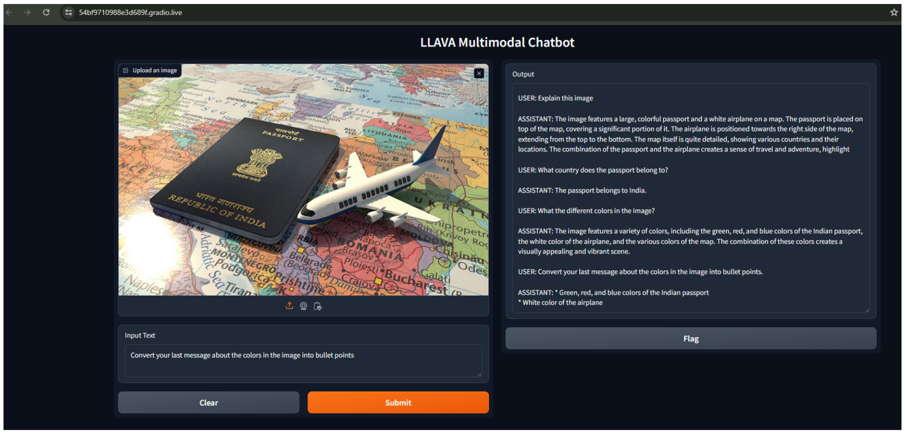
- LORA rank 64
- LORA alpha 16
- LORA dropout=0.1

2 Conclusion

The training of LLMs or LMMs is a resource hungry venture. There are few techniques and methods to mitigate or reduce the compute and memory requirements. However reduction is achieved by these methods the required resources to achieve fine tunning is still expensive. The complex relationships and generalizations can only be captured accurately by such large model with billions of parameters. Thanks to the open source culture everyone has access to the large foundational models which can be fine tuned for down stream tasks.

The learning's from this assignment are valuable. I would like to thank the AI Planet team for giving me this assignment.

3 UI Gradio



Convert your last message about the colors in the image into bullet points

USER: Explain this image

ASSISTANT: The image features a large, colorful passport and a white airplane on a map. The passport is placed on top of the map, covering a significant portion of it. The airplane is positioned towards the right side of the map, extending from the top to the bottom. The map itself is quite detailed, showing various countries and their locations. The combination of the passport and the airplane creates a sense of travel and adventure, highlight

USER: What country does the passport belong to?

ASSISTANT: The passport belongs to India.

USER: What the different colors in the image?

ASSISTANT: The image features a variety of colors, including the green, red, and blue colors of the Indian passport, the white color of the airplane, and the various colors of the map. The combination of these colors creates a visually appealing and vibrant scene.

USER: Convert your last message about the colors in the image into bullet points.

ASSISTANT: * Green, red, and blue colors of the Indian passport
* White color of the airplane

Figure 4: Gradio UI