Optimizing LLaVA: Measuring Mistral Against LLaMA for Visual Instruction Tuning

Albert Miao amiao@ucsd.edu
Ge Gao ggao@ucsd.edu
Aritra Das ardas@ucsd.edu
Hao Zhang haozhang@ucsd.edu

Abstract

Large Language Models (LLMs) have gone from relative obscurity to ubiquitous application in just a few years. Recently, they have demonstrated remarkable ability to generalize to downstream tasks, including visual modality. Our group has modified the Large Language-and-Vision Assistant (LLaVA) to use Mistral as a language model rather than LLaMA. Doing so allows LLaVA to be run at a significant computational discount, reducing model size by 46 percent, while maintaining comparable results. We attribute this improvement to Mistral's advancements in flash attention through sliding-window and grouped query attention. In addition, we provide two subsets of the ScienceQA dataset, reducing the training set from 12726 prompts to 1000 of the most educating through hand-selection and maximization of L2 distance between hidden state representations. Unfortunately, our final performance does not show a significant improvement as opposed to the pertaining stage. However, further exploration into the fine-tuning pipeline should show significant improvement. In the future, we intend to explore further modifications to the LLM training as well as to the visual encoder.

Website: https://ari-03.github.io/LLaVA/
Code: https://github.com/Ari-03/LLaVA

1 Introduction

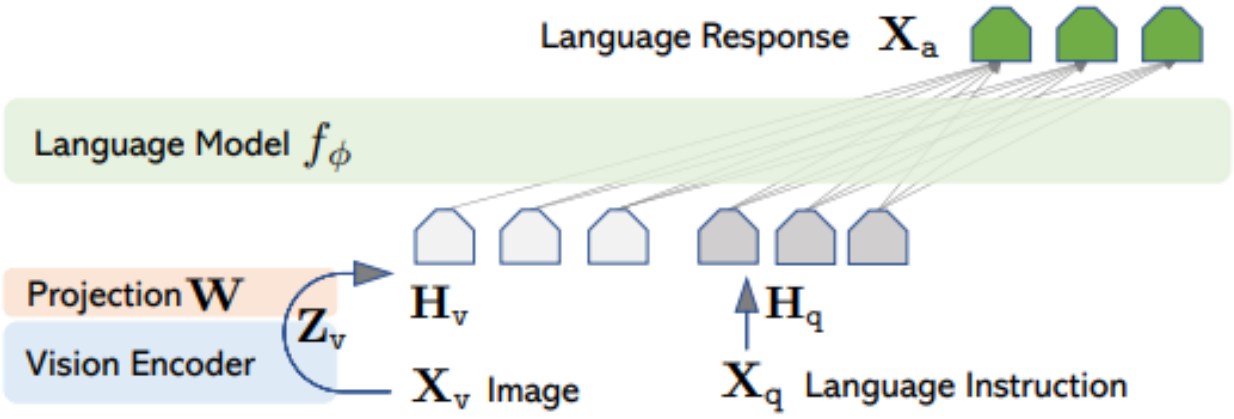
In the rapidly evolving landscape of artificial intelligence, Large Language Models (LLMs) have transitioned from being a niche area of research to becoming a cornerstone of technological advancements across various domains. The versatility of LLMs has been significantly highlighted through their successful application in not only understanding and generating human-like text but also in their ability to extend these capabilities to incorporate visual modalities. This expansion underscores the transformative potential of LLMs, enabling them to interpret and interact with a blend of textual and visual information, thus paving the way for more intuitive and multifaceted AI systems.

Liu et al. (2023) introduces a groundbreaking method for visual instruction tuning that utilizes GPT-4's capabilities to generate data that follows language-image instructions. It presents a Large Language-and-Vision Assistant (LLaVA), an expansive multimodal model that merges a vision encoder with a language model, enhancing overall visual and linguistic comprehension. It explains how a specialized dataset was developed to train and test the model on a wide range of tasks, marking a considerable improvement over previous models. LLaVA's effectiveness is confirmed through benchmarks, such as a multimodal chatbot and the ScienceQA dataset, where it achieves leading accuracy rates, highlighting its exceptional multimodal understanding and ability to follow instructions.

We delve into presenting an inventive modification of the Large Language-and-Vision Assistant (LLaVA), which previously used LLaMA as its foundational language model, now to incorporate Mistral 7B (Jiang et al. 2023). This strategic modification has led to a groundbreaking reduction in computational requirements, slashing the model size by an impressive 46 percent without compromising on performance efficacy. The secret behind this efficiency lies in leveraging Mistral's cutting-edge advancements in flash attention mechanisms, specifically through the implementation of sliding windows and grouped query attention techniques. Not only do we reduce training time by modifying the size of the model, but we show that properly pruning a target dataset can allow a model to achieve comparable performance with a limited size of a dataset. Although our study does not present significant improvement over the pertaining stage, we feel the potential of this approach is palpable, especially given more time for hyperparameter tuning and training. Looking ahead, we aim to delve deeper into refining LLM training methodologies and enhancing the visual encoder component, signaling a promising direction for future research in making LLM more accessible and efficient.

2 Model Architecture

2.1 LLaVA



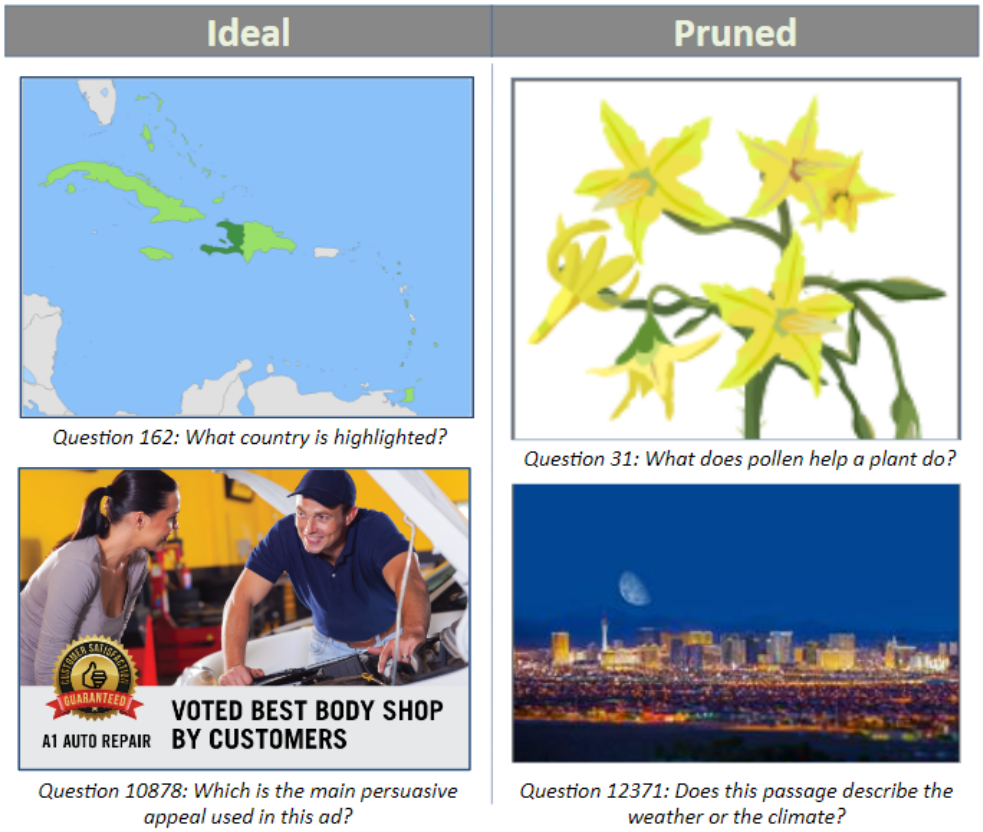
Although the introduction of multimodality to a large language model is a novel and nuanced subject, the actual design is relatively straightforward. The initial image data is passed through a pretrained visual encoder to produce a hidden representation of the information. This visual representation is then acted upon by a large language model, given input prompt, to interpret the information. The process then returns to a typical LLM decoder schema, outputting the language response.

2.2 Mistral and LLaMA comparisons

The selection of the LLM to use has significant effects on the output. In the case of LLaVA, previous experiments have used LLaMA and Vicuna as their foundational model base. In our case, we elect to use Mistral, as the model has previously demonstrated a significant improvement in performance and inference speed when compared to its counterparts. These improvements came about as through a series of improvements to compute efficiency through sliding window attention and grouped query attention.

Sliding window attention, as opposed to previous iterations, minimizes the number of features looked at for each layer, for each token. For example, a token on layer k will only attend to tokens within d distance on layer $k-1$. This allows the transformer to have tokens only focus on local tokens when computing attention score, while also maintain long-range dependencies across all layers. This improvement to inference efficiency drastically lowers inference and training time. Similarly, the introduction of grouped query attention removes previous redundancies of language transformer networks by introducing duplicated key and value weights when performing self-attention. This hierarchical head structure reduces the number of computations necessary to achieve a larger number of heads, balancing quality and speed.

3 Dataset



In order to further reduce training time, we show that our new model can achieve comparable performance to the original with one-sixth of the original fine-tuning dataset size. We reduce the training dataset from 12726 to 1000 data points by hand-selecting data that follows these three guidelines. These guidelines are created to produce the most 'educating' training prompts for the model.

1. **Emphasizes reliance on reference images**: Since the large language model is pre-trained, our expectation for the reduced dataset is to focus on the multimodality of the questions. That is, questions that rely on image data should be included much more often than questions that do not.
2. **Variety**: The dataset must have examples of each possible question to be asked in the test set. This means that each subject should be tested at least once, even if it goes against the other guidelines. In addition, the more a topic is in the raw dataset, the more of that topic should be in the reduced dataset (albeit at a smaller proportion).
3. **Difficulty**: More difficult problems will better train the model. Aspects that make a question more difficult include: text in the reference image, need for careful examination of reference image, complex or arithmetic reasoning, or reliance on uncommon knowledge.

In addition, to better extend this methodology to a larger scale, we propose another method to find the 1000 most semantically distinct training prompts of ScienceQA. First, we perform

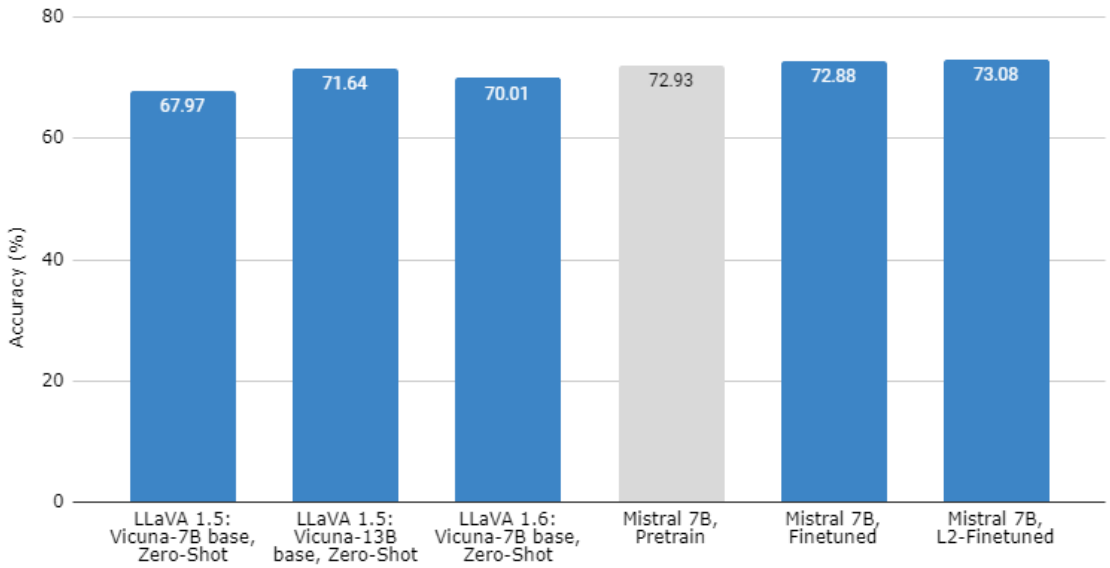
inference on all training prompts using another pretrained multimodal model, in this case using an already trained version of LLaVA with a LLaMA LLM. We record initial hidden states from this method, and select 1000 points from these points that maximizes the minimum L2 distance between their respective hidden states. We compare the results of another model fine-tuned on this secondary dataset.

4 Training

As per LLaVA's recommended training pipeline, we freeze the pretrained weights in both the visual encoder and language model. Then, we train the projection matrix given an input image, instruction, and ground truth output. Doing so allows the projection matrix to align the visual encoder's output with the word embedding of the LLM. We select the 558k language-image instruction-following samples used in LLaVA for pretraining. Following this, we unfreeze the language model weights for fine-tuning, using the two aforementioned datasets.

We train our models using one A100, following each LLM's respective hyperparameters. Our compared model is trained for 1 epoch with an initial learning rate of 1e-3 and a batch size of 64, fine-tuning for 15 epochs with an initial learning rate of 2e-5 and a batch size of 32. We use the Adam optimizer with no weight decay.

5 Results



Unfortunately, our final results were not able to reflect the improvement possible by this methodology. Although we do see a marginal increase in accuracy, it does not reflect an improved understanding of the dataset by the model.

There are several possible explanations. Most immediately, the ScienceQA dataset may contain too complex relationships between prompts and answers that cannot be easily discerned. This issue would suggest that in order to train models properly, a sizable dataset, as per Chinchilla's scaling law, is necessary to see the emergent behavior we've come to expect from large models. Although not all training data given by ScienceQA is fully educative, past a certain point, quantity matters over quality.

However, this issue deserves further examination. Given our empirical understanding of large models, it would be hasty to claim that hand-selection or L2 distance maximization are irresponsible data-pruning methods. Should we be able to extend this project outside the scope of ten weeks, we might be able to refine our fine-tuning pipeline to produce more inspiring results. Given the profound effects that significantly reducing training time can have — providing more accessible models, faster turnaround on development, and deeper insight into the theoretical foundations behind foundational models — we find that this project is worth further exploration.

References

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. "Mistral 7B." https://mistral.ai/news/announcing-mistral-7b/. [Link]

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. "Visual Instruction Tuning."