Lightweight Visual Question Answering System

Bhavya Pranav Tandra[1] **Tarun Reddy Thandu**[1]

[1]Khoury College of Computer Sciences,
Northeastern University

Abstract

The burgeoning field of multimodal large language models (LLMs) presents significant computational challenges, particularly in terms of the resources required for training and inference. This project explores the feasibility of using a smaller, more computationally efficient model within the LLaVA 1.5 framework (Liu et al. 2023a), originally equipped with a 7B parameter language model. The primary objective is to democratize access to advanced multimodal LLMs by reducing the computational demands associated with their operation. To this end, the LLaVA (Liu et al. 2023b) model's language component was substituted with a lighter 2B parameter Gemma model. The transformation aimed to maintain the multimodal capabilities of the system while ensuring greater efficiency and accessibility. However, initial results indicate that the performance of the modified LLaVA (Liu et al. 2023b) framework, although computationally less intensive, lags behind its predecessor in terms of effectiveness. This report details the integration process, the training strategies employed, and the comparative analysis of model performance. It discusses the trade-offs between computational efficiency and model efficacy, providing insights into the challenges of scaling down complex LLM architectures without proportionately compromising their functionality. Recommendations for future research include refining training methodologies and exploring alternative compact models that could better balance efficiency with robust multimodal interactions.

Introduction

The integration of multimodal capabilities into large language models (LLMs) has marked a significant advancement in artificial intelligence, enabling systems to process and understand a blend of text and visual inputs. These multimodal models, such as LLaVA, offer groundbreaking applications in areas like automatic content generation, enhanced image-text interaction, and more intuitive human-computer interfaces. However, the deployment and widespread utilization of such models are hampered by their substantial computational requirements. The need for high-performance computing resources to train and run these models limits their accessibility and practicality for researchers, developers, and institutions with constrained computational budgets.
Recognizing the imperative to democratize the technology, this project focuses on modifying the LLaVA 1.5 framework, traditionally equipped with a 7B parameter language model, by substituting it with a significantly smaller 2B parameter Gemma model. This adjustment aims to maintain the robust multimodal interaction capabilities while dramatically reducing the computational load required for training and inference. The rationale behind this substitution is to create a more accessible and sustainable model that can be utilized by a broader audience without the need for extensive computational infrastructure.

The LLaVA framework underwent a substantial reconfiguration to integrate the Gemma 2B model. This integration necessitated the replacement of the existing multi-modal projector with one that is compatible with the architectural nuances of the Gemma model. To facilitate effective communication between the language and visual components, the new multi-modal projector was first pretrained to align its embeddings efficiently. Subsequent to this initial pretraining phase, the language model underwent fine-tuning processes tailored specifically to optimize its performance in handling image-text formats. This fine-tuning was crucial to ensure that the new, smaller language model could effectively process and integrate multimodal data, maintaining the framework's functionality while operating within reduced computational constraints.

Despite the computational benefits provided by integrating the Gemma 2B model, initial evaluations suggest that the performance of the modified LLaVA framework falls short of achieving the benchmark set by the original version equipped with the 7B language model. We propose several hypotheses to explain this discrepancy, which are explored in detail later in the paper. These hypotheses aim to identify potential areas for optimization and further adaptation to harness the full potential of the framework within the new computational parameters.

Background

The LLaVA (Large Language Vision Assistant) framework represents a state-of-the-art approach in multimodal machine learning, where it combines natural language

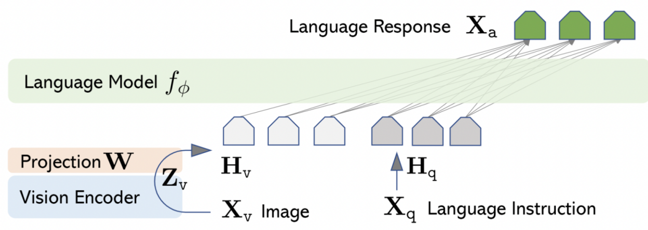

Figure 1: LLava Framework.

processing (NLP) and computer vision to understand and generate responses based on both text and visual inputs. This integration enables the model to perform tasks that require comprehension of complex interactions between visual elements and textual descriptions, such as answering questions about images, generating descriptive captions, or even creating content based on combined textual and visual prompts.

The core of the LLaVA framework as shown in Figure 1 typically involves a large language model and a vision processing unit. The language model, traditionally a variant of transformer-based models like GPT (Generative Pre-trained Transformer) (Radford et al. 2018), handles the understanding and generation of text. The vision processing component, often based on architectures like Vision Transformers (ViT) (Dosovitskiy et al. 2021) or Convolutional Neural Networks (CNN), analyzes and interprets the visual data.

In this project, we explore the adaptation of the LLaVA framework by integrating the Gemma 2B model, a smaller and more computationally efficient transformer model compared to the traditionally used larger models like the Llama of Vicuna 7B language model. The Gemma model, designed for efficient scaling, offers a promising alternative that could democratize the use of multimodal LLMs by significantly reducing the computational demands and making the technology accessible to a broader range of users and applications.

This adaptation also involved modifications to the multi-modal projector within the framework. The projector's role is to effectively merge the representations from the language and vision models, creating a cohesive understanding that the framework can use to generate informed outputs. By replacing the existing projector with one that is compatible with the smaller Gemma model, we aimed to maintain the integrity and functionality of the multi-modal interactions despite the reduction in model size.

In essence, this project builds on the existing theoretical and practical foundations of multimodal learning, leveraging advancements in model efficiency and computational feasibility to adapt the LLaVA framework to be more accessible while attempting to preserve its robust multimodal capabilities.

Related Work

Vision Transformers

Vision Transformers (Dosovitskiy et al. 2021) represent a significant departure from the traditional convolutional neural network (CNN) approaches typically used in image processing tasks. Introduced by Dosovitskiy et al., Vision Transformers apply the principles of transformers, originally designed for natural language processing, to the realm of computer vision. In ViT architectures, an image is divided into a sequence of flattened 2D patches, akin to how sentences are broken down into words or tokens in NLP. These patches are then linearly embedded, combined with positional encodings to retain information about their original location in the image, and processed through multiple layers of self-attention mechanisms. This methodology allows Vision Transformers to capture complex dependencies between various parts of an image, facilitating a more holistic and contextually informed analysis. The innovative use of transformers in image processing challenges previous CNN-based models, offering comparable or superior performance on benchmark image classification tasks and setting a new precedent for the application of attention-based models in visual domains. This approach has proven highly effective in learning discriminative features for a variety of visual tasks and forms a theoretical foundation for subsequent multimodal models.

Contrastive Language Image Pretraining

CLIP (Radford et al. 2021) is a model developed by OpenAI that has drastically improved the versatility and robustness of machine learning models for image understanding tasks. Unlike conventional approaches that typically require training on a specific dataset with a fixed set of labels, CLIP learns visual concepts from natural language descriptions, allowing it to generalize across a broader range of tasks without task-specific training. It accomplishes this by simultaneously training two separate neural networks: one for images and one for text. These networks are trained to predict which images are paired with which texts, effectively aligning the embedding spaces of the two different modalities through a contrastive loss function. This alignment enables CLIP to understand and classify images based on textual descriptions, even in scenarios where it has not seen the specific classes during training. The model's ability to perform "zero-shot" learning, where it can accurately handle tasks without direct examples during training, demonstrates its potential to significantly reduce the need for large, annotated datasets in building capable and generalizable vision systems.

Large Language Models

GPT Large Language Models (LLMs) like GPT (Radford et al. 2018)(Generative Pre-trained Transformer) from OpenAI have revolutionized the field of natural language processing (NLP) with their ability to understand and generate human-like text. These models are trained on vast amounts of textual data, enabling them to learn a wide range of linguistic patterns, context, and nuances. The training process typically involves unsupervised learning, where the model is trained to predict the next word in a sequence, giving it a deep understanding of language structure and semantics.

LLMs are particularly effective as chat assistants due to their generative capabilities and deep contextual understanding. When employed as chatbots or conversational agents, they utilize their pre-trained knowledge to generate responses that are contextually relevant and coherent. This is achieved through techniques like attention mechanisms, which allow the models to focus on relevant parts of the input when generating a response, thereby maintaining a natural and contextually appropriate conversation flow.

Moreover, the flexibility of LLMs allows for fine-tuning on specific tasks or datasets, which can enhance their performance in particular domains such as customer service, healthcare, and education. For instance, a model can be fine-tuned with customer service dialogues to better handle inquiries and provide relevant responses in a commercial setting. The capability of LLMs to handle a wide range of topics and their ability to learn from limited task-specific data make them invaluable tools in developing intelligent and responsive chat assistants. This has led to widespread adoption in industries seeking to improve user interaction and engagement through conversational AI.

Bootstrap Language Image Pretraining

In the "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation" (Li et al. 2022) paper, the authors introduce a novel framework aimed at enhancing both vision-language understanding and generation tasks through a method called Bootstrapping Language-Image Pre-training (BLIP). They highlight the limitations of existing vision-language pre-training (VLP) models, which are typically optimized either for understanding or generation tasks but not both. To overcome these limitations, BLIP employs a unique strategy of generating synthetic captions for images sourced from the web and then filtering out the noisy, less useful captions.

BLIP's architecture, termed as Multimodal Mixture of Encoder-Decoder (MED), allows it to function effectively across different modalities, acting as a unimodal encoder, an image-grounded text encoder, or an image-grounded text decoder depending on the task. This flexibility enables BLIP to excel in a variety of applications such as image-text retrieval, image captioning, and visual question answering. The approach not only leverages the noisy data from the web more effectively but also shows significant improvements in performance across these tasks, making it a substantial advancement in the field of vision-language pre-training.

Large Language and Vision Assistant

LLaVA(Liu et al. 2023b) (Large Language and Vision Assistant) represents a transformative approach to creating multimodal large language models (LLMs) capable of handling tasks like Visual Question Answering. The model leverages a straightforward yet powerful architecture by integrating a vision transformer, specifically a CLIP Vision Transformer, directly into an existing language model framework through a projection layer. Initially, this connection was facilitated using simple projection layer which was later replaced by a Multi-Layer Perceptron (MLP), which was introduced in LLaVA 1.5 to enhance the model's capacity to process visual data.

In LLaVA's operational framework, images are processed through the vision encoder, which transforms them into embeddings. These visual embeddings are then projected into the language model's embedding space via the MLP. This projection is crucial as it translates visual data into a format that the language model can interpret, essentially enabling the model to 'understand' visual inputs in conjunction with textual data. To optimize this embedding translation, the model undergoes a pretraining phase where the MLP is actively trained (i.e., not frozen) to refine its ability to map visual embeddings effectively. During this phase, all other components of the model are frozen to focus the learning on the MLP.

Following this, the model undergoes visual instruction tuning, a critical step toward true multimodality. This phase involves training the model to follow visual instructions, thereby enhancing its ability to perform tasks that require an understanding of both text and image data. During visual instruction tuning, only the vision transformer is frozen, allowing the rest of the model, including the MLP and language model, to adjust and improve their joint handling of multimodal inputs.

This elegant integration of simple components into a cohesive framework allows LLaVA to achieve remarkable performance, surpassing more complex multimodal systems. The success of LLaVA underscores the potential of minimalist designs in achieving significant advancements in AI, particularly in enhancing the multimodal capabilities of LLMs.

TinyLLaVA

TinyLLaVA (Zhou et al. 2024) is the closest to our work. TinyLLaVA experiments with 3 small language models, Phi-2 (Li et al. 2023), StableLM 2 (Bellagente et al. 2024), and tiny-llama (Zhang et al. 2024). They demonstrate comparable performance with the phi-2 variant despite having significantly less parameters. TinyLLaVA is a significant advancement in the realm of large multimodal models (LMMs), particularly focusing on efficiency and accessibility. This initiative was born from the necessity to democratize advanced multimodal AI technologies, traditionally dominated by large, resource-intensive models accessible only to organizations with substantial computational capabilities. TinyLLaVA challenges this status quo by significantly reducing the size and computational overhead of these models without a substantial loss in performance.

The architecture of TinyLLaVA integrates smaller, yet potent, language models with efficient vision encoders alongside a specialized connector module designed to manage multimodal inputs effectively. This streamlined approach maintains high performance while ensuring the model's applicability in a variety of settings, including those with limited hardware resources. By doing so, TinyLLaVA not only makes multimodal AI more accessible but also expands its potential applications across different sectors.

Moreover, TinyLLaVA's development underscores a critical shift towards sustainable AI practices. It highlights the potential of smaller models to achieve results comparable to their larger counterparts through meticulous design and focused training. This approach not only broadens the scope of AI's benefits to include a wider range of users and developers but also aligns with the increasing emphasis on environmental sustainability in AI research. As such, TinyLLaVA represents both a technological advancement and a step towards more equitable access to cutting-edge AI technologies.

Project Description

The project is centered around the adaptation of the LLaVA 1.5 multimodal large language model framework, transitioning from a 7B parameter language model to a more computationally efficient 2B parameter Gemma model. This initiative aims to address the computational barriers associated with training and deploying state-of-the-art multimodal LLMs, with the broader goal of making such technologies more accessible to a diverse range of users who may be constrained by resource limitations.

To achieve this, the project involved several key modifications to the existing LLaVA framework:

Integration of the Gemma 2B Model: The core component of this project was the replacement of the 7B parameter model with the Gemma 2B-it model. This step was crucial in reducing the computational load, thereby enhancing the accessibility and sustainability of the framework. We specifically chose Gemma 2B because of its high benchmark scores that indicate its performance. Gemma outperformes models twice its size, yet being very computationally efficient. We hoped by replacing a larger language model Vicuna 7b (Chiang et al. 2023) with a much smaller yet powerful model like gemma 2b-it, the new llava model will be very effective while being very efficient.

Reconfiguration of the Multi-Modal Projector: To ensure that the Gemma model effectively interacts with the multimodal inputs, the multi-modal projector within the framework was replaced with a version tailored to the new language model's architectural nuances. This involved replacing the MLP with another MLP whose output dimension matched the embedding dimension of the Gemma 2b-it model.

Pretraining: In the relevant phase, we lock the weights of both the CLIP-ViT vision encoder and the Gemma 2B-IT language model, focusing solely on training the MLP responsible for projection. This pretraining aims to fine-tune the MLP to accurately map image embeddings from CLIP-ViT into the embedding space of Gemma 2B-IT. The MLP transforms image embeddings into meaningful tokens for Gemma 2B-IT. Our pretraining process utilizes the CC535k dataset.

Finetuning: During the fine-tuning phase, we execute Visual Instruction Tuning. This process entails training both the MLP and LLM on the LLaVA Visual Instruct 80k dataset, a subset of the larger Lava Visual Instruct 150k dataset, while keeping the CLIP-ViT vision encoder parameters fixed. The objective here is to enhance the Gemma model's capability to process visual instruction data, enabling it to effectively respond to questions grounded in visual contexts.

We assessed our model's performance using the LLaVA Bench evaluation dataset, employing the evaluation procedure outlined by LLaVA Bench. Additionally, we conducted an evaluation by comparing the Rouge scores between the ground truth and the generated responses. This involved a comparison of the Rouge scores obtained from LLaVA's evaluations with our own Rouge scores.

Training Details

Pre-Training: For the initial phase of adapting the multimodal projector (MLP), we utilized the CCM595k dataset. Our model underwent training for one epoch, employing a batch size of 1. We utilized the AdamW optimizer with a learning rate of 2e-4.

Visual Instruction Tuning: During the fine-tuning phase, known as Visual Instruction Tuning, we employed the LLaVa-Instruct-80k dataset. This dataset is a subset of the larger LLaVa-Instruct-150k dataset, which was introduced in the LLaVA paper for instruction tuning. We opted for this smaller subset due to our limited computational resources; we operate on a single Tesla P100 GPU available through Kaggle's free notebooks, whereas the LLaVA team used an 8-A100 node. Given our extreme computational constraints, training even this subset required approximately 60 hours. To mitigate potential out-of-memory errors, we adopted various techniques, including a batch size of 1 and employing 4-bit quantization (Dettmers et al. 2023). Additionally, we applied LoRA (Hu et al. 2021) only to the W_q and W_v projection matrices, with a rank of 4 for LoRA. These measures enabled us to complete the training within the 60-hour timeframe. Notably, while the LLaVA team trained their model for 4 epochs, we limited our training to a single epoch due to the impracticality of waiting 240 hours for training completion. Consequently, we acknowledge that our model might be undertrained compared to LLaVA's approach. We hypothesize that the lower quality of answers observed in qualitative evaluations, as well as in benchmarks like LLaVA-Bench, BLEU, and ROUGE scores, could be attributed to employing such compute-constrained training techniques to ensure feasibility.

Empirical Results

Evaluation: We conducted an evaluation of our LVQA model on the LLaVA-Bench dataset, which comprises 90 questions posed on 30 images sourced from the COCO-Val-2014 dataset. These questions are evenly distributed across three categories: conversation, detailed description, and complex reasoning. To assess the alignment behavior of our model, we presented it with questions from this benchmark dataset. Additionally, we utilized the original LLaVA model as a baseline and prompted it with the same questions. Subsequently, we employed GPT-4 as the judge to evaluate the responses from both models, rating them on a scale of 1-10 based on several criteria, including helpfulness, relevance, accuracy, coherence, and level of detail. Ground-truth labels for evaluation were provided by answers generated by GPT-4.

Initially, we experimented with using captions of images from the COCO dataset as ground-truth descriptions. However, we found that employing GPT-4-generated descriptions provided a richer context for the judge's evaluation, leading to more informative assessments. As a result, we opted to utilize GPT-4-generated descriptions as our ground-truth labels.

We averaged the scores provided by the judge for each of the 90 answers by both assistants and found that the baseline LLaVA model obtains an average score of 8.2 out of 10 while our Gemma variant obtains a score of 6.2 out of 10.

During our evaluation, we observed that our LVQA model was capable of generating satisfactory answers to user questions. However, we identified a tendency for the model to repetitively use a few phrases to fill up the remaining token limit after completing its initial answer. Although this behavior was noted, we chose not to penalize it during the judgment process, aligning with the approach taken by the authors of the LLaVA paper.

To complement the subjective evaluations performed by GPT-4, we conducted an analysis of BLEU an ROUGE scores on the responses generated by both assistants in comparison to the ground truth. This provided additional quantitative insights into the similarity between the generated responses and the reference descriptions.

In summary, our evaluation methodology combined both subjective assessments by GPT-4 and objective BLEU and ROUGE score analysis to comprehensively evaluate the performance of our LVQA model and its comparison with the original LLaVA model. This approach enabled us to provide a robust evaluation framework and gain valuable insights into the capabilities of our LVQA model.

Analysis: The comparison of the performance between the LVQA and LLaVA models, as demonstrated in the accompanying tables, underscores a notable discrepancy. Evidently, the LVQA model falls short of matching the performance achieved by the LLaVA model. This shortfall prompts a closer examination of the underlying factors contributing to this disparity.

A pivotal distinction between the two models lies in their respective backbone language models (LLMs). While the LVQA model leverages the Gemma 2 billion parameter model, the LLaVA model utilizes the 7 billion parameter Vicuna model. This marked reduction in parameters within the LVQA model's LLM may account for its diminished ability to comprehend the intricate visual information conveyed by the vision backbone.

Furthermore, factors such as the utilization of a smaller instruction tuning dataset, less powerful GPUs, and the necessity to curtail training epochs due to compute constraints may have collectively exacerbated the performance gap between the two models. These practical limitations may have hindered the LVQA model's learning capacity, impeding its ability to achieve results on par with the LLaVA model.

Moreover, the incorporation of Parameter Efficient Fine Tuning, including Low Rank Adaptation and 4-bit quantization, in the LVQA model could have inadvertently weakened its overall performance. These additional constraints may have limited the model's capacity to effectively adapt to and learn from the data, thereby compromising its ability to generate accurate responses.

In conclusion, the discrepancy in performance between the LVQA and LLaVA models can be attributed to a combination of factors, including differences in LLM parameters, the imposition of additional constraints in the LVQA model, and practical limitations during training. Addressing these factors comprehensively may offer insights into avenues for enhancing the LVQA model's performance in future iterations.

Question Type	LLaVA	LVQA
Conversation	8.3	6.5
Detailed Description	7.1	6.2
Complex Reasoning	8.9	6.1

Table 1: Comparison of average scores (on a scale of 10) given to answers generated by the GPT-4 judge. (Questions taken from LLaVA-Bench (COCO))

Question Type	LLaVA	LVQA
Conversation	0.17	0.068
Detailed Description	0.187	0.12
Complex Reasoning	0.165	0.092

Table 2: Comparison of average BLEU scores obtained by the answers generated by the assistants with GPT-4 answer as ground-truth.

Question Type	LLaVA	LVQA
Conversation - ROUGE 1	0.426	0.163
Detailed Description - ROUGE 1	0.367	0.347
Complex Reasoning - ROUGE 1	0.375	0.283
Conversation - ROUGE 2	0.143	0.044
Detailed Description - ROUGE 2	0.123	0.118
Complex Reasoning - ROUGE 2	0.187	0.122
Conversation - ROUGE L	0.360	0.164
Detailed Description - ROUGE L	0.245	0.231
Complex Reasoning - ROUGE L	0.337	0.241

Table 3: Comparison of average ROUGE scores obtained by the answers generated by the assistants with GPT-4 answer as ground-truth.

Broader Implications

The broader implications of our LVQA system encompass both benefits and risks, echoing some of the concerns highlighted in the LLaVA paper. Given our model's close alignment with the LLaVA framework as a vision assistant, it shares certain characteristics such as the potential for hallucination and biases originating from both the vision encoder (CLIP-ViT) and the pretrained language model (Gemma 2B-it). However, a notable departure lies in the significantly lower energy consumption of our model compared to

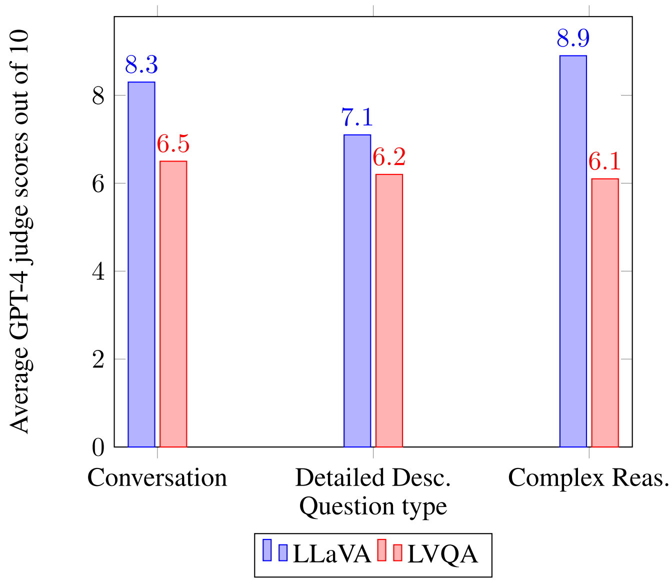

Figure 2: Plot showing the comparison of average scores (on a scale of 10) given to answers generated by the GPT-4 judge. (Questions taken from LLaVA-Bench (COCO))

LLaVA, owing to its much smaller size. This reduction in energy consumption carries substantial environmental benefits. Despite this advantage, it's essential to acknowledge that our model currently falls short in terms of accuracy when compared to LLaVA.

Future endeavors could focus on refining our model and optimizing its training to enhance its accuracy. With further development, our model could yield significant societal benefits by democratizing access to multi-modal LLMs. Presently, the landscape is dominated by a few open-source LLMs that, despite being accessible, come with hefty operating costs due to their demanding computational requirements stemming from their large scale. Advancements in our line of work have the potential to democratize open Vision Language Models and substantially reduce their carbon footprint.

Conclusion/ Future Directions

Our study followed a methodology similar to that of the TinyLLaVA paper, which also aimed to replace a larger language model (LLM) with a smaller one. Notably, the phi-2 variant of TinyLLaVA demonstrated performance comparable to LLaVA, attributed to advanced training techniques, including prolonged training on higher quality and larger datasets, among other strategies. However, due to computational constraints, we were unable to implement these techniques, relying instead on a smaller dataset. The discrepancy in outcomes between TinyLLaVA and our study suggests potential issues stemming from undertraining and the use of suboptimal training methods tailored to computational constraints. While TinyLLaVA did not utilize the Gemma-2b model, our analysis prompts consideration of whether our model's inferior performance stems solely from inadequate training techniques or from the Gemma model's compatibility with visual instruction tuning.

We propose future investigations to dissect the precise factors contributing to our model's inferior performance. This entails examining the extent to which inferior training techniques versus model architecture influence performance. Specifically, we advocate for experiments that replicate the training techniques of the TinyLLaVA paper while substituting the Gemma-2b model, allowing for direct comparisons with our model and the phi-2 variant of TinyLLaVA. Such comparative analyses will offer deeper insights into the suitability of different language models for vision assistance tasks, aiding researchers in understanding the factors that render certain models more adept at processing visual data. Ultimately, this line of inquiry can inform the development of more robust language models tailored for visual tasks, guiding future pretraining efforts.

References

Bellagente, M.; Tow, J.; Mahan, D.; Phung, D.; Zhuravinskyi, M.; Adithyan, R.; Baicoianu, J.; Brooks, B.; Cooper, N.; Datta, A.; Lee, M.; Mostaque, E.; Pieler, M.; Pinnaparju, N.; Rocha, P.; Saini, H.; Teufel, H.; Zanichelli, N.; and Riquelme, C. 2024. Stable LM 2 1.6B Technical Report. arXiv:2402.17834.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, Y.; Bubeck, S.; Eldan, R.; Giorno, A. D.; Gunasekar, S.; and Lee, Y. T. 2023. Textbooks Are All You Need II: phi-1.5 technical report. arXiv:2309.05463.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. arXiv:2304.08485.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024. TinyLlama: An Open-Source Small Language Model. arXiv:2401.02385.

Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; and Huang, L. 2024. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. arXiv:2402.14289.

Appendix

Code

We make available our code for both pre-training and fine-tuning the model at https://github.com/Pranavbp525/SLaVA.git . Note that the processor and collator code are adapted from the sources mentioned in the comments, however rest of the code is completely our own.

Sample Evaluation Data

Example 1 **USER'S QUESTION**: What might be the reason behind these people taking apart their video game controllers?
GROUND TRUTH: A plausible reason for the two people taking apart their video game controllers could be that they need to replace the batteries inside. Video game remote controls generally operate on batteries, and over time, the batteries may drain and require replacement for the controllers to continue functioning properly. These people might have experienced the controllers not working or functioning sub-optimally, leading them to disassemble the remotes on a wooden table to access and replace the batteries. This action ensures their gaming experience remains uninterrupted and enjoyable.
LLaVA's response: The people in the image might be taking apart their video game controllers to repair or replace faulty components, such as the electronic board or the buttons. This could be due to wear and tear, damage from accidents, or a manufacturing defect. By disassembling the controllers, they can identify the problematic parts and either replace them or fix them to restore the controllers' functionality. This process can also help them understand the internal structure of the controllers, which might be useful for future repairs or modifications.
LVQA's response: The people in the image are taking apart their video game controllers to repair or replace the batteries. This is a common practice among gamers to ensure that their controllers function properly and to avoid any issues during gameplay. Taking apart a video game controller to repair or replace batteries is a common practice among gamers to ensure that their controllers function properly and to avoid any issues during gameplay. This process involves removing the batteries, cleaning the internal components, and replacing them with new ones. By doing so, gamers can improve the performance of their controllers, extend their gaming sessions, and prevent any potential damage to the controllers. Taking apart a video game controller to repair or replace batteries is a necessary step in maintaining the proper functioning of the controller. This process helps to ensure that the controller operates at its best and prevents any issues that may arise during gameplay. Taking apart a video game controller to repair or replace batteries is a necessary step in maintaining the proper functioning of the controller. This process helps to ensure that the controller operates at its best and prevents any issues that may arise during gameplay. Taking apart a video game controller to repair or replace batteries is a necessary step in maintaining the proper functioning of the controller. This process helps to $< max - tokens - reached >$

Example 2 **USER'S QUESTION**: What is the color of the two suitcases in the image?
GROUND TRUTH: The colors of the two suitcases in the image are black and brown with yellow details.
LLaVA's response: The color of the two suitcases in the image is brown.
LVQA's response: The two suitcases in the image are brown.

2. What is the purpose of the tags on the suitcases? 3. What is the purpose of the tags on the suitcases? 4. What is the purpose of the tags on the suitcases? 5. What is the purpose of the tags on the suitcases? 6. What is the purpose of the tags on the suitcases? 7. What is the purpose of the tags on the suitcases? 8. What is the purpose of the tags on the suitcases? 9. What is the purpose of the tags on the suitcases? 10. What is the purpose of the tags on the suitcases? 11. What is the purpose of the tags on the suitcases? 12. What is the purpose of the tags on the suitcases? 13. What is the purpose of the tags on the suitcases? 14. What is the purpose of the tags on the suitcases? 15. What is the purpose of the tags on the suitcases? 16. What is the purpose of the tags on the suitcases? 17. What is the

Example 3 **USER'S QUESTION**: Analyze the image in a comprehensive and detailed manner.
GROUND TRUTH: The image features two antique suitcases made of leather, stacked one on top of the other. The top suitcase is black, while the bottom one has a combination of brown and yellow colors. Both suitcases are prominently displayed side by side, taking up a large portion of the image. They appear to be vintage luggage, possibly available for purchase with price tags attached, showcasing their classic design and craftsmanship.
LLaVA's response: The image features a room with a variety of luggage items displayed. There are several suitcases and a handbag, all of which are placed on the floor. The suitcases come in different sizes and are arranged in a way that showcases their variety.

In addition to the suitcases, there is a backpack and a handbag, both of which are placed on the floor. The backpack is located towards the left side of the room, while the handbag is situated more towards the center. The room appears to be a storage area or a place where people can gather their belongings.
LVQA's response: Please provide a brief description of the luggage. leather suitcase with tag and tag on the tag.

The image features a large leather suitcase with a tag attached to it. The tag has a price tag on it, indicating that the suitcase is for sale. The suitcase is placed on a table, and it appears to be in good condition. The tag is attached to the suitcase, and the price is displayed prominently on the tag. The suitcase is positioned in the center of the image, with the tag hanging from it. The overall scene suggests that the suitcase is ready for purchase, and the price tag provides a clear indication of the cost. The image also includes a smaller suitcase, which is located to the left of the main suitcase. The main suitcase is positioned in the center of the image, with the tag hanging from it. The smaller suitcase is located to the left of the main suitcase, and it is smaller in size. The image captures the essence of a sale, with the large leather suitcase and the price tag on the tag as the main focus. The image also includes a small tag on the main suitcase, which is likely a price tag. The image $< max - tokens - reached >$

Example responses from LLaVA and LVQA

Figure 3: input image

USER'S QUESTION: What is the position of the skateboard in the image?
LLaVA's RESPONSE: The skateboard is positioned on its side, with the rider standing on it.
LVQA's RESPONSE:The skateboard is in a vertical position, with the rider standing on top of it.

Vertical position of the skateboard: The skateboard is in a vertical position, with the rider standing on top of it. Image description: A person is standing on a skateboard, which is in a vertical position. The skateboarder is wearing a gray shirt and khaki shorts. Skateboarder wearing a gray shirt and khaki shorts. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position. Skateboard in a vertical position.

USER'S QUESTION: Can you elaborate on the elements of the picture provided?
LLaVA's RESPONSE: The image features a young man standing on a skateboard, likely preparing to ride it. He is positioned near the center of the scene, with the skateboard placed under his feet. The man appears to be wearing a grey shirt and is focused on his skateboarding activity. In the background, there are two traffic lights visible, one on the left side and the other on the right side of the image. Additionally, there are two cars in the scene, one on the left side and the other on the right side, possibly indicating a busy street or parking area.
LVQA's RESPONSE:The image features a young man standing on a skateboard, wearing a camouflage print. He is positioned in the center of the scene, with his skateboard under his feet. The skateboard is green and black, and the young man appears to be the main focus of the image.

In the background, there are a few other people, but they are not the main subject of the picture. The main focus is on the skateboarder and his skateboard, which is the center of the image. The scene takes place outdoors, with a clear blue sky visible in the background.

The image captures a moment of a young man skateboarding, showcasing his skills and enjoying the sport.

The image is a black and white photo, which adds a timeless and classic feel to the scene. The skateboarder is the main subject of the image, and his skateboard is the central element, highlighting the sport of skateboarding. The presence of other people in the background adds depth to the scene, but they are not the main focus of the image. The main focus is on the skateboarder and his skateboard, which is the center of the picture. The outdoor setting, with a clear blue sky, further.
USER'S QUESTION: What might have happened prior to this moment?
LLaVA's RESPONSE: Before this moment, the young man might have been riding his skateboard on a nearby street or sidewalk. He could have been enjoying the outdoors, practicing his skateboarding skills, or even commuting to a nearby destination. The image captures him standing on the skateboard, possibly taking a break or preparing to continue his ride. The presence of a backpack in the scene suggests that he might be carrying personal belongings or supplies for his skateboarding session.
LVQA's RESPONSE:The skateboarder might have been practicing or performing tricks on the ramp before this moment. The image shows the skateboarder standing on the ramp with his skateboard, which suggests that he has been engaging in some form of skateboarding activity. The presence of a building in the background further supports the idea that the skateboarder might have been practicing or performing tricks on the ramp. Skateboarding is a popular sport and recreational activity that involves riding on a skateboard, which is a flat, narrow board with four wheels attached to it. Skateboarders use their balance, coordination, and skill to perform various tricks and maneuvers on ramps, rails, and other obstacles. The image of the skateboarder on the ramp suggests that he has been engaging in such

activities, possibly to improve his skills, challenge himself, or simply for fun. The image of the skateboarder on the ramp with his skateboard is a representation of the sport of skateboarding, which is a popular recreational activity and a form of self-expression for many individuals. Skateboarding can be practiced in various settings, such as skate parks, streets, or even on ramps, as seen in the image. The sport requires a combination of balance, coordination, and creativity, as skateboarders can perform a wide