Visual Question Answering via Prompt Tuning

Written by Samavedam Manikhanta Praphul*[1], Poorna Chandra Vemula*[1]

[1] Khoury College of Computer Sciences, Northeastern University
440 Huntington Avenue
Boston, Massachusetts 02118 USA
samavedam.m@northeastern.edu, vemula.p@northeastern.edu

Abstract

Visual Question Answering is the task of answering the posed question on the image. This is particularly challenging as the model has to know about the image content and then use the information to answer the question. It is significant in the field of LLMs as it will allow the LLM models to digest image data. This was a serious concern in the pre-training of several LLM models. Several people have worked on this area and have published models such as BLIP, CLIP, LLava, where they have trained the models for the task of working with image data along with textual data. Our project is about exploring the possibility to address the task of visual question answering (VQA) using the existing models through different pipelines, prompts and different number of in-context examples. We have found that certain prompts outperform other prompts and adding lot of in-context examples can impact the performance of the model pipelines under our setup conditions. Our project will wave paths for exploring this non fine-tuning based approach for the task of VQA making progress in the multi-modal capabilities.

Introduction

Visual Question Answering (VQA) constitutes an advanced domain at the intersection of computer vision and natural language processing (NLP). This problem involves designing systems that can accurately answer queries about images, necessitating a great understanding of both visual data such as images and its contextual textual relationships. The core challenge in VQA is not merely recognizing objects within an image but involves parsing complex visual scenes and their semantic relationships to generate contextually relevant responses to textual questions.

The evolution of VQA tasks is pivotal as it facilitates a more human-like interaction with technology, wherein machines can comprehend and respond to queries about their visual environment. Such capabilities require an integrated approach that combines robust object recognition, attribute discernment, and relational understanding between elements within an image. These tasks go beyond traditional image processing by demanding a seamless integration of visual and linguistic components.

This research focuses on the utilization of advanced pre-trained models that amalgamate capabilities in both language and visual domains. Traditionally, large language models (LLMs) have been pre-trained predominantly on textual data, which imposes significant limitations on their applicability to multi-modal tasks. Our project seeks to explore and enhance the adaptability of these large models by adding smaller models on top of them to better solve the problem of VQA.

Background

To better understand our work, it is necessary to understand about different models such as BLIP, YOLO, Llama and Mistral.

The BLIP image captioning model is a pre-trained model based on Bootstrapping Language and Image Pre-training framework. This framework is flexible for both vision-language understanding and generation tasks. This model is designed specifically for image captioning task.

The BLIP model is a pre-trained model, that is trained on Common objects in context (COCO) dataset containing images and captions. When provided with an image, BLIP outputs a descriptive caption, employing an attention mechanism to concentrate on various image segments before formulating a word sequence to create the caption.

In terms of structure, the BLIP model includes a patch embedding layer, a transformer encoder, and a language decoder. Initially, the patch embedding layer processes the input image, transforming it into a series of patch embeddings. Following this, the transformer encoder uses multiple self-attention layers to discern the interrelations among these patches, thus forming a contextual image representation. Subsequently, the language decoder utilizes this representation to produce a word sequence that comprises the image's caption.

YOLO (You Only Look Once), is a state of the art realtime object detection model. It is a one stage object detector. This model takes in image as an input and produces the bounding boxes for each of the objects present in the image with their corresponding class names.

Llama 2 is a Large language model released by Meta that operates as an auto-regressive language model, incorporating an optimized transformer architecture. It employs techniques like supervised fine-tuning (SFT) and reinforcement

*These authors contributed equally.

learning with human feedback (RLHF) to better align its outputs with human standards of helpfulness and safety. The model we are using in our work is a 7b fine-tuned model that is optimized for dialogue use cases and converted for Hugging face transformers format (meta-llama/Llama-2-7b-chat-hf).

We are also using an instruction tuned 7b parameter large language model released by Mistral AI. The Mistral-7B-Instruct-v0.2 Large Language Model (LLM) is an instruct fine-tuned version of the Mistral-7B-v0.2.

Related Work

Visual question answering generally involves encoding the image and text, and processing them together. Different approaches with various architecture backbones were proposed over the years.

Malinowski et al[1] proposed an approach where RNNs(LSTMs) handle text inputs and outputs, i.e questions and answers. A CNN pre-trained for object detection is used to produce image features. Both the image and question features are passed first through an encoder LSTM and then the output of this passed through decoder LSTM. The decoder auto-regressively generates the answers.

In a slightly different approach Ren et al. [2] pass the feature vector with image features from two different backbones and text features to a encoder with bidirectional LSTM and the pass it's outputs into a classifier to get a single word from the vocabulary there by formulating the problem as a classification task rather than a sequence generation task.

Recent vision language models mostly use transformer encoder and decoders with various pre-training objectives. Contrastive Learning is used extensively as a pre-training objective for vision models but now been also proved its effectiveness with vision-language models. Recent works including CLIP, ALIGN, CLOOB [3,4] have a text and image encoder linked with a contrastive loss. Here, the objective is to map both modalities(images, texts) to the same feature space such that the corresponding distance between embeddings is minimized.

Further, more recently we have models like BLIP [5] where they use an image encoder with VIT[6] backbone and then a transformer encoder for the question, which is linked to the image encoder through cross attention. Further, there is a transformer decoder again linked with cross attention to the encoder. This has been a very successful architecture and different versions were released fine tuning this model on image captioning, visual question answering.

For the scope of this project, we will be using the BLIP model trained on image captioning and combine it with any of the open source LLMS such as LLaMA, Mistral to accomplish visual question answering. There is a very recent model BLIP2[7] that basically pre-trains, BLIP and Flan-T5 together but we will be working majorly on prompt-tuning due to lack of compute/memory resources. Our Idea is very much inspired by the BLIP2 model, However we want to see what just prompt tuning can do instead of fine tuning the model.

Project Description

We experimented with various pipelines for the task of visual question answering. The dataset that we are considering is the balanced real images of the VQA-2 dataset from the official website visualqa.org which is being maintained by Virginia Tech and Georgia Tech universities. The following are the different approaches we implemented:

1. BLIP and LLM (Llama and Mistral)
2. YOLO and LLM (Llama and Mistral)
3. BLIP,YOLO and a LLM(Llama and Mistral)

1. Pipeline combining BLIP and LLM

   The idea is to use captions of the image from blip-image-captioning model. These captions provide a brief description of the image. Further, a large language model can use these captions as context to answer the question.

   We feed the question and the caption of the corresponding in a prompt and ask the Llama 7b and Mistral 7b model to answer the question based on the image caption.

2. Pipeline combining YOLO and a LLM

   The idea is to use the detections of the objects in the image from a YOLO model. These detections provide a brief description of the image. Further, a large language model can use these detections as context to answer the question. The object detections provide information about differnt object present in the image and their locations. This information helps better answer questions about image, specifically questions related to objects and their relative locations.

   Further, we add the object detections from the image with the given question and prompt the Llama and Mistral 7b instruction tuned models to answer the question using object detections as context.

3. Pipeline combining BLIP,YOLO and a LLM

   In this case, we want to pass both the object detections from YOLO model and image captions from the BLIP model with the given question. Both the detections and image captions should provide greater context about the image for the Language model to answer the given question.

Exploring prompt templates

Further, we experimented with various prompt templates to combine the captions/object detections together with the question in different ways. These are the different prompt templates we are using:

1. Prompt with specific model used

```
"Based on the image caption(provided
by BLIP model) as '{caption}' and
detections(provided by Yolo) as
'{detections}', answer in a single
word the question based on the image
details as question: '{question}'
Answer:"
```

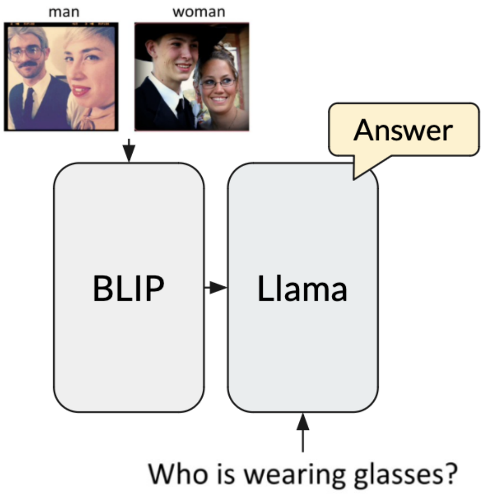


Figure 1: BLIP and Llama pipeline

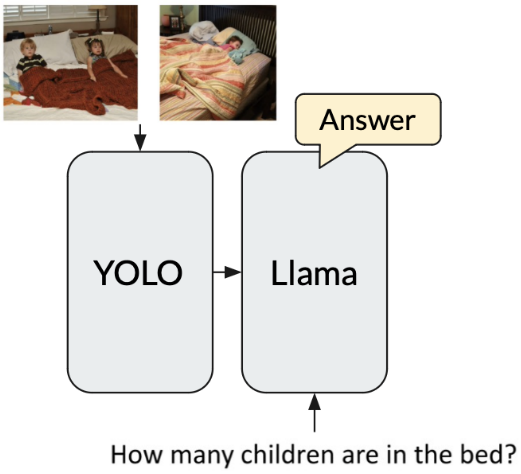


Figure 2: YOLO and Llama pipeline

2. Prompts with Explicit Instruction for Single-Word Answer

```
"Using the image caption '{caption}'
and detected objects '{detections}',
answer the following question with a
single word: '{question}'.
Answer: "
```

3. Direct Question Format

```
"Caption: '{caption}'. Detected
Objects: '{detections}'. What is the
one-word answer to this question about
the image: '{question}'?
Answer: "
```

4. Focused on Image Details

```
"Given the description '{caption}' and
identified elements '{detections}',
provide a one-word response to this
inquiry about the image: '{question}'.
Answer: "
```

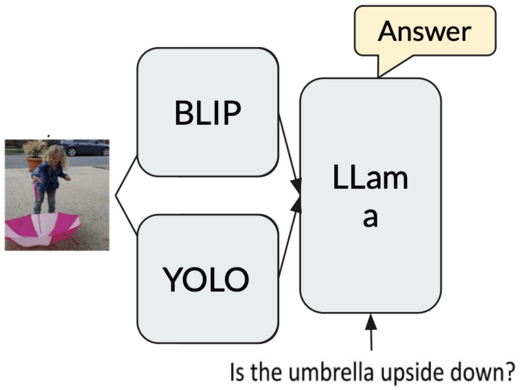


Figure 3: BLIP, YOLO and Llama pipeline

5. Simplified Question-Response Style

```
"From the image caption '{caption}'
and object detections '{detections}',
find the answer to: '{question}'.
Respond in just one word.
Answer: "
```

6. Structured as a Challenge

```
"Challenge: With the caption
'{caption}' and objects detected
as '{detections}', determine the
single-word answer to the question:
'{question}'.
Answer: "
```

7. Asking to answer question using Image content

```
"Answer in a single word for the
question: '{question}' using image
caption: '{caption}' and object
detections: '{detections}'.
Answer:"
```

Each of these prompts, combine the captions, object detections and question in different ways. Some prompts pose it as a challenge and other prompts pose it as a simplified question-response style, direct question format etc. as described above.

In-context learning

In-context learning is a crucial aspect of our project, allowing models to adapt to specific tasks without explicit retraining or fine-tuning. This technique leverages the large language models' inherent ability to generalize from similar examples presented in the prompt. For the Visual Question Answering (VQA) task, we experimented with incorporating different numbers of in-context examples to determine their impact on performance.

To implement in-context learning, we provided the models with a set of examples where each example consisted of prompt with image caption, object detections (if applicable), and question with the answer. These examples, in our experimentation were selected randomly from the dataset so that atleast one of the examppples match the type of questions and visual context that our system would encounter in the testing phase. While in-context learning provides significant advantages, it also presents challenges. The quality of the model's performance can heavily depend on the design of the input prompts and the examples provided.

Results

In this section, we will indicate the results against different experimental setups and key take-aways from it. Note that the results are based on 10k samples and not on the complete dataset due to resource constraints.

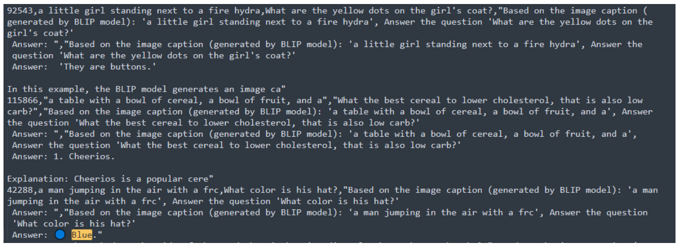
1. Generation Configuration exploration
2. Pipeline exploration
3. Template exploration
4. In Contextual Learning(ICL) performance

Only in case of ICL performance measurement, we have considered a smaller sample size of 100 samples, running for these 100 samples itself took significant time.

Generation Configuration exploration

In this setup, we had experimented with generation configuration and template to be used for the answer generation.

Starting with the default generation configuration and basic simple template stating that you need to answer the question based on the information provided. We have observed that the large language models LLama and Mistral generated the results like a sentence as evidenced for BLIP + LLama in figure 4. This raised two issues, firstly we had to process the sentence long answer for the evaluation and secondly the results would have emoticons, text with cont color as that of the answer as observed in figure 4.



```
92543,a little girl standing next to a fire hydra,What are the yellow dots on the girl's coat?,"Based on the image caption (
generated by BLIP model): 'a little girl standing next to a fire hydra', Answer the question 'What are the yellow dots on the
girl's coat?'
 Answer: ","Based on the image caption (generated by BLIP model): 'a little girl standing next to a fire hydra', Answer the
 question 'What are the yellow dots on the girl's coat?'
 Answer:  'They are buttons.'

In this example, the BLIP model generates an image ca"
115866,"a table with a bowl of cereal, a bowl of fruit, and a","What the best cereal to lower cholesterol, that is also low
carb?","Based on the image caption (generated by BLIP model): 'a table with a bowl of cereal, a bowl of fruit, and a', Answer
the question 'What the best cereal to lower cholesterol, that is also low carb?'
 Answer: ","Based on the image caption (generated by BLIP model): 'a table with a bowl of cereal, a bowl of fruit, and a',
 Answer the question 'What the best cereal to lower cholesterol, that is also low carb?'
 Answer: 1. Cheerios.

Explanation: Cheerios is a popular cere"
42288,a man jumping in the air with a frc,What color is his hat?,"Based on the image caption (generated by BLIP model): 'a man
jumping in the air with a frc', Answer the question 'What color is his hat?'
 Answer: ","Based on the image caption (generated by BLIP model): 'a man jumping in the air with a frc', Answer the question
 'What color is his hat?'
 Answer: 🔵 Blue."
```

Figure 4: Default configuration answers

In order to address this issue, we have modified the configuration to generate only 3 tokens. This was required as few answers were two words at max and this configuration allows us to properly evaluate our pipeline approach. This was later further improved by explicitly stating that the answer needs to be provided in a single word, this change has improved the performance of the pipeline as the model is fed in advance that the model needs to output answer in a single word in addition to the generation restriction. This has significantly improved the performance as captured in figure 1.

Configuration	Exact Match Accuracy	Semantic Match Accuracy
Default	0.1288	0.1522
Max 3 tokens	0.2159	0.2577
Max 3 tokens + single word in prompt	0.4293	0.5209

Table 1: Performance of various configuration of Yolo + Mistral

As the last step, we have added in our prompt that the output should not have any emoticons, special characters apart from plain simple English language words. This change majorly assisted in improving the performances; however, sometimes the performance decreased.

Pipeline exploration

We have explored the different pipelines for the task of visual question answering (VQA) with the configuration of restricting the model to generate for three new tokens only and mentioning that the answer needs to be in a single word, we have explored the different pipelines as the template exploration can be performed in the next stage of experiments. Thus having a clear step by step approach and de-linking the prompt engineering from generation configuration.

Firstly, we have explored the pipeline of BLIP + LLama, YOLO + LLama, BLIP + YOLO + Llama and the results are captured in table 2

Pipeline	Exact Match Accuracy	Semantic Match Accuracy
BLIP + LLama	0.4542	0.5927
YOLO + LLama	0.471	0.63
BLIP + YOLO + LLama	0.461	0.6423

Table 2: Exploration of performance of various Llama pipelines

Similar results for the pipelines of BLIP + Mistral, YOLO + Mistral, BLIP + YOLO + Mistral and the results are captured in table 3

Pipeline	Exact Match Accuracy	Semantic Match Accuracy
BLIP + Mistral	0.4293	0.5209
YOLO + Mistral	0.2923	0.3961
BLIP + YOLO + Mistral	0.4542	0.5927

Table 3: Exploration of performance of various Mistral pipelines

Overall the combined results for both LLama and Mistral langauge models will be figure 5

Based on these results, it is clear that the majority of the performance of the models is through image captioning and large language model. The performance of detections and

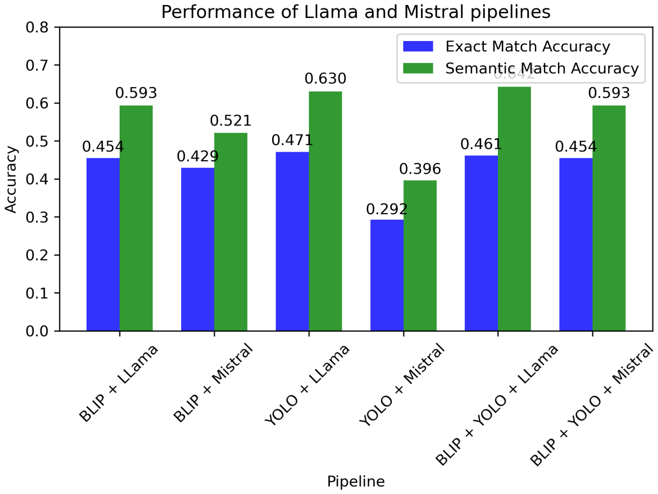


Figure 5: Bar charts of pipeline models

large language model for the task VQA is not so great, this is probably because questions are more focused about the scene in the image in general with little focus about specific objects in the image. Hence we see that the majority of the performance of the pipelines is present for the pipeline consisting of image captions using BLIP and LLama/Mistral to answer the question based on the image captions. However, we can observe that adding object detections as additional information has improved the performance as pipelines containing both image captioning and object detection models has outperformed that has either of them. So with result that it is better to have both BLIP and YOLO for image captioning and object detection in the image respectively should be the way to explore further.

Template exploration

As observed in the previous subsection, we have noted that using both the image captions and object detected in the image will be helpful in the improved performance of the model i.e. pipeline would be BLIP + YOLO + LLM (LLama2 or Mistral). So establishing on this understanding, in this experimentation we had explored 7 different prompt templates as discussed in section of Project Description.

In the table 4 we have the results of both quanitzed and unquantized versions of LLama2
From these results, we observe that template 1 has provided the best performance and template 7 has the worst performance for the LLama2 model. In template 7 we have mentioned the task at the beginning itself and the details are provided at the end, while the template 1 has details at the beginning and task is provided at the end.

Similarly the table 5 has the results for quantized and unquantized versions of the Mistral. However we do not find such patterns in the results with the Mistral and on the contrary to LLama results, the template 7 has the best performance.

Template	Quantized	Exact Match Accuracy	Semantic Match Accuracy
1	Yes	0.43	0.57
1	No	0.43	0.58
2	Yes	0.3	0.41
2	No	0.3	0.39
3	Yes	0.26	0.35
3	No	0.24	0.35
4	Yes	0.39	0.52
4	No	0.39	0.52
5	Yes	0.33	0.43
5	No	0.33	0.43
6	Yes	0.41	0.52
6	No	0.41	0.53
7	Yes	0.11	0.17
7	No	0.11	0.18

Table 4: Exploration of performance of Llama pipelines

Template	Quantized	Exact Match Accuracy	Semantic Match Accuracy
1	Yes	0.476	0.491
1	No	0.476	0.491
2	Yes	0.483	0.501
2	No	0.484	0.102
3	Yes	0.478	0.492
3	No	0.479	0.493
4	Yes	0.472	0.488
4	No	0.473	0.488
5	Yes	0.484	0.498
5	No	0.486	0.501
6	Yes	0.472	0.488
6	No	0.473	0.489
7	Yes	0.516	0.549
7	No	0.519	0.551

Table 5: Exploration of performance of Mistral pipelines

In Contextual Learning(ICL) performance

The large language models have the ability to follow the task which is famously known as in-context learning. Usually providing few examples in the context enhances the model performance, so we plan to explore the performance improvement using the same approach by providing some in context examples. In this exploration, we have explored with 1, 3, 5 in-context examples for YOLO + BLIP + LLama2 using different templates on 100 samples, the results are captured in the table 6 Common observable pattern is that the performance of the model has actually decreased by increasing the number of in-context examples from 3 to 5 along for almost all the templates except the last template, where the performance has increased.

Finally the comparison of our best model against the BLIP which is trained on this dataset for the purpose of the visual question answering and contrast it with the best pipelines that we have added. The results are captured in the table 7. BLIP model is trained on this dataset, so definitely, the

Examples #	Accuracy	Template 1	Template 2	Template 3	Template 4	Template 5	Template 6	Template 7
1	Exact Match	0.46	0.44	0.36	0.44	0.46	0.48	0.37
	Semantic Match	0.57	0.58	0.52	0.56	0.57	0.59	0.5
3	Exact Match	0.44	0.34	0.35	0.44	0.34	0.43	0.23
	Semantic Match	0.53	0.42	0.41	0.54	0.41	0.51	0.34
5	Exact Match	0.17	0.2	0.14	0.22	0.16	0.15	0.24
	Semantic Match	0.23	0.26	0.19	0.24	0.18	0.22	0.29

Table 6: Exploration of ICL examples performance of BLIP + YOLO + LLama2

Pipeline	Exact Match Accuracy	Semantic Match Accuracy
BLIP-VQA	0.9053	0.9611
BLIP + YOLO + LLama2	0.461	0.6423
BLIP + YOLO + LLama2 - 1 ICL	0.48	0.59
BLIP + YOLO + Mistral	0.519	0.551

Table 7: Contrast of our approach vs BLIP baseline

performance of the model is far superior as compared to our approach, this is absolutely possible as the fine-tuned or model trained on this dataset will outperform general model pipelines. But training and fine-tuning models for specific dataset may not be viable solution, so the other alternative is to use the pre-trained general models for the task of VQA on different datasets making them more viable solution.

Broader Implications

As the saying goes 'Picture speaks a thousand words', by exploring the image data, the systems would be able to work a rich data for designing better large systems for society. Specifically, the task of visual question answering would assist the user to query about the product.

By extending it to videos, system would be answer the user queries with ease cutting the tedious task of going through lengthy documentation with the augmented image data along with the video transcript. This task of extending the question answering on image data to video is called as 'Video Question Answering' which focuses on answering the question about the video. In video question answering, we can fetch the specific timestamps for questions such as what time did stranger enter the house in long camera footage, so that the investigation time can be heavily reduced.

By improving the systems to perform the task of visual question answering (VQA), we can have improved search systems with faster and better image retrieval capabilities as the model is now able to augment the user's search along with the rich image data based on specific characteristics of the image. For example, when queried about if the shop is shutdown, the search system can leverage the image data to answer the question rather than following the shop timings being mentioned explicitly on the website.

Similarly, these solutions can be extended to answer the curious questions of students at archaeological sites, museums etc about the history of the excavated object.Most importantly, VQA models would be helpful in creating an inclusive step for the visually impaired individuals by providing information about images and the real world and thus reducing the visual barrier.

Our project explores the idea of using the existing models for the purpose of the VQA, rather than fine-tuning the large models. This approach is significant in terms of reducing the carbon footprint, as stated in the previous papers, the amount of resources used for training the models is far more than the resources used at the inference time. So instead of training/fine-tuning the model for a particular dataset, our approach will aid to navigate the task of visual question answering with lower resource utilization. This will be crucial for environment and will pave way for such similar approaches to equivalent issues.

Conclusion and Future Works

In conclusion, we propose to use a pipeline involving both the description of the image and objects detected in the image for the task of visual question answering as it is observed that it has superior performance compared to pipeline having only either of the data. Based on exploration of template using Llama2, it may be imperative that a template will outperform over the other templates, however based on our results we propose that we need to explore other models as well before coming to such conclusions as different models performs better based on different templates cannot simply disregard template based on single model performance on the templates.

Overall, we achieved the project goal of exploring the capability to solve the task of Visual Question Answering(VQA) by combining SOTA vision models with Large language models without any fine-tuning.

Based on our results, we observe that our pipeline performance is majorly dependent on the image descriptions given by the BLIP model. So improving the description of the image to include the color and location of the objects in the image instead of simple caption will enhance the performance of our pipeline significantly. So, one of the future direction is to generate rich descriptions of these images using larger existing LLMs such as Gemini-vision-pro, which can be used to train the small model pipeline of ours for improved performance.

Further, if we look at our results overall; there is still a significant difference in performance with respect to the current state of the art multi-modal models such as GPT-4, Gemini models. Apart from these models being very large, the other important factor is that these models are trained with multimodal data during pre-training phase.

As discussed earlier, some of the open source models such as BLIP2 and Llava are recently proposed whose fundamental idea is similar to our project. The only difference being that, in both of the above models, then combined a Large language model such as Flan T5 in case of BLIP2 and Llama (or Mistral); in case of Llava with a vision pre-trained model such as BLIP/CLIP and trained them together on vision-language data. This way of combining and training them together drastically improved performance of these models in solving tasks such as VQA where greater visual-language understanding is required.

Our future work would be to come up with a novel architecture which combines a large language model with a vision pre-trained model that can surpass the performance of Llava on various vision-language benchmarks.

Source Code

All the code is available under the link.
https://github.com/PraphulSamavedam/VisualLLM

References

1. Malinowski, M., Rohrbach, M., Fritz, M., 2015. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In: Proc. IEEE Int. Conf. Comp. Vis.
2. Ren, M., Kiros, R., Zemel, R., 2015. Image question answering: a visual semantic em- bedding model and a new dataset. In: Proc. Advances in Neural Inf. Process. Syst.
3. Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).
4. Jia, Chao et al. "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision." ArXiv abs/2102.05918 (2021)
5. Li, Junnan et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." International Conference on Machine Learning (2022).
6. Dosovitskiy, Alexey et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ArXiv abs/2010.11929 (2020): n. Pag.
7. Li, Junnan et al. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." International Conference on Machine Learning (2023).