On the Robustness of Multimodal Large Language Models

Alejandro Aparcedo
University of Central Florida
al402390@ucf.edu

Young Kyun Jang
Meta AI
youngkyun@meta.com

Ser-Nam Lim
University of Central Florida
sernam@ucf.edu

Abstract

Visual Large Language Models (VLLMs) have shown promising capabilities in understanding visual context. In this study, we investigate the performance of a VLLM, LLaVA in a visual question answering task after augmenting the input image with noise, a rotation, crop, etc. We further probe the resilience of VLLMs under adversarial conditions, specifically when the vision encoder is subjected to adversarial attacks. Our findings reveal that our VLLM's ability to understanding visual context is minimally impacted by augmenting the input image. We discover that our VLLM exhibits reduced susceptibility to adversarial attacks. This crucial insight suggests that the integration of a Large Language Model (LLM) as a language decoder, coupled with a vision encoder, could potentially serve as a countermeasure against adversarial attacks.

1. Introduction

Large Language Models such as GPT3 and PaLM have garnered significant attention in the field for their state-of-the-art natural language understanding capabilities [1][3]. However, these models, along with the larger and multimodal model GPT4, remain unavailable to the open source community[8].

Fortunately for the academic community we have open source models such as LLAMA. LLAMA is a LLM that outperforms GPT3 in a variety of benchmarks including zero- and few-shot free-form generation and multiple choice tasks[10]. In the multimodal domain we have models such as Contrastive Language-Image Pre-training (CLIP). CLIP is an open source model that is able to learn image representations through image captioning in pre-training [9].

Given the known vulnerability of CLIP to adversarial attacks we evaluate the image classification abilities of LLaVA when its vision encoder is subjected to these attacks [6]. We perform the Fast Gradient Signed Method on CLIP's vision encoder to obtain adversarial examples [4]. The FGSM works by using the gradients of the network to create an adversarial example. For an input image, the method uses the gradients of the loss with respect to the input image to create a new image that maximises the loss. This maximization of the loss results in CLIP incorrectly classify the input image. We can see how the effect of such an attack on CLIP and LLaVA in Table 2.

In our work we aim to advance the understanding of open source multimodal models. To achieve this we explore an alternative: a small, open-source Visual Large Language Model (VLLM) known as the Large Language and Visual Assistant (LLaVA) [5]. LLaVA connects CLIP as its vision encoder and LLAMA as its language decoder for general visual and language understanding. An illustration of LLaVA's architecture can be seen in Figure 2. LLaVA is a relatively small model as it contains 13× less parameters than GPT3 (175B). Our paper makes the following contributions:

- **Quantitative Evaluation (Visual Question Answering):** Our findings indicate that LLaVA's visual context understanding abilities are minimally affected when the input image is digitally augmented.

- **Quantitative Evaluation (Zero-shot classification)** We find that compared to CLIP, LLaVA displays reduced susceptibility to adversarial attacks.

In the Related Work section we provide provide additional information on the models used. The methods section contains our methodology for benchmarking and evaluating LLaVA's performance. In the experiments we provide explanation for the experiment setup and discuss our results. In our discussion we conclude on the significance of our results and provide insights into future work.

2. Related Work

LLaVA is able to take text and image input and output a text response. LLaVA's performance demonstrates that fine-tuning on a small amount of data can drastically improve performance on the overall task. The task is Visual Question Answering (VQA), LLaVA has been fine-tuned on a VQA dataset and demonstrates improved understanding of visual
Figure 1. Example showing necessary to generate visual question answering samples with text-only GPT4. A customized system prompt is used to prompt GPT4 to generate the question and answer pair. The prompt can be found in the LLaVA GitHub repository.

**Context type 1: Captions**

A group of people standing outside of a black vehicle with various luggage. Luggage surrounds a vehicle in an underground parking area. People try to fit all of their luggage in an SUV. The sport utility vehicle is parked in the public garage, being packed for a trip. Some people with luggage near a van that is transporting it.

**Context type 2: Boxes**

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.262, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

**Response type 1: conversation**

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

**Response type 2: detailed description**

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle. Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

**Response type 3: complex reasoning**

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver’s view or make the passengers uncomfortable during the trip.

context. In Fig. 1 we show an example of the necessary context that is provided to GPT4 to obtain a reference or ground-truth answer for each type of prompt, this process is repeated with 158k image-question pairs. That is, the GPT4 answer serves as the upper bound for VQA performance.

LLAMA is an open source LLM that outperforms GPT3 in a variety of benchmarks while being 10× smaller. LLAMA obtains its ability by training on trillions of tokens and show that it is possible to train a capable LLM with publicly available data.

A major breakthrough in the multimodal domain occurred with the release of Contrastive Language-Image Pre-training (CLIP). CLIP is an open-source model that is able to learn image representations through image captioning in pre-training and transfer its language-image understanding abilities to other tasks such as zero-shot image classification.
To utilize CLIP's zero-shot image classification abilities we create a dictionary of classes for CLIP to choose from and an input image. From this we can obtain CLIP's top-1 prediction for the input image. This process can be seen in the top part of Figure 4. We input an image of a dog along with a dictionary of classes and we obtain CLIP's top-1 prediction "dog". We use this method of zero-shot classification for the evaluation described in sections 3.2 and 4.2.

Researchers have also explored different methods to improve CLIP's zero-shot image classification accuracy. The paper Visual Classification via Description from Large Language Models describes a method which they call classification by description. This method uses descriptors for classification instead of class names. The idea being that by only using the class name, CLIP neglects make use of the rich context of additional information that language affords. This method of classification by description can be explored with VLLMs such as LLaVA to improve its zero-shot classification accuracy.

3. Methods

To investigate the robustness of LLaVA we will evaluate using two metrics: VQA performance and zero-shot image classification. We will use a pre-trained version of LLaVA that has also been fine-tuned on a VQA dataset. We do not fine-tune LLaVA for any of our experiments.

3.1. GPT Evaluation

The robustness of LLaVA is evaluated from analyzing visual question answering performance after we modify the input image through data augmentation. In our study, we utilize the data augmentation techniques outlined in the paper A Simple Framework for Contrastive Learning of Visual Representations, well-established techniques used for self-supervised learning purposes. The techniques we use include crop and resize, gaussian noise, color jitter, etc. For a complete list of all the data augmentation techniques used see our results in Table 1. We expand on these results in section 3.1.

We leverage the power of large GPT models for our evaluation. We use answers provided in the original LLaVA paper, generated with text-only GPT4 as our desired response. For comparison we prompt LLaVA with the same question and associated image. We provide the question and both answers to GPT3.5 and obtain a score in Figure 2 we show an example evaluation using a "conversation" type prompt from the LLaVA VQA dataset. GPT3.5 provides a score based on the helpfulness, relevance, accuracy, and level of detail of the responses. For more information about the evaluation process please see Visual Instruction Tuning.

3.2. Adversarial Evaluation - Classification

To evaluate adversarial robustness, we employ the Fast Gradient Signed Method to perform an adversarial attack on LLaVA's vision encoder (CLIP). To evaluate the effectiveness of our adversarial attack we test our models on several image classification tasks. We test their ability to accurately classify Common Objects in Context in the MSCOCO dataset, different species of animals in ImageNet100, and food items in Food101. Accuracy is recorded both pre- and post-attack for CLIP, providing a measure of its resilience to such perturbations. The outcomes from these tests are compared with the performance of LLaVA on an identical classification task, offering a quantitative assessment of the relative robustness of the two models.

4. Experiments

LLaVA model comes in two small sizes, LLaVA-7B and LLaVA-13B. For both of our experiments we utilize LLaVA-13B. CLIP comes in difference sizes as well, for our experiments we use ViT-L/14 model from HuggingFace library. For our experiments we need to inference LLaVA, to achieve this we utilized 2 H100 80GB GPUs. LLaVA-13B is able to run on a single 80GB H100 but it will take significantly longer to inference. All the necessary versions and installation steps for LLaVA can be found in the LLaVA GitHub repository.
“What is the color of the two suitcases in the image?”

Figure 3. Framework for GPT-assisted evaluation. The “detail” type question and a set of captioned bounding boxes is used to generate our ideal response using text-only GPT4. We prompt LLaVA with the same question and associated image and obtain a response. We input the original question and both answers to GPT3.5 and obtain a relative score for our question.

Assistant 1 provides a helpful and relevant response by correctly identifying the colors of the two suitcases as black and brown with yellow details. The answer is accurate, although it lacks some level of detail in describing the specific colors of the suitcases. Assistant 2, on the other hand, provides a less accurate response by only identifying the color of one suitcase as brown. It fails to mention the presence of a second suitcase and the additional color details. The level of detail in the response is also limited. While the answer is somewhat relevant, it falls short in terms of accuracy and level of detail compared to Assistant 1. Based on these evaluations, Assistant 1 outperforms Assistant 2 in terms of helpfulness, relevance, accuracy, and level of detail. Therefore, Assistant 1 receives a higher overall score of 8, while Assistant 2 receives a score of 6.

Figure 4. Framework for evaluating classification accuracy for CLIP and LLaVA. Input an image and a dictionary of classes into CLIP and to obtain CLIP top-1 prediction. Input the same image and a classification prompt, as seen in the figure into LLaVA and obtain text output. Then, calculate the cosine similarity of LLaVA’s output with each of the same classes we used for CLIP. The class with the highest cosine similarity is considered LLaVA’s top-1 prediction.

This technique is implemented individually, generating a distinct score for each question as well as a relative score for each question type, as determined by the GPT3.5-turbo model.

Our Table 1 shows the scores for each prompt type with no augmentation as well as the scores obtain when we augment the input image. The scores on the table are obtained by averaging the scores of GPT4 for that category and dividing it by the average of the scores for LLaVA in the same category and multiplying it times a 100.

\[
\text{Score}_{\text{category, technique}} = \left( \frac{\text{Average LLaVA score}}{\text{Average GPT4 score}} \right) \times 100
\]
Table 1. Comparison of GPT-4 performance against the original (unmodified) images or one of the data augmentation techniques

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Color Jitter</th>
<th>Crop-Resize</th>
<th>Crop-Resize-Flip</th>
<th>Cutout</th>
<th>Gaussian Blur</th>
<th>Gaussian Noise</th>
<th>Grayscale</th>
<th>Rotation 90</th>
<th>Rotation 180</th>
<th>Rotation 270</th>
<th>Sobel Filler</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversational</td>
<td>94.6</td>
<td>86.4</td>
<td>86.4</td>
<td>86.4</td>
<td>91.1</td>
<td>88.8</td>
<td>90.3</td>
<td>100.4</td>
<td>92.4</td>
<td>91.7</td>
<td>92.4</td>
<td>83.3</td>
<td>89.86</td>
</tr>
<tr>
<td>Detail</td>
<td>95.3</td>
<td>93.9</td>
<td>89.5</td>
<td>83.6</td>
<td>81.7</td>
<td>93.3</td>
<td>78.3</td>
<td>93.9</td>
<td>88.5</td>
<td>84.4</td>
<td>85.6</td>
<td>66.7</td>
<td>85.4</td>
</tr>
<tr>
<td>Complex Reasoning</td>
<td>108.0</td>
<td>107.6</td>
<td>102.9</td>
<td>105.6</td>
<td>104.6</td>
<td>107.6</td>
<td>107.8</td>
<td>107.8</td>
<td>109.8</td>
<td>107.8</td>
<td>106.5</td>
<td>105.9</td>
<td>106.72</td>
</tr>
<tr>
<td>All</td>
<td>99.3</td>
<td>96.7</td>
<td>93.0</td>
<td>90.8</td>
<td>92.5</td>
<td>96.5</td>
<td>92.1</td>
<td>100.7</td>
<td>96.9</td>
<td>94.6</td>
<td>94.8</td>
<td>85.3</td>
<td>93.99</td>
</tr>
</tbody>
</table>

That is, the score for the detail prompts with no data augmentation can be obtained by diving the average LLaVA score by the average GPT4 score (7.567 / 8) * 100 = 94.6

\[ \text{Score}_{\text{detail, original}} = \left( \frac{7.567}{8} \right) \times 100 = 94.6 \]

As evidenced in Table 1, the introduction of augmentation techniques incites only a minimal decrement in performance. Among the question types, ‘complex’ is observed to be the least susceptible to performance degradation in response to these techniques. However, ‘detail’ prompts exhibit a more considerable decline in performance, surpassing 10%. This performance deviation can be attributed to the inherent nature of the prompt type. Upon application of certain data augmentations, such as Cutout, Gaussian Noise, and Sobel Filtering, a significant portion of details can be obliterated, which correlates to the observed performance degradation.

4.2. Adversarial Evaluation - Classification

For the second experiment, we evaluate the performance of both CLIP and LLaVA on several zero-shot image classification tasks. We compare the performance of CLIP to LLaVA to evaluate how effective LLaVA is against adversarial attack. We prompt LLaVA with “Can you fill in the blank, This is a photo of a [ ]” to obtain a response. Using the CLIP tokenizer we obtain a text-embedding from the LLaVA output. We calculate cosine similarity between LLaVA output embedding and each class embedding. The pair with highest cosine similarity is LLaVA Top-1 prediction. This process is outlined in Figure 4, the figure also shows how we obtain CLIP prediction for the image.

In Table 2 we can see that CLIP performs close to its benchmark pre-attack and is greatly affected post-attack. LLaVA on the other hand is not near the benchmark for CLIP but performs better than CLIP during post-attack evaluation. This finding indicates that similar to how classification by description works, LLaVA makes use of additional visual context to generate its response [7]. The impressive accuracy for LLaVA with MSCOCO is likely because LLaVA was fine-tuned on the MSCOCO VQA dataset.

Table 2. CLIP vs LLaVA accuracy before and after adversarial attack

<table>
<thead>
<tr>
<th></th>
<th>CLIP Pre-attack</th>
<th>CLIP Post-attack</th>
<th>LLaVA Pre-attack</th>
<th>LLaVA Post-attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCOCO</td>
<td>93%</td>
<td>2%</td>
<td>83%</td>
<td>73%</td>
</tr>
<tr>
<td>ImageNet100</td>
<td>82%</td>
<td>2%</td>
<td>33%</td>
<td>21%</td>
</tr>
<tr>
<td>Food101</td>
<td>87%</td>
<td>0.5%</td>
<td>32%</td>
<td>18%</td>
</tr>
</tbody>
</table>

5. Discussion

In this study, we have provided empirical evidence demonstrating the robustness of LLaVA ability to understand visual and language context in the face of various data augmentation techniques. We provide quantitative evidence indicating that LLaVA’s performance is minimally impacted because of augmentation. The exploration of the robustness of VLLMs can be deepened by broadening the scope of investigation to include a wider variety of such models, datasets, and data augmentation techniques.

Our secondary experiment demonstrates that a vision encoder like CLIP can operate in conjunction with a LLM language decoder to provide a defense mechanism against adversarial attacks. It is important to note that in its current state, despite its demonstrated robustness, LLaVA does not serve as a direct replacement for CLIP in classification tasks, nor does it present a definitive solution against adversarial attacks.

Nonetheless, there exists a significant potential for improvement, particularly in enhancing classification accuracy with our VLLMs. At the bottom of our Related Work section discuss classification by description which could possibly improve zero-shot classification accuracy for LLava. Should such improvements be realized, VLLMs could potentially evolve to become a viable solution against adversarial attacks. The insights garnered from this research provide an essential foundation for future work in this direction.

References

Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.


