Shifted Reality: Navigating Altered Visual Inputs with Multimodal LLMs

Yuvanshu Agarwal Peya Mowar
Carnegie Mellon University
Pittsburgh, PA
{ypagarwa, pmowar}@andrew.cmu.edu

Abstract

Blind people increasingly use multimodal LLMs, such as in Be My Eyes, to answer their visual questions. However, they often capture irrelevant images with blurry or partial subjects due to a lack of image verification, rendering their questions unanswerable. Through an empirical study, we look into responses by GPT4-V for such "shifted" images. The model, biased by language priors, may inadvertently reveal harmful or misinformed responses to blind people.

1. Motivation

Blind and visually impaired people struggle to identify objects in their surroundings which they encounter in everyday life. To assist them, solutions like the Danish mobile app, Be My Eyes, have incorporated the services of OpenAI's Generative Pre-trained Transformer (GPT) 4-Vision within its Be My AI feature. By simply taking and uploading a picture in the app, a visually impaired person can receive a detailed description of the image and chat further and ask questions for more information.

As multimodal LLMs such as GPT4-V become increasingly prevalent, it is essential to thoroughly examine their image understanding capabilities. Previous research has demonstrated that multimodal machine learning models often exhibit superficial image comprehension, largely influenced by biases in language priors [1]. This issue is particularly pronounced in large language models and poses a significant risk to blind users, who are unable to independently verify the images they capture.

Thus, in this paper, we test GPT4-V against various augmentations of images collected from the VizWiz dataset. We study potential harms and biases in these responses and document interesting observations.

2. Empirical Methods

We formulate the given task: given an image *I* and its augmentation *I'*, we qualitatively evaluate the differences in GPT4-V generated answers for question *Q, A* and *A'*, and contrast them against the ground truth *A**. As shown in Table 1, we produce augmented images from a variety of augmentation factors and their settings. We test images corresponding to each VizWiz answer category and carefully select them to mitigate the effects of any confounding variables that could influence GPT4-V's answer aside from the selected image augmentation control variable.

Images	Factor	Setting			
16	Rotation	0°	90°	180°	270°
16	Brightness	100%	75%	50%	25%
16	Blur Radius	0	5	10	20
8	Cropping			Yes	No

Table 1. Image Augmentation Experiment Parameters

3. Discussion

GPT4-V was resistant to almost all rotations except for the 180 degree rotation. As shown in Table 2, the "Hulk" was incorrectly identified as the "Green Goblin" when rotated 180 degrees. We also observed that GPT4-V identified the number of calories in the V8 V-Fusion beverage even when the number of calories was cropped off. This suggests GPT4-V could leverage previous knowledge in the absence of visual cues to answer questions. While correct in this case, GPT4-V's previous knowledge may be inaccurate and the lack of clarification for the source of its response is concerning. For changes in brightness and Gaussian blurring, no harmful differences in GPT4-V's response have yet been identified. GPT4-V admitted when it was unsure of its response to avoid hallucinations and successfully performed at OCR-related tasks in low brightness like identifying the labelled brand of a swim cap.

References

[1] Douglas Summers-Stay Dhruv Batra Devi Parikh Yash Goyal, Tejas Khot. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2016. 1