Paper Review: Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding

By: Lam Nguyen

1. SUMMARY

This paper discusses how to mitigate object hallucinations in Large Vision-Language Models (LVLMs) using a simple and training free method called Visual Contrastive Decoding (VCD). VCD doesn't require training and is simple. It contrasts output distributions between the original undistorted input images and distorted input images.

VCD reduces over-reliance on statistical bias and unimodal priors, which are the main cause of hallucinations.

In the context of this paper Object hallucination is where LVLMs generate texts based off the image that doesn't actually describe the ground truth image.

The paper then gives a history of the various approaches that have been used before to reduce object hallucination. However, this paper will focus on Visual Contrastive Decoding.

Visual uncertainty in the input image increases the chances of hallucination when generating text based off an image. Examples of uncertainty can be a blurry, dark, or low-resolution image. These kinds of images increase the chances of hallucination. What VCD does is to contrast text outputs from the original image with the outputs from distorted images. This difference can be calculated and used to correct for hallucinations.

The problem with VCD is that indiscriminately applying this correction factor on all outputs can penalize non-hallucinatory output. To address this, am Adaptive Plausibility constraint was used based on a confidence level.

To test the effectiveness of the approach, POPE, MME, LLaVa-Bench and LVLM Baselines were used as test datasets and/or evaluation metrics. These tests showed that improvements were made in reducing hallucination.

2. STRENGTHS

- Does not require training to improve the results. This reduces the need for more computational resources.
- No external tools were required to calibrate the model's output.
- Able to improve accuracy by a few percentage points without requiring more expensive GPU Compute resources.

3. WEAKNESSES

- The dataset had uncertainty introduced only by using Gaussian noise. Other forms of uncertainty should be used such as curating images that are low contrast, darkening the image, reducing resolution, etc…. This would further reduce bias in the outputs.
- The Method could be applied to other types of visual media such as video or 3D models to test whether VCD is a more general method that can be used on all visual media formats
- A discussion on the increase in inference time or an increase in oher forms of computational usage should be discussed between VCD and regular was not made.

4. TECHNICAL EXTENSIONS

- Apply VCD to video
- Apply VCD to 3D Models and simulation
- Add more forms of image distortion to further fine-tune this method

5. OVERALL REVIEW

VCD to reduce object hallucination is a viable method that doesn't require fine-tuning or training the model. Further testing and refinement can be done with this approach by using different forms of data augmentation besides adding noise… As well as testing the principle on other forms of visual media such as 3d Models or Video.