Paper Review: Can We Talk Models Into Seeing the World Differently?

By: Lam Nguyen

1. SUMMARY

Vision Language Models (VLM) offer a way to work with visual content through text/ speech prompting by combining a Large Language Model with a vision encoder.

LLMs and Vision encoders have their own biases on how they perceive information. And this bias is often different from humans.

This paper tested if and how VLMs could have their biases in how they perceive information changed through text prompts. To test this out, the paper did a test on a well-known bias of vision encoders where they focus on more the texture of an object instead of its shape and where small parts of an image are focused on instead of the holistic view of the image. Humans are focused on the exact opposite; in order to recognize what they see, humans first see the shape of an object and also focus on the big picture of an image before zooming into the details.

Through experimentation, it has been shown that through text prompting, a VLM can be steered towards focusing on shape over texture when recognizing an image input. However compared to humans, this bias towards shape is still much lower…

2. STRENGTHS

- It has been shown that through simple text prompt steering, a VLM can be trained to perceive images and produce images differently. No expensive training or fine-tuning is required.
- Opens up avenues to create multiple forms of vision encoders that can be used for specialized tasks from just one VLM. For example, a vision encoder could be used to focus on edges of an image. Another could focus on creating normal maps. Another could focus on texture. Another could focus on high and low contrast areas of an image.

3. WEAKNESSES

- As of now, text steering is still imprecise. For a specialized application using machine learning models where generalization is not required, it is better to train and deploy a specialized machine learning model to perform a specific task as this will be more consistent and precise then a generalized VLM.
- Text Steering to perform specialized vision tasks isn't as efficient and computationally cheap as having a Machine Learning vision encoder that performs a specialized task or set of tasks. For example, some neural networks for image recognition can be run locally on a small and power efficient computer such as a Raspberry Pi. However this is impossible to do for a good Vision Language Model which might require multiple GPUs and a lot of electricity to run.

4. TECHNICAL EXTENSIONS

Beyond Texture/Shape Bias, other facets could be tested such as steering a VLM to create normal maps of an image, focus on lighting, certain colors, gender or any aspects of an image. Possibly even for images that contain multiple spectrums of visual data beyond just visible light such as images that contain Infrared, Near-Infrared, and/or X-Ray… A VLM could be steered towards focusing on certain spectrum or combining certain spectrums of EMF that are picked up by a camera sensor to produce new combinations of images that might be extremely useful to humans, vehicles, or whatever other life forms.

5. OVERALL REVIEW

Vision Language Models can be steered to change their biases towards certain parts of an image. This is very interesting in that it opens up new avenues of study in shifting the bias of a model without having to retrain. This not only applies to low-level image data but also towards things like focusing on different racial biases in image and video, focusing on different speech and sound patterns, focus on different content in text etc…