LVLM Interpret: An Interpretability Tool for Large Vision-Language Models

Gabriela Ben Melech Stan*, Raanan Yehezkel Rohekar*, Yaniv Gurwicz*, Matthew Lyle Olson*
Anahita Bhiwandiwalla*, Estelle Aflalo*, Chenfei Wu+, Nan Duan+, Shao-Yen Tseng*, Vasudev Lal*
*Intel Labs +Microsoft Research Asia

CVPR JUNE 17-21, 2024 SEATTLE, WA

Abstract

We present a novel interactive application aimed towards understanding the internal mechanisms of Large Vision-Language models. Our interface is designed to enhance the interpretability of the image patches, which are instrumental in generating an answer, and assess the efficacy of the language model in grounding its output in the image.

With our application, a user can systematically investigate the model and uncover system limitations, paving the way for enhancements in system capabilities. This work also presents a case study of how our application can aid in understanding failure mechanisms in large multi-modal models.

User Interface

- Gradio interface for multimodal chat
- Supports LLaVa-based models from Transformers
- Interpretability methods for interactive analysis of
  - Raw attention
  - Relevancy map
  - Causal interpretation

Attention Visualization

Attention heatmaps

Visualization of per layer/head raw attention for each generated output token

Relevancy maps

- Per token relevancy map
- Image-text relevancy comparison

Causal Interpretation

From causal interpretation of the attention mechanism, identify a set of image tokens that causally explain a generated text token. Here, explain 'yellow' in response to the question 'what is the color of the man's shirt?".

Case Study

Instances where LLaVA mainly attends to the text tokens and less to the image tokens indicates the model becomes susceptible to input manipulation, altering its responses based on the query with low regard to the image content.

When the generated outputs exhibit a greater relevance to image tokens than to input text, the response appears to remain unaffected by how the question is phrased.