Investigating Multimodal Interaction in Vision Large Language Models

2310041 WEI Houjing

Vision **L**arge **L**anguage **M**odels (VLLMs) extend **L**arge **L**anguage **M**odels (LLMs) by equipping them with the ability to perceive and process both textual and visual data, enabling impressive capabilities such as drafting stories based on images and building a website based on handcrafted images. In recent years, the development of VLLMs has advanced rapidly, yielding a substantial body of remarkable research. For instance, MiniGPT-4 introduces a simple linear mapping to align visual information from a pre-trained vision encoder with a frozen large language model. such linear mapping can be regarded as a **modality connector**. Meanwhile, the LLaVA series follows a similar architecture, utilizing a linear layer or a multilayer perceptron as the modality connector. Through fine-tuned on visual instruction-tuning data, the LLaVA model achieves remarkable multimodal conversational capabilities, setting new state-of-the-art benchmarks in multimodal reasoning tasks. In addition, other works such as InstructBLIP utilize a **Q**uery trans**F**ormer (Q-Former) as the cross-modal interface, wherein the query-based mechanism enables a more selective extraction of visual features tailored to language instruction. When combined with vision-language instruction tuning, the model attains impressive zero-shot performance acrossvar ious vision-language tasks.

As mentioned before, a common architectural paradigm of these VLLMs is a concatenate of a pre-trained **L**arge **L**anguage **M**odel (LLM) with a frozen visual encoder and a learnable cross-modal projector leading to the alignment between multimodal representations. Such framework, referred to as bridge-style architecture, is the basis of most modern VLLMs. In detail, during the inference, VLLMs are fed with both visual and textual inputs, and (1) the frozen image encoder first encodes the image into a set of visual representations, then (2) the visual representations are transferred by a cross-modal projector aiming an alignment with the distribution of typical text token representation, and (3) the projected visual representations (tokens) are then concatenated with some instruction tokens (if any), and fed into the pre-trained LLM for a causal language modeling operation.

Given the remarkable progress of VLLMs on various vision-language tasks, another line of work has emerged, focusing on investigating the inner work of the VLLMs. A pioneering study approached this problem by identifying multimodal neurons within the Transformer's MLP layers and mapping them to semantically related text. Their experiments empirically showed that image tokens, which have been projected into LM embedding space, do not effectively encode interpretable semantics. Similarly, another study found that

language models inherently capture domain-specific visual attributes, while fine-tuning the cross-modal projector does not enhance this capability. More recent research adopts a mechanistic interpretation approach to examine the internal processes of VLLMs. Other works demonstrate that VLLMs encode factual associations within early multi-layer perceptron (MLP) layers and subsequently transfer this information to the final position token through intermediate Multi-Head Self Attention(MHSA) modules. Additional studies contribute to this research domain by investigating the internal workings of VQA in LLaVA, employing methodologies such as log-probability analysis, parameter projections into the unembedding space, and the examination of multimodal information flow across LM layers. These studies have significantly advanced our understanding of the internal mechanisms of VLLMs.

In VLLMs, the LLM is fed with a concatenation of visual token representations and textual token embedding to perform the causal language modeling operation, indicating that the internal mechanisms of the language model, particularly the attention module, are required to leverage information from the visual modality to refine representations. However, existing research has primarily focused on interpreting the projected image tokens or examining how multimodal information flows throughout the text decoder, leaving interaction between the vision token and text token (multimodal interaction) unexplored. Moreover, considering that the image tokens are obtained from encoders pre-trained exclusively on visual data, how they are progressively processed within the LLM representation space is a crucial indicator for revealing the aforementioned multimodal interaction.

Thus, this thesis investigates the interaction between the image token and text token, especially focusing on how image representations evolve along Transformer-based autoregressive text decoders in modern VLLMs. To this end, we first map the projected visual representations into textual tokens by LM heads in LLMs (*logit lens*) to examine how encoded representations from the visual encoder are progressively transformed into language semantics along the layers of the LLMs. Our experiments reveal two key findings:

1. In mid-to-late layers, the hidden states of visual tokens become more semantically aligned with the textual modality compared to the early layers.

2. The correctness of the LM decoding of the visual token's hidden states appears largely independent of the instruction tokens.

Next, we employ the cosine similarity between the hidden states of visual tokens and textual tokens to characterize the magnitude of the multimodal

interaction, using the aligning dynamics of visual token representations toward text token embeddings as an indicator. Specifically, our experimental results, conducted on four models across two datasets, reveal the following findings:

1. Despite differences in designs of cross-modal projectors and the size of LM components, these models exhibit consistent trends in their inter-modal similarity curves, suggesting a general aligning dynamics of visual token representations towards textual token embeddings.

2. Similarity curves exhibit a bimodal pattern and increase rapidly in mid-to-late layers, suggesting a three-stage multimodal interaction dynamics.

3. Regardless of varying types of textual prompt tokens, the layer-wise changes in inter-modal similarity values remain consistent, suggesting that the inter-modal interaction dynamics are largely independent of the specific prompt used.

Moreover, from another perspective, we conducted a layer-wise attention visualization analysis. Our analysis reveals that: (1) Attention scores from instruction tokens (as attention queries) to visual tokens strengthen starting from the middle layers. (2) Certain visual tokens at specific positions receive significantly higher attention than others from textual tokens. Such observation motivated our investigation into the relationship between the number of visual tokens and language modeling loss, aiming to provide empirical insights for balancing the effectiveness (lower forward loss) and efficiency (fewer image tokens) during model inference. In detail, we investigate the impact of varying the number of visual tokens on the model's forward computation loss. Extensive experiments reveal that

1. Once the quantity of image tokens surpasses a certain threshold, loss reduction goes slowly or even stops, indicating that subsequent image tokens may have limited contribution and could be redundant.

2. Across different paraphrased prompts, the curves follow a similar downward trend, suggesting that VLLMs are robust to minor textual variations during inference.

Additional experiments show that the contribution of visual tokens to loss reduction is not uniform, with certain tokens at specific positions playing a significantly greater role. This finding provides valuable insights for optimizing and accelerating inference in VLLMs.

Keywords: Interpretability, Vision Large Language Models, Inference Dynamics.