Challanges on Vision Language Multi-modal model

Juneau Jung, (TA) Jaeseok Byun, (Advisor) Prof. Taesup Moon
M.IN.D LAB
Dept. of Electrical and Computer Engineering
Seoul National University
{sean2ie, wotjr3868, tsmoon}@snu.ac.kr

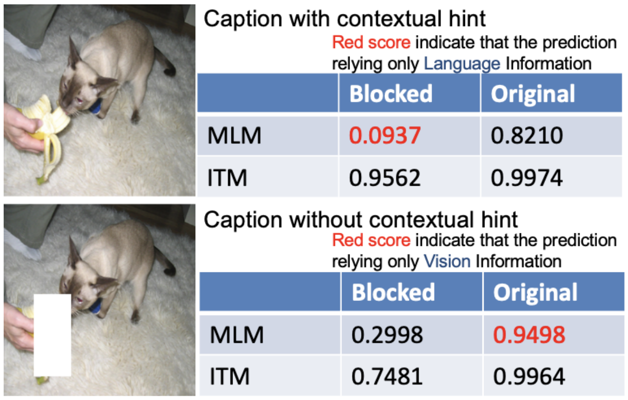
Abstract

Vision-Language pre-trained models have demonstrated remarkable performance in various downstream tasks. However, previous research has predominantly emphasized achieving high performance without conducting comprehensive analyses of the factors contributing to their success or identifying potential limitations. This research aims to bridge this gap by conducting a series of experiments to evaluate the feasibility of the state-of-the-art model. Firstly, we investigate the relative importance of vision-text alignment which has been infeasible by previous approaches. Then, we aim to identify any weaknesses including colors and positional skew within the model that may be attributed to factors originating from the pre-trained network or procedures employed. Through these experiments, we intend to shed light on the underlying reasons behind the models' success and uncover any limitations they may possess.

1. Introduction

In recent years, the development of Vision-Language multi-modal models has witnessed significant progress, revolutionizing various fields such as image captioning, visual question answering, and image-text retrieval. These models [4, 11, 9, 16, 7, 6], by combining visual and textual information, have demonstrated remarkable performance in understanding and generating content that bridges the gap between vision and language domains.

Despite the significant progress and successes achieved by Vision-Language multi-modal models, recent research has raised concerns regarding their limitations and the need for a deeper understanding of their inner workings. Many of these models can be regarded as "black boxes" where their impressive performance is observed, but the underlying reasons remain elusive. Researcher have started to question whether the models are truly comprehending the semantic connections between vision and language or if they are relying on superficial correlations.



Caption with contextual hint
Red score indicate that the prediction relying only Language Information

	Blocked	Original
MLM	0.0937	0.8210
ITM	0.9562	0.9974

Caption without contextual hint
Red score indicate that the prediction relying only Vision Information

	Blocked	Original
MLM	0.2998	0.9498
ITM	0.7481	0.9964

Figure 1. Our surrogate vision-language importance comparison analysis comparing blocked image against the non-blocked image using ALBEF model. (a) Vision Information with contextual hint, (b) Identical vision information without contextual hint

To address these concerns, several studies have delved into investigating the characteristics of these models and the nature of their cross-modal influence. For example, [5] explored the question of whether Vision-Language models primarily excel in Vision-for-Language tasks generating language descriptions for given images or if they can equally perform Language-for-Vision tasks inferring visual concepts from textual descriptions. Besides, [1] proposed a suite visual-language understanding to investigate whether vision and language-based pretraining can enhance performance on text-only tasks that involve implicit visual reasoning. These findings shed light on the biases and limitations present in these models, emphasizing the need for a more comprehensive analysis.

In this paper, we aim to contribute to the existing body of research by conducting an investigation into the feasibility of Vision-Language multi-modal model. We recognize the importance of understanding the underlying reasons behind the success of these models and identifying any potential limitations they may have. To achieve this, we have designed a series of experiments that aim to explore two critical aspects: the relative importance of vision-text alignment and the identification of weak points within the state-of-the-art model.

The contribution of this paper is summarized as follows:

• We conduct a Vision-Language Importance Comparison analysis. With surrogate evaluation of the relative importance between vision and text information in such model, we gain insights into the extent to which accurate alignment of visual and textual information contributes to overall performance.

• We investigate the presence of color bias within a Vision-Language model. We discover that these models often exhibit an overemphasis on color information when processing visual inputs. This bias can potentially lead to skewed interpretations and reliance on superficial visual cues, rather than capturing the semantic essence of the content.

• We examine the existence of positional bias in a Vision-Language model. Positional bias refers to the model's sensitivity to the spatial arrangement of visual elements or textual tokens. Through our experiments, we identify instances where the models exhibit preferential treatment towards certain positions, leading to disproportionate attention allocation or inconsistent performance across different regions of the input.

2. Backgorund and Related Works

2.1. Vision-Language Multi-modal models

ViT [4] made a significant breakthrough by applying the Transformer architecture to image classification tasks. The model divided images into patches and leveraged self-attention mechanisms to capture interdependencies between visual elements. This approach surpassed the previous convolutional neural network (CNN) based methods on several image recognition benchmarks, highlighting the potential of utilizing transformer-based architectures in vision tasks. Similarly, CLIP [11] adopted a contrasting objective to jointly train a vision encoder and a language encoder. This approach allowed the model to learn meaningful representations that could match images and their associated textual descriptions.

In addition to the aforementioned models, other approaches have emerged in the field of Vision-Language multi-modal models, further advancing the understanding and reasoning capabilities across visual and textual modalities. ViLBERT [9] utilizes co-attention mechanism to enable joint reasoning over visual and textual modalities and LXMERT [16] uses pre-trained cross-modality models to learn intermodal connections. These models, extending upon the foundation laid by BERT [3], employ innovative techniques to enable joint reasoning and capture intermodal connections.

Furthermore, [7] focuses on aligning language and visual features through the use of an adaptive attention mechanism. By effectively aligning these modalities, ALBEF enhances its understanding and reasoning capabilities, resulting in state-of-the-art performance on diverse tasks. Another notable approach [6], employs contextual modulation to enhance the vision-language pre-training process. By considering the interdependencies between visual and textual modalities, BLIP achieves improved performance on a range of downstream tasks, including image classification and visual question answering.

These models, along with the previously mentioned examples, collectively showcase the remarkable capability of Vision-Language multi-modal models to comprehend and reason over multi-modal inputs. By incorporating innovative techniques and leveraging the strengths of pre-training and attention mechanisms, these models have consistently achieved state-of-the-art results on diverse benchmarks.

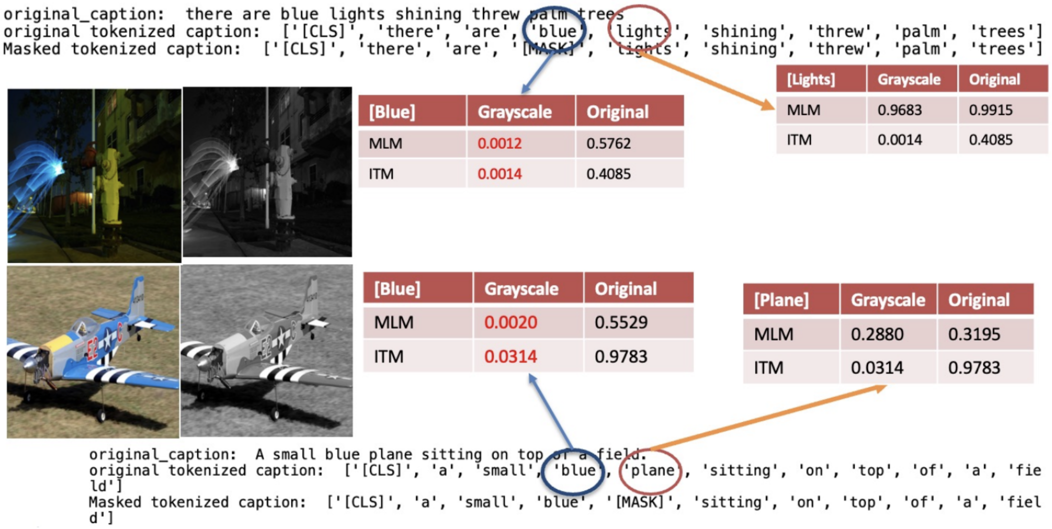
2.2. Feasibility Inspections

Explainable Methods plays a crucial role in understanding the inner workings of multi-modal models, and one of the most intuitive approaches to gain insights into their functioning is through the utilization of saliency maps. [13] involved measuring the gradient of the predicted class with respect to the input image and generating a saliency map by identifying spatial locations with large gradient magnitudes. This method was further enhanced by [14, 15], resulting in sharper saliency maps. These gradient-based saliency methods have demonstrated generalized performance on various datasets.

Class Activation Mapping (CAM) [17] was a significant advancement in generating coarse localization heatmaps. It employed a global average pooling (GAP) layer to compute gradients flowing into the final convolution layer, producing a heatmap that highlights important regions. Grad-CAM [12] extended CAM by eliminating the need for a GAP layer, enabling the computation of fine-grained localization heatmaps. Grad-CAM utilizes gradient information to generate visual explanations from deep networks, providing insights into the model's decision-making process.

Another approaches such as Integrated Gradients (IG) [15], and Layer-wise Relevance Propagation (LRP) [2] have also been employed to unravel the model's interpretation of images, shedding light on the decision-making process and providing valuable insights into the model's behavior.

Cross-modal Influence Addressing concerns regarding the characteristics of Vision-Language multi-modal models and their cross-modal influence, several studies have been



```
original_caption:  there are blue lights shining threw palm trees
original tokenized caption:  ['[CLS]', 'there', 'are', 'blue', 'lights', 'shining', 'threw', 'palm', 'trees']
Masked tokenized caption:  ['[CLS]', 'there', 'are', '[MASK]', 'lights', 'shining', 'threw', 'palm', 'trees']
```

[Blue]	Grayscale	Original
MLM	0.0012	0.5762
ITM	0.0014	0.4085

[Lights]	Grayscale	Original
MLM	0.9683	0.9915
ITM	0.0014	0.4085

[Blue]	Grayscale	Original
MLM	0.0020	0.5529
ITM	0.0314	0.9783

[Plane]	Grayscale	Original
MLM	0.2880	0.3195
ITM	0.0314	0.9783

```
original_caption:  A small blue plane sitting on top of a field.
original tokenized caption:  ['[CLS]', 'a', 'small', 'blue', 'plane', 'sitting', 'on', 'top', 'of', 'a', 'fie
ld']
Masked tokenized caption:  ['[CLS]', 'a', 'small', 'blue', '[MASK]', 'sitting', 'on', 'top', 'of', 'a', 'fiel
d']
```

Figure 2. Our investigation on the color bias in the ALBEF model with comparing original image against gray-scaled images. The MLM scores for grayscaled images dramatically drops even though it has semantics unchanged.

conducted to explore these aspects comprehensively. MM-SHAP (Multimodal Shapley) [10] has been introduced to quantitatively evaluate the extent to which a multimodal model utilizes individual modalities. It leverages Shapley values to provide a performance-agnostic multimodality score, offering a reliable measurement of the proportions in which a multimodal model employs different modalities. By applying MM-SHAP to various Vision-Language models across multiple tasks and datasets, insights into the degree and direction of unimodal collapse where a unimodal model achieves similar accuracy to a multimodal model were obtained.

Additionally [5] delved into investigating whether Vision-Language models predominantly excel in Vision-for-Language tasks, or if they possess equal proficiency in Language-for-Vision tasks, which involve inferring visual concepts from textual descriptions. Besides, [1] proposed a suite visual-language understanding tasks specifically designed to investigate the impact of vision and language-based pre-training on text-only tasks that involve implicit visual reasoning.

However, the traditional vision-language ablation task is not applicable to mainstream models like ALBEF and BLIP because these models align vision and language tokens before fusion, unlike V-L cross-modal models such as LXMERT. Therefore, we suggest an alternative method to explain the influence of vision-language interactions. Moreover, existing models [4, 11, 9, 16] often prioritize achieving higher performance scores on specific evaluation sets, overlooking other problematic aspects. They may rely on shallow cues and fail to consider the deeper semantic meaning of the input. To gain a more comprehensive understanding of how these models work, we also investigate other issues related to multi-modality, such as color and positional information processing.

By delving into these aspects, we aim to uncover the nuances and limitations of vision-Language multi-modal models, facilitating a deeper understanding of their workings and paving the way for further improvements in the field of multi-modal research.

3. Proposed Methods and Experiment Setup

To assess the feasibility and characteristics of the ALBEF model [7], which is currently considered the state-of-the-art vision-language model, we conducted a series of feasibility check experiments using a subset of 50 samples from the COCO-Test dataset [8]. We employed a masked language modeling (MLM) score, similar to BERT [3], to examine the model's contextual text-based token predictions and gain insights into its fine-grained characteristics and image text matching (ITM) score as a metric for representing vision-and-language information alignment.

In masked language modeling, let $\hat{T}$ denote a masked text, and $p^{\text{msk}}(I, \hat{T})$ denote the model's predicted probability for makes token. MLM minimizes a binary cross-entropy loss.

$$L_{\text{mlm}} = \mathbb{E}_{(I,\hat{T})\sim\mathcal{D}} H(y^{\text{msk}}, p^{\text{msk}}(I, \hat{T})) \tag{1}$$

$$L_{\text{itm}} = \mathbb{E}_{(I,T)\sim\mathcal{D}} H(y^{\text{itm}}, p^{\text{itm}}(I, T)) \tag{2}$$

For the image text matching, $p^{\text{itm}}(I, T))$ denotes the predicted a two-class probability, where y^{itm} is a 2-dimensional one-hot vector representing the ground-truth label. These notations and definitions are proposed in ALBEF [7] which we adopted as our baseline model.

4. Experiments

4.1. Vision-for-Language Diagnostic

In this experiment, we aimed to understand the influence of vision-language information in the ALBEF model [7]. We compared the model's performance when provided with visual inputs (images) versus contextual hints (masked tokens) related to the images. Figure 1 examines the model's responses and predictions in these two scenarios, we sought to assess the importance and effectiveness of visual cues in driving the model's language generation capabilities.

4.2. Color Information Comprehension

To investigate the ALBEF model's understanding of color information, we conducted experiments focusing on information loss when presented with grayscale images. By evaluating the model's responses to grayscale inputs as Figure 2, we aimed to assess its ability to comprehend and leverage color cues for various vision-language tasks. Additionally, we detected instances of skewed-color detection, where the model exhibited biases or imbalances in its interpretation and utilization of color information.

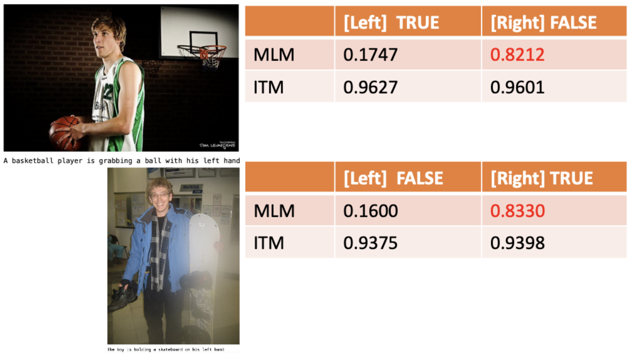
4.3. Positional Information Comprehension

Understanding the ALBEF model's comprehension of positional information was another area of focus in our research. We conducted experiments to assess the model's ability to understand and leverage locational information within images. By examining its responses and predictions related to the spatial arrangement of objects or textual tokens as Figure 3, we aimed to identify any instances of skewed interpretations or inconsistencies in the model's handling of positional information, particularly when dealing with reflected or mirrored images.

Through these proposed methods, we aimed to gain insights into the feasibility and performance characteristics of the ALBEF model. By investigating the influence of vision-language information, comprehending color information, and assessing positional information comprehension, we aimed to identify strengths, weaknesses, biases, and limitations within the model. These findings contribute to a more comprehensive understanding of the ALBEF model and pave the way for future advancements in vision-language multi-modal models.

5. Results

Vision-Language Dominance The results presented in Figure 1 provide clear evidence regarding the feasibility of the model. Intuitively, it can be observed that the importance of vision information outweighs the significance of textual information. This is evident from the substantial decrease in the surrogate ablation score when the images are blocked and accompanied by hint captions, indicating a collapse in performance. In contrast, the non-blocked images without textual hints exhibit minimal signs of collapse. These findings highlight the dominant role of vision information in driving the model's performance, suggesting that it heavily relies on visual cues for effective comprehension and decision-making.


TOM LEWNJEANS
A basketball player is grabbing a ball with his left hand
The boy is holding a skateboard on his left hand

	[Left] TRUE	[Right] FALSE
MLM	0.1747	0.8212
ITM	0.9627	0.9601

	[Left] FALSE	[Right] TRUE
MLM	0.1600	0.8330
ITM	0.9375	0.9398

Figure 3. Our investigation on positional skew in ALBEF model. The MLM prediction score lies more reliability on the false label for the symmetric *mirror-image skew* images

Color-Bias The results in Figure 2 of our analysis indicate that when presented with grayscale images, there is a significant loss of information in terms of the metric used, despite the preservation of object semantics with only the color component being removed. From this observation, we can infer that the model exhibits an excessive fixation on color information, placing a disproportionate emphasis on it in its decision-making process. This finding suggests that the model's performance and decision outcomes might be biased or overly reliant on color cues, potentially overshadowing other important visual features and leading to imbalances in its overall comprehension and reasoning abilities.

Positional Mirror Image Skew One notable observation is the presence of positional *mirror-image skew* in the model's performance. When presented with images featuring symmetrical objects, such as humans with two hands, and two legs, the model tends to struggle in accurately recognizing the position of these symmetrical elements. A clear example of this can be seen in Figure 3, where a person is depicted holding a ski in their left hand. However, due to the image being captured from a mirrored perspective, the model erroneously identifies the object as being in the right hand, leading to a false interpretation.

This finding suggests that the model encounters difficulties in accurately understanding the positional information of symmetrical objects. It highlights a potential limitation in the model's ability to discern and interpret the correct spatial orientation of such elements within an image. Further investigations into this phenomenon can provide valuable insights for improving the model's positional understanding and enhancing its overall performance in scenarios involving symmetrical objects or scenes.

6. Conclusions

The weaknesses observed in the model's performance on downstream tasks raise concerns regarding the potential presence of pre-trained biases. To address these concerns and further improve the model, future work should focus on conducting a comprehensive investigation of the dataset to identify and assess the existence of biases. This exploration will contribute to a deeper understanding of the limitations and challenges associated with the model's training data.

Furthermore, efforts should be directed towards the development of a more robust model that can effectively mitigate pre-trained biases. By enhancing the model's ability to handle biases, we can promote fair and unbiased decision-making processes and improve the overall performance of the vision-language multi-modal system.

In addition to these research directions, there may be other avenues worth exploring to advance the field. These could include investigating novel techniques for interpretability and explainability in vision-language models, exploring alternative training strategies, or exploring the potential of transfer learning to address the identified limitations.

By addressing the concerns related to biases and striving for a more robust and unbiased model, we can unlock the full potential of vision-language multi-modal models and facilitate their deployment in a wide range of real-world applications.

References

[1] Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6778–6788, 2023.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[7] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[10] Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. *arXiv preprint arXiv:2212.08158*, 2022.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[14] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[16] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[17] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.