VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers

Estelle Aflalo*
Intel Labs
estelle.aflalo@intel.com

Meng Du*
Intel Labs, UCLA
mengdu@ucla.edu

Shao-Yen Tseng
Intel Labs
shao-yen.tseng@intel.com

Yongfei Liu
Microsoft Research
liuyf3@shanghaitech.edu.cn

Chenfei Wu
Microsoft Research
chewu@microsoft.com

Nan Duan
Microsoft Research
nanduan@microsoft.com

Vasudev Lal
Intel Labs
vasudev.lal@intel.com

Abstract

Breakthroughs in transformer-based models have revolutionized not only the NLP field, but also vision and multimodal systems. However, although visualization and interpretability tools have become available for NLP models, internal mechanisms of vision and multimodal transformers remain largely opaque. With the success of these transformers, it is increasingly critical to understand their inner workings, as unraveling these black-boxes will lead to more capable and trustworthy models. To contribute to this quest, we propose VL-InterpreT, which provides novel interactive visualizations for interpreting the attentions and hidden representations in multimodal transformers. VL-InterpreT is a task agnostic and integrated tool that (1) tracks a variety of statistics in attention heads throughout all layers for both vision and language components, (2) visualizes cross-modal and intra-modal attentions through easily readable heatmaps, and (3) plots the hidden representations of vision and language tokens as they pass through the transformer layers. In this paper, we demonstrate the functionalities of VL-InterpreT through the analysis of KD-VLP, an end-to-end pretraining vision-language multimodal transformer-based model, in the tasks of Visual Commonsense Reasoning (VCR) and WebQA, two visual question answering benchmarks. Furthermore, we also present a few interesting findings about multimodal transformer behaviors that were learned through our tool.

1. Introduction

Since transformers were introduced in Vaswani *et al.* [30], not only have they seen massive success in NLP applications, their impact on computer vision and multimodal problems has also become increasingly disruptive. However, the internal mechanisms of transformers that lead to such successes are not well understood. Although efforts have been made to interpret the attentions [8] and hidden states [22] of transformers for NLP, such as BERT [12], investigations in the mechanisms of vision and multimodal transformers are relatively scarce, and tools for probing such transformers are also limited. Given the fast-growing number of successful vision and multimodal transformers (*e.g.*, ViT [13] and CLIP [25]), enhanced interpretability of these models is needed to guide better designs in the future.

Past research has shown the importance of interpreting the inner mechanisms of transformers. For example, Clark *et al.* [8] found certain BERT attention heads specialized in handling certain syntactic relations, as well as interesting ways in which BERT attention utilizes special tokens and punctuation marks. Additionally, Lin *et al.* [21] showed that the linguistic information encoded in BERT becomes increasingly abstract and hierarchical in later layers. These studies provide valuable insights into the functions of various elements in transformer architecture for NLP, and shed light on their limitations.

This paper presents VL-InterpreT[1], which is an interactive visualization tool for interpreting the attentions and hid-

*Equal Contributions

[1] A screencast of our application is available at https://www.youtube.com/watch?v=4Rj15Hi_Pdo. Source code and a link to a live demo: https://github.com/IntelLabs/VL-InterpreT

den representations of vision-language (VL) transformers. Importantly, it is a single system that analyzes and visualizes several aspects of multimodal transformers: first, it tracks behaviors of both vision and language attention components in attention heads throughout all layers, as well as the interactions across the two modalities. Second, it also visualizes the hidden representations of vision and language tokens as they pass through transformer layers.

The main contributions of our work are:

- Our tool allows interactive visualization for probing hidden representations of tokens in VL transformers.
- Our tool allows systematic analysis, interpretation, and interactive visualization of cross- and intra-modal components of attention in VL transformers.
- As an application of VL-InterpreT, we demonstrate multimodal coreference in two analyses: 1) how contextualized tokens in different modalities referring to the same concept are mapped to proximate representations, and 2) how attention components capture the conceptual alignment within and across modalities.

2. Related Work

As deep learning models flourish, many tools and methods have been proposed to offer insight into their inner workings. Some methods are general-purpose [24, 32], while others are nuanced for specific models such as CNNs [35, 36] or RNNs [16, 17, 27]. In transformers, the introduction of attention not only helped improve performance, but also served as an additional component towards interpretability.

Interpretability of NLP transformers was initially approached through the analysis of attention to capture its alignments with syntactic or semantic relationships [8, 31]. Following this, subsequent works introduced additional functionalities including visualizations of hidden representations, task matching of attention heads, aggregate metrics, and interactive datapoint analysis [15, 18, 20, 28]. While common in allowing a user to understand the inner workings of transformers, each tool introduces novel applications. For example, LIT [28] enables probing for bias through examination of coreferences in counterfactuals. InterpreT [18] allows tracking of token representations through the layers and offers users the ability to define new metrics to identify coreference relationships in attention heads. Additionally, T^3-Vis [20] focused on allowing users to improve transformer training by integrating the training dynamics in their visualization tool.

Interpreting vision transformers, such as those for object detection [4, 13] or image captioning [10, 19], has also become increasingly popular. Cordonnier *et al*. [9] showed that the first few layers in transformers can learn to behave similarly to convolutional layers, and demonstrated the filter patterns through visualization of image-to-image attention. As illustrated in this paper, the attention mechanism in transformers is a natural gateway to understanding vision models, as the heatmaps of attention can be used to highlight salient image regions. Furthermore, Chefer *et al*. [7] proposed a method for visualizing self-attention models by calculating a LRP [1]-based relevancy score for each attention head in each layer, and propagating relevancies through the network. The end result is a class-specific visualization of image regions that led to the classification outcome.

Multimodal interpretability has, up until now, mainly entailed using probing tasks to study the impact of each modality on the responses generated by the model. These probing tasks aim to quantify the information captured in the hidden representations by training classifiers, or applying metrics to embeddings at different points in a model. For instance, Cao *et al*. [3] proposed various probing tasks to analyze VL transformers, where the authors observed modality importance during inference, and identified attention heads tailored for cross-modal interactions as well as alignments between image and text representations. Additionally, other works have proposed probing tasks to interpret VL transformers for aspects such as visual-semantics [11], verb understanding [14], and other concepts such as shape and size [26]. However, a disadvantage of probing tasks is the amount of work: additional training of the classifiers is often required, and specific task objectives must be defined to capture different embedded concepts. Finally, most aforementioned works require image-caption pairs as input, and are therefore not best suited for interpreting multimodal transformers in tasks such as visual question answering.

Recently, a first attempt at explaining predictions by a VL transformer was proposed in [6]. There the authors constructed a relevancy map using the model's attention layers to track the interactions between modalities. The relevancy map is updated by a set of update rules that back-propagates relevancies of the prediction result back to the input. This map is very useful in understanding how model decisions are formed, but a more comprehensive interpretation for other aspects of VL transformers is still needed.

Our proposed tool, VL-InterpreT, differs from previous works in that it interprets various aspects of multimodal transformers in a single interface. This interactive interface allows users to explore interactions between tokens in each modality from a bottom-up perspective, without tying to task-specific inputs and outcomes. To the best of our knowledge, this is the first interactive tool for interpreting multimodal transformers.

3. System Design

3.1. Workflow

VL-InterpreT is designed as a two-stage workflow: First, the attentions and hidden states of a given multimodal model are generated and saved for a set of examples (in this case, 100 examples). Next, the saved data, along with the metadata of the corresponding examples, are loaded into our tool to enable visualizations of the inner workings of the model. The workflow of VL-InterpreT is shown in Figure 1, and the user interface is shown in Figure 2. Different from interpretability tools for NLP transformers, VL-InterpreT addresses the analysis and interpretation of the following properties of multimodal transformers:

- **Input:** In general, multimodal transformers are able to process inputs originating from different modalities, *e.g.*, video, audio, or language. Here, we only consider VL transformers where the input is composed of visual and textual tokens. These tokens are mapped into a shared space, allowing for concept-level alignment between the two modalities.
- **Attention:** Because of the bi-modal nature of the input, the resulting attention can be split into four components: language-to-language, vision-to-vision, vision-to-language, and language-to-vision. An illustration of these components is shown in Figure 1.
- **Hidden states:** In each layer, the transformer produces as many hidden states as the number of input tokens. Each input token is embedded as a *d*-dimension vector after processed by each layer.

3.2. Visualizations

3.2.1 Attention heads components

In this section, we describe each attention component and how VL-InterpreT visualizes them interactively. The attention matrix in a VL transformer, as they are loaded in VL-InterpreT, is of size $(N_{layers}, N_{heads}, L_v + L_l, L_v + L_l)$, where N_{layers} and N_{heads} correspond to the number of layers and heads, respectively, L_v to the number of visual tokens, and L_l to the number of text tokens.

- The **Language-to-Vision** attention component (L2V) of size $(N_{layers}, N_{heads}, L_l, L_v)$ reflects the text tokens' dependency on the visual tokens. This partition of the attention contains the attention scores calculated from the dot product of the query vector based on the selected image patch and the key vectors from text tokens. These attention scores are the weights given to value vectors of text tokens when summing for the updated representation of a specific image patch in the next layer. A user can select any attention head and image patch in the interface of our tool, and the corresponding L2V attention weights are displayed as a heatmap overlaid on the input text, to visualize how much each text token contributes to the updated image patch representation (see Figure 7).
- The **Vision-to-Language** attention component (V2L) of size $(N_{layers}, N_{heads}, L_v, L_l)$ reflects the visual tokens' dependency on the text tokens. This component of attention in a VL transformer arises through the query-key dot product where the query vectors are computed from text token embeddings, and the key vectors are computed from the image token embeddings. A user can select a specific attention head and a text token, and the corresponding attention scores will be overlaid onto the image as a heatmap (see Figure 6). This visualization helps users understand the relative contributions of various parts of the image to the updated representations of the text tokens. The interactive application also allows users to play an animation, in which the heatmap over the image is automatically displayed in a sequence for each word.
- The **Language-to-Language** attention component (L2L) of size $(N_{layers}, N_{heads}, L_l, L_l)$ corresponds to the attention mapping in NLP Transformers. This component visualizes how all text tokens attends to each individual token of the input sentence (see Figure 6).
- The **Vision-to-Vision** attention component (V2V) of size $(N_{layers}, N_{heads}, L_v, L_v)$ is analogous to language-to-language attention, but in the visual space. It represents the attention between one visual tokens and all visual tokens, including itself. Similar to the V2L component, an attention vector (of size $(L_v, 1)$) here in a given head can also be translated into a heatmap and overlaid onto the image. This visualization is useful for identifying the contributions of different image patches to the updated representation of the image patch selected by a user.

3.2.2 Attention head summary

This functionality allows users to visualize a head summary plot containing statistics of the attentions calculated for all heads and layers. For an attention matrix of size $(N_{layers}, N_{heads}, L_v + L_l, L_v + L_l)$, the head summary computes statistical metrics over the last two dimensions, resulting in a plot of size (N_{layers}, N_{heads}).

- The **mean** attention in an attention head is generated by calculating the average of the corresponding attention matrix, while some tokens can be excluded from

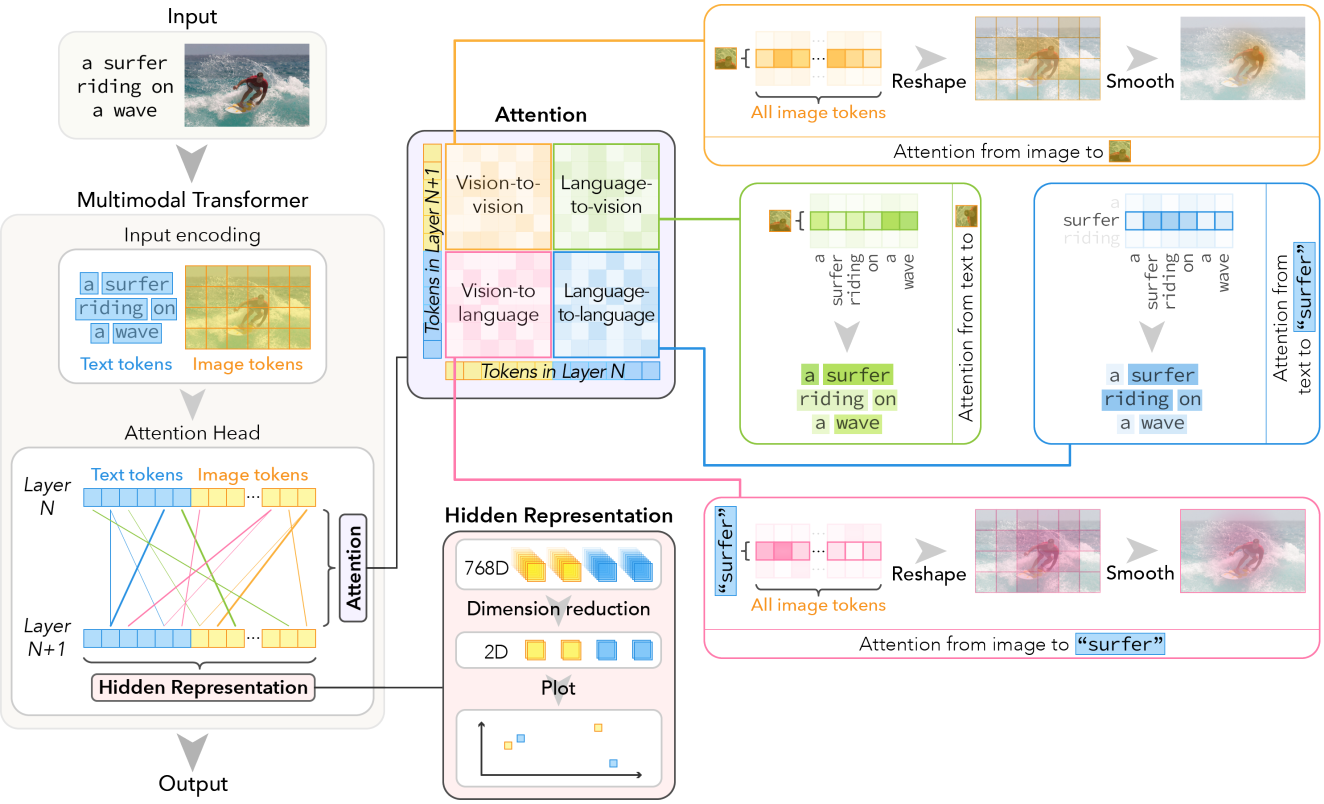
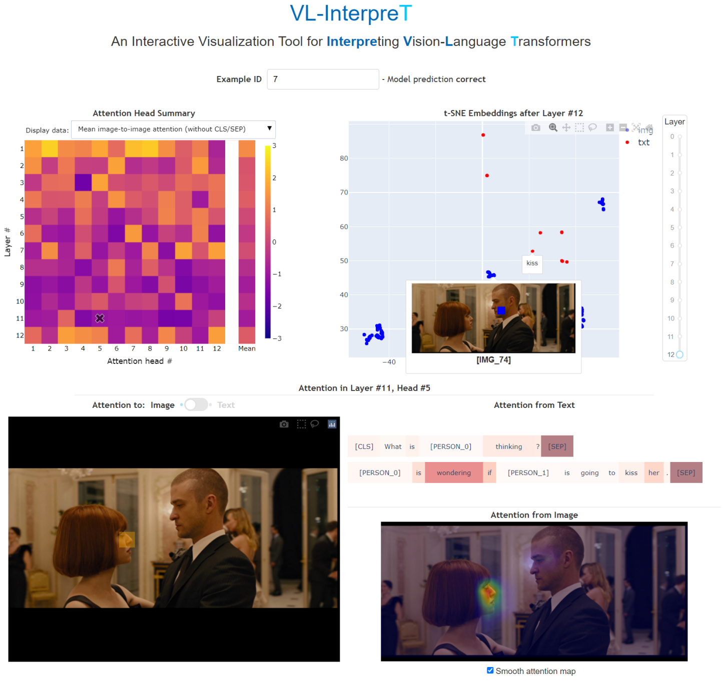


Figure 1. VL-InterpreT workflow. Data shown in this figure are for illustration purposes only.



VL-InterpreT

An Interactive Visualization Tool for Interpreting Vision-Language Transformers

Example ID 7 - Model prediction correct

Attention Head Summary

Display data: Mean image-to-image attention (without CLS/SEP)

t-SNE Embeddings after Layer #12

Attention in Layer #11, Head #5

Attention to: Image Text

Attention from Text

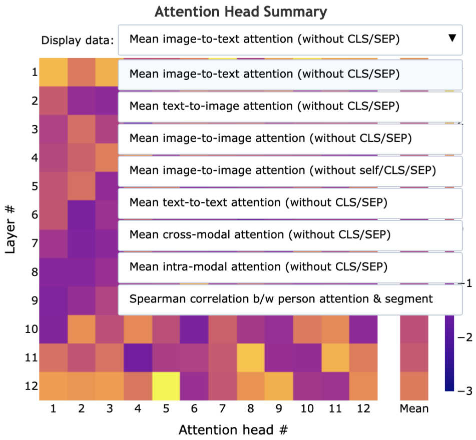
[CLS] What is [PERSON_0] thinking ? [SEP]

[PERSON_0] is wondering if [PERSON_1] is going to kiss her . [SEP]

Attention from Image

☑ Smooth attention map

Figure 2. The VL-InterpreT user interface (rearranged for print).



Attention Head Summary

Display data: Mean image-to-text attention (without CLS/SEP)

Mean image-to-text attention (without CLS/SEP)
Mean text-to-image attention (without CLS/SEP)
Mean image-to-image attention (without CLS/SEP)
Mean image-to-image attention (without self/CLS/SEP)
Mean text-to-text attention (without CLS/SEP)
Mean cross-modal attention (without CLS/SEP)
Mean intra-modal attention (without CLS/SEP)
Spearman correlation b/w person attention & segment
Layer #
1
2
3
4
5
6
7
8
9
10
11
12
Attention head #
Mean
−1
−2
−3

Figure 3. The Attention Head Summary plot colored by the metrics selected from the dropdown menu

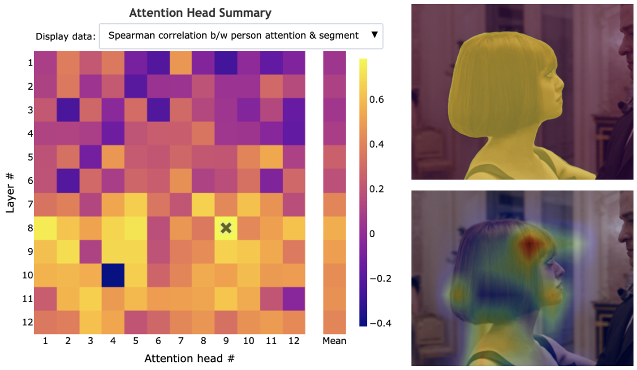


Figure 4. Custom metric: Average correlation between the V2L attention to the [*PERSON*] tokens (bottom right) and the person's panoptic segmentation mask (top right).

this calculation. A user can select a summary metric from the list of options. Each metric can be restricted to different components of the attention. For example, instead of computing the mean of all input tokens regardless of modality, the calculation can be limited to the vision-to-language part or the vision-to-vision part of the attention. Additionally, users may also focus on both cross-modal components (i.e., V2L and L2V averaged) or both intra-modal ones (i.e., V2V and V2L averaged). See the drop down menu in Figure 3.

- **Custom metrics:** Based on users' interests, custom metrics can be integrated in this plot to show relevant attention heads. For example, we created a custom metric to look for the attention heads responsible for aligning the same person between vision and language modalities, based on the V2L component. This metric, labeled *Spearman correlation b/w person attention and segment* (see Figure 3), is the Spearman correlation between the panoptic segmentation mask for a person in the image (generated by Maskformer model [2]), and the attention heatmap to the corresponding person token in the Vision-to-Language attention component.

  This metric allows a user to identify attention heads that perform a function similar to panoptic segmentation for people (see Figure 4).

Finally, for each metric, a mean is automatically computed for each layer and shown in the rightmost column of the attention head summary (see Figures 2 and 4). This column visualizes the general behavior in each layer for the selected metrics, and shows its evolution throughout the layers of the transformer.

3.2.3 Hidden state representation

For each input token, a d-dimensional hidden representation is produced by the transformer after each layer (in our setup, $d = 768$). This pool of hidden representations is then filtered by two criteria: (1) if the related text is a stop word and (2) if the related image patch comes from a part of the background (e.g., wall, ground, etc.). In order to visualize the remaining hidden representations in a readable form, t-distributed Stochastic Neighbor Embedding (t-SNE) [29] was applied to reduce dimensions and create disjoint t-SNE spaces for different layers. This way, given a selected example, VL-InterpreT tracks the hidden representations both before the first layer and after each subsequent layer, and plots them in two-dimensional spaces.

Figure 8a shows the data points representing the visual (in blue) and textual (in red) tokens from a given example. When hovering on a data point from language, the corresponding text is displayed. When hovering on a data point representing visual tokens, the image is shown with a highlighted blue patch corresponding to the visual token. Further observations on the hidden states often reveal the concept-level vision-text alignments that are learnt in this multimodal setup (see Section 4.2.3). To help further understand this alignment, VL-InterpreT allows a user to select a token (text or image patch) from a given example, and shows the nearest token in the other modality from the whole subset of examples, marked with a green star.

4. Case Studies

To demonstrate the functionalities of VL-InterpreT, we analyze an end-to-end VL transformer model, KD-VLP [23], on two benchmarks: Visual Commonsense Reasoning (VCR) [34] and WebQA [5]. Nonetheless, our tool is generally applicable to a variety of multimodal transformer configurations and types of VL datasets.

4.1. Model

The **KD-VLP model** used in our case study is a transformer for end-to-end vision-language processing. This model utilizes a ResNet backbone for visual inputs, and is pretrained using text-oriented, image-oriented, and cross-modal tasks in the form of masked language modeling, object-guided masked vision modeling, and phrase-region alignment, respectively. Depending on the application, the KD-VLP model can be fine-tuned for classification or generation tasks given bi-modal input of image and text.

4.2. Analysis on VCR

The **VCR benchmark** consists of 290K multiple choice QA problems derived from 110K movie scenes. This dataset is uniquely valuable in that it requires higher-order cognition and commonsense reasoning about the world. Given an image and a question, the objective is to select an appropriate answer from four possible choices, and then provide the rationale. To predict the correct answer and rationale, the KD-VLP model is fed with the image, the question, and each answer or rationale individually. The predicted answer a_p is the answer/rationale choice that receives the highest probability score, *i.e.*,

$$a_p = \underset{a_j}{\operatorname{argmax}} f(v, q, a_j) \tag{1}$$

where f is the KD-VLP model, v is the image, q the question, and $\{a_j | \; j \in [1, 2, 3, 4]\}$ are the possible answer choices.

The functionalities of VL-InterpreT are demonstrated using example *val-445* from the VCR validation set, or example ID 89 in the VL-InterpreT live demo. This data sample, shown in Figure 6, comprises an image showing a little girl running to a man and a woman in a garden. The question is:

Where is [*PERSON_0*] *running to?*

where [*PERSON_0*] corresponds to the little girl on the right side of the image (the locations of persons are provided in the VCR dataset and also passed to the model). We also analyze the answer predicted by the model (which in this case is correct):

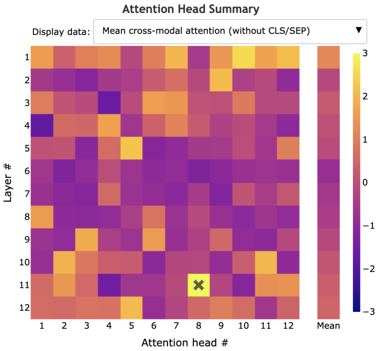
[*PERSON_0*] *is running to help* [*PERSON_1*] *and* [*PERSON_2*] *with the plants.*

where [*PERSON_1*] and [*PERSON_2*] refer to the couple.

The following analyses on this example will highlight the visualization capabilities that our application provides.

4.2.1 Attention head summary

This functionality allows a user to identify interesting heads based on various metrics. By selecting *Mean cross-modal attention* from the dropdown menu (see Figure 3), a user can identify attention heads specialized in cross-modal attention. For instance, in Figure 5 the eighth attention head in layer 11 (denoted as (11, 8)) has, on average, the highest attention across modalities. Thus, we focus on this specific head and plot its cross-modal (V2L and L2V) attention. Apart from cross-modal attention, specific heads could also been identified through *Mean image-to-text attention* for analyzing V2L attentions, or *Mean text-to-image attention* for L2V attentions. Analogous procedure also applies to L2L and V2V attention components – for instance, the metric *Mean image-to-image attention (without self)* shows the heads' attentions from every image patch to the every other image patches, excluding each patch itself and its neighbors. As described in section 3.2.2, a custom metric was used to identify attention heads with V2L components for [*PERSON*] tokens highly correlated with their panoptic segmentations. As shown in Figure 4, head (8, 9) scores particularly high in this metric. With head (8, 9) selected, the V2L attention attention heatmap for the person token (Figure 4, bottom right) indeed aligns well with the panoptic segmentation of the corresponding woman (Figure 4, top right). Furthermore, this correlation metric exhibits a trend where middle and later layers tend to have higher correlation coefficients on average (above 0.5) than early layers (less than 0.1), showing the evolution of attention patterns as the transformer layers grow deeper.


Attention Head Summary

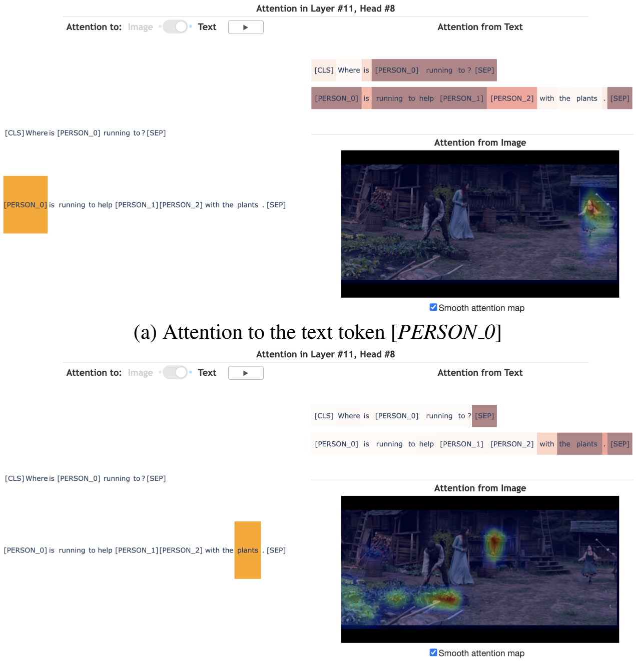
Display data: Mean cross-modal attention (without CLS/SEP) ▼

Figure 5. Attention Head Summary

4.2.2 Attention components

As described in Section 3.2.1, Figure 6 shows the heatmaps over the image and the text generated by the **Language-to-Language** and **Vision-to-Language** attention components of head (11, 8). This figure shows how attentions differ for two selected tokens: [*PERSON_0*] and *plants*. The V2L components are represented as heatmaps over the images at the bottom right of Figures 6a and 6b. It can be observed that the attention is concentrated on regions corresponding to the text, namely the little girl for the attention to [*PERSON_0*] (Figure 6a) and the plants for the attention to the *plants* token (Figure 6b). The L2L components are on the top right of these figures. For both selected text tokens, related text tokens (including themselves) are highlighted in the heatmaps. For the example in Figure 6a, the attention to [*PERSON_0*] is mostly from [*PERSON_0*], *running to*, and *running to help* [*PERSON_1*]. In the other example in Figure 6b, the attention to *plants* is mostly from the *plants* token itself.

Figure 7 visualizes the attentions to vision tokens, i.e., the **Language-to-Vision** and the **Vision-to-Vision** attentions. In this example, we select an image patch that is a part of the plants on the left image for analysis. Accordingly, the L2V component (top right of Figure 7) shows that the attention to this image patch is mainly from the [*CLS*] token as well as the *plants* token, which aligns with the concept behind the selected region. As for the V2V component (bottom right of the figure), it is also interesting that most of the regions containing plants in the image attend to this


Attention in Layer #11, Head #8
Attention to: Image Text
Attention from Text
[CLS] Where is [PERSON_0] running to ? [SEP]
[PERSON_0] is running to help [PERSON_1] [PERSON_2] with the plants . [SEP]
[CLS] Where is [PERSON_0] running to ? [SEP]
[PERSON_0] is running to help [PERSON_1] [PERSON_2] with the plants . [SEP]
Attention from Image
☑ Smooth attention map

(a) Attention to the text token [*PERSON_0*]

(b) Attention to the text token *plants*

Figure 6. Two selected text tokens and the corresponding L2L (top right) and V2L (bottom right) attention to them.

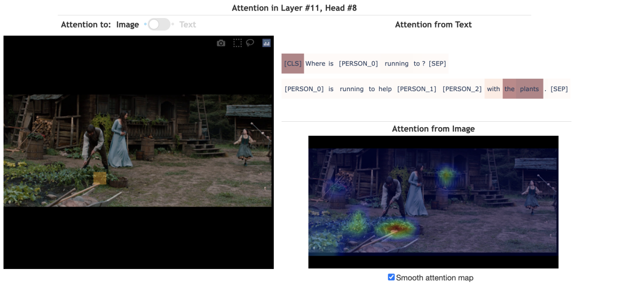
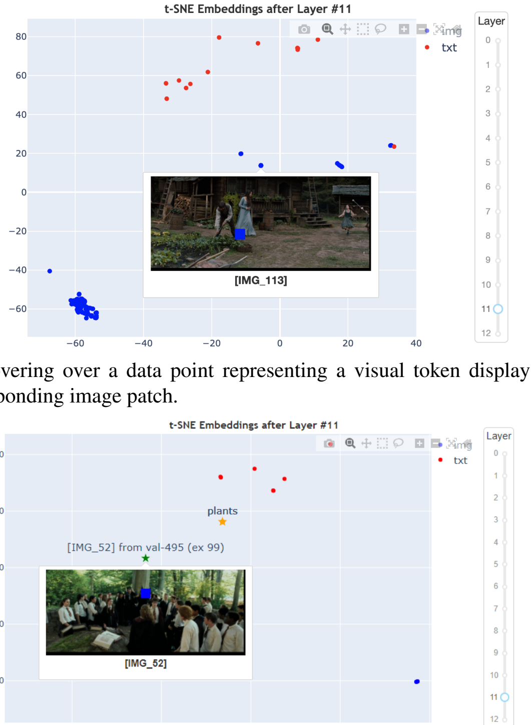

Attention in Layer #11, Head #8
Attention to: Image Text
Attention from Text
[CLS] Where is [PERSON_0] running to ? [SEP]
[PERSON_0] is running to help [PERSON_1] [PERSON_2] with the plants . [SEP]
Attention from Image
☑ Smooth attention map

Figure 7. Attentions to vision for a selected image patch in the left image. The heatmaps show the attention from the text (top right) as well as other patches in the image (bottom right).

specific patch of plants, which again shows a conceptual alignment. In summary, this example shows evidence of a unified concept of "plants", where the attention has a consistent pattern between intra- and cross-modal components.



(a) Hovering over a data point representing a visual token displays the corresponding image patch.

(b) Clicking on the text token *plants* marks it orange, and displays the closest image token from the whole dataset, marked as a green star.

Figure 8. t-SNE plot from the hidden representations of the selected example.

4.2.3 Hidden states

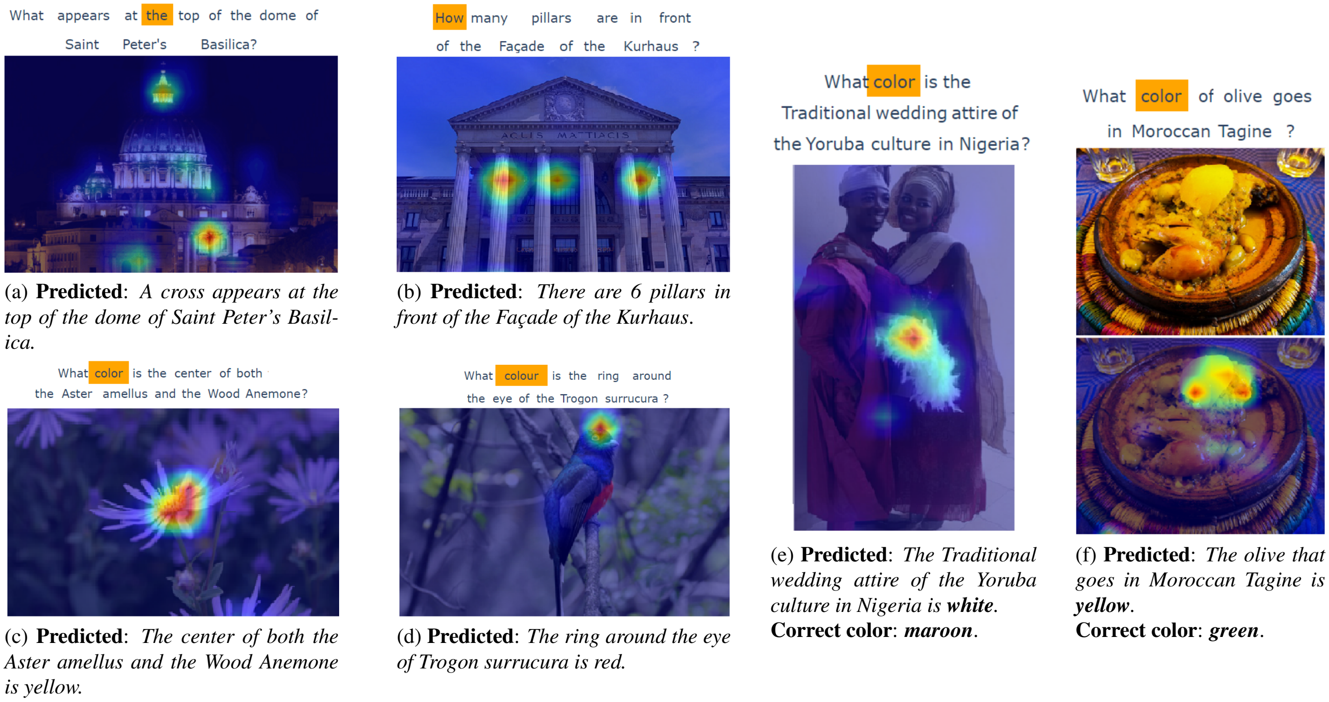
Figure 8 demonstrates the capabilities of the VL-InterpreT for visualizing hidden states. When selecting the text token *plants* (marked in orange) from the previous example *val-445* (ex 89), Figure 8b shows that in layer 11, the closest image patch from the whole pool of examples is the [*IMG_52*] (marked with a green star) from *val-495* (ex 99). It is interesting that even when it comes from a very different example, this token also refers to the plants in the image. Other than layer 11, users can select different layers on the right to see other closest image tokens to *plants* throughout layers.

Sections 4.2.2 and 4.2.3 show evidence of alignment between visual and textual concepts of *plants*. We see that the concept of "plants" in both modalities and across examples is captured by proximate representations. As such, VL-InterpreT allows studying how such sense of objectness emerges by probing the attentions and hidden states across all layers of the transformer.

4.3. Analysis on WebQA

The **WebQA benchmark** focuses on multimodal, multihop reasoning for open-domain question answering. This benchmark emulates a knowledge-seeking query to a search engine for information which may be contained in either text-based articles or images. Given a query, the goal is to identify which information is relevant across modalities, and to generate a full natural language answer based on the selected sources. The dataset contains 50k QA pairs, half of which are text-based and the other half are image-based.

We use the KD-VLP model to first select relevant sources using a classification head. Then, by adding a decoder in a Fusion-in-Decoder manner as in [33], a predicted answer is generated based on the retrieved sources. Similar behaviors can be studied for WebQA as in the previous sec-


What appears at the top of the dome of Saint Peter's Basilica?
How many pillars are in front of the Façade of the Kurhaus ?
What color is the Traditional wedding attire of the Yoruba culture in Nigeria?
What color of olive goes in Moroccan Tagine ?
What color is the center of both the Aster amellus and the Wood Anemone?
What colour is the ring around the eye of the Trogon surrucura ?

(a) **Predicted**: *A cross appears at the top of the dome of Saint Peter's Basilica.*

(b) **Predicted**: *There are 6 pillars in front of the Façade of the Kurhaus.*

(c) **Predicted**: *The center of both the Aster amellus and the Wood Anemone is yellow.*

(d) **Predicted**: *The ring around the eye of Trogon surrucura is red.*

(e) **Predicted**: *The Traditional wedding attire of the Yoruba culture in Nigeria is **white**.*
Correct color: ***maroon***.

(f) **Predicted**: *The olive that goes in Moroccan Tagine is **yellow**.*
Correct color: ***green***.

Figure 9. V2L attention heatmaps for WebQA examples and generated answers. Figures have been rearranged for print. (a)-(d) are correct predictions and (e) and (f) are incorrect.

tion. The following analyses will focus on the L2V attention in the KD-VLP encoder, as these visualizations helps in understanding how the model generates answers.

For this analysis, attention head (11, 5), identified in the same way as previous analyses, was selected. Figure 9 shows the V2L attention from image to the highlighted word above the pictures. These heatmaps show that the model attends more to the regions that help answer the question. For instance, when the question is about pillars, Figure 9b shows that the model attention comes particularly from the columns in the picture. In order to determine the color of the center of the flower in Figure 9c, the model exhibits attention from the flower center in the image and generates the correct color (yellow). Furthermore, these visualizations can also be generated for incorrectly answered questions, providing insights into the reason why incorrect answers were generated. For example, one may imagine why the model answered incorrectly in Figure 9e and 9f: In 9e, the attire that the question asks about was misidentified. That is, instead of getting attention from patches of the maroon dress, the model focuses on the white feathers. A similar behavior is also seen in 9f, where the model fails to locate the olives but focuses on the yellow vegetable in the tagine. In both cases, the model answers according to the identified object color (i.e., ”white” feathers and ”yellow” olive). These interpretations of attention helps users identify why a model fails in certain cases, and provides guidance for future efforts in improving model accuracy.

5. Conclusions and Future Directions

In this paper we presented VL-InterpreT, an interactive visualization tool for interpreting vision-language transformers. This tool allows for interactive analysis of attention and hidden representations in each layer of any VL transformer. VL-InterpreT can be used to freely explore the interactions between and within different modalities to better understand the inner mechanisms of a transformer model, and to obtain insight into why certain predictions are made. Through case studies, we demonstrated how VL-InterpreT can be used to validate the learning of cross-modal concepts, as well as to “explain” cases of failure.

In the latest version, VL-Interpret is able to run a live model to process user-generated examples in real time. This allows interactive manipulations of inputs, including both text and image, to study their effects on the attention and hidden representations.

For future work, we would like to include aggregated metrics and visualizations over multiple samples to obtain a more comprehensive understanding of model operation. In addition, we hope to experiment with any additional functionalities that will assist users in interpreting multimodal transformers, and continue to enhance this interpretability tool.

References

[1] 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. 2

[2] Alexander Kirillov Bowen Cheng, Alexander G. Schwing. Per-pixel classification is not all you need for semantic segmentation. 2021. 5

[3] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[5] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. *arXiv preprint arXiv:2109.00590*, 2021. 5

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, October 2021. 2

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 2

[8] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. In *BlackboxNLP@ACL*, 2019. 1, 2

[9] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020. 2

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 2

[11] Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2

[14] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online, Aug. 2021. Association for Computational Linguistics. 2

[15] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online, July 2020. Association for Computational Linguistics. 2

[16] Bo-Jian Hou and Zhi-Hua Zhou. Learning with interpretable structure from gated rnn. *IEEE transactions on neural networks and learning systems*, 31(7):2267–2279, 2020. 2

[17] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015. 2

[18] Vasudev Lal, Arden Ma, Estelle Aflalo, Phillip Howard, Ana Simoes, Daniel Korat, Oren Pereg, Gadi Singer, and Moshe Wasserblat. InterpreT: An interactive visualization tool for interpreting transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 135–142, Online, Apr. 2021. Association for Computational Linguistics. 2

[19] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8928–8937, 2019. 2

[20] Raymond Li, Wen Xiao, Lanjun Wang, Hyeju Jang, and Giuseppe Carenini. T3-vis: visual analytic for training and fine-tuning transformers in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 220–230, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[21] Yongjie Lin, Yi Chern Tan, and Robert Frank. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, Aug. 2019. Association for Computational Linguistics. 1

[22] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 1073–1094, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[23] Yongfei Liu, Chenfei Wu, Shao-Yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation, 2021. 5

[24] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[26] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? A probing perspective. In *AAAI 2022*, 2022. 2

[27] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2017. 2

[28] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online, Oct. 2020. Association for Computational Linguistics. 2

[29] Laurens van der Maaten. Learning a parametric embedding by preserving local structure. In *Proceeding of AISTATS 2009.*, 2009. 5

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[31] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, 2019. 2

[32] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019. 2

[33] Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. KG-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering, 2022. 7

[34] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 5

[35] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[36] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6261–6270, 2019. 2