COSE474-2024F: Final Project Report
Understanding the Role of Self-Attention on CLIP's Multimodal Performance : A Layer-Wise Analysis

Youjin Kim

1. Introduction

Motivation

Image-caption matching has become a cornerstone of multimodal learning, enabling a wide range of applications such as caption generation, visual search, and content understanding. These tasks require deep semantic alignment between visual and textual representations, a challenge that has been effectively addressed by models like CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021). CLIP achieves this by leveraging a transformer-based architecture and self-attention mechanisms to align textual and visual information in a shared embedding space. This approach has demonstrated remarkable performance across various benchmarks, cementing CLIP's role as a foundational model in multimodal learning.

Despite its success, it is unclear how much each self-attention layer contributes to the model's ability to semantically align text and image features. Self-attention, a core component of transformers, enables the model to capture contextual relationships between different input tokens. Inspired by discussions from class about the transformer architecture, I aimed to understand how attention mechanisms operate within each layer and how they affect the model's overall performance. This study investigates the specific role of self-attention by systematically modifying and disabling its layers. This project provides a deeper understanding of the significance of self-attention in CLIP and highlight the potential pitfalls of relying on structural complexity over semantic understanding.

Problem Definition

The main goal addressed in this study is twofold:
(1) to evaluate the extent to which self-attention contributes to semantic alignment between text and image features in CLIP, and (2) to analyze the failure cases and performance degradation when self-attention is disabled or replaced with simpler mechanisms. This work focuses on understanding the criticality of self-attention by examining the behavior of a modified CLIP model under various experimental conditions.

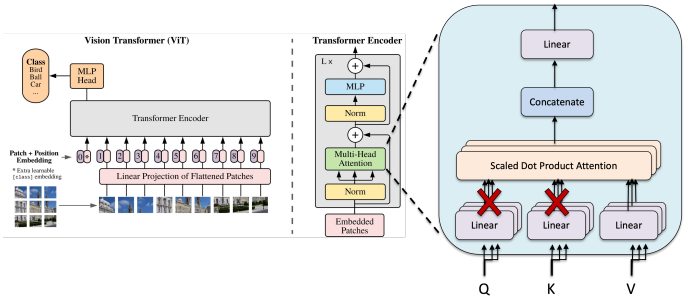

Figure 1. Canceling the self attention of each layer. A layer-wise analysis by applying this method to each different layer was conducted.

Contribution

This project presents a systematic investigation of self-attention in CLIP, with the following contributions:
1. A layer-wise analysis of self-attention, quantifying its impact on text-image matching performance.
2. Identification of failure modes where modified CLIP selects semantically irrelevant captions.
3. Novel insights into the reliance on structural text complexity when self-attention is removed.

2. Methods

This study was inspired by the perturbed-attention guidance method introduced by (Ahn et al., 2024), where certain self-attention maps in a diffusion U-Net are replaced with identity matrices. This approach generates structurally perturbed intermediate samples, leveraging the structural capturing ability of self-attention to guide the denoising process. By perturbing the self-attention, the method helps the model avoid undesired solutions and ultimately improves the diffusion model's output quality.

Building on this concept, this work investigates the roles of query and key in self-attention by modifying the vision model in CLIP. Specifically, in this study, the self-attention mechanism of the CLIP vision model was systematically altered to remove query and key computations. The self-

attention output was forced to rely solely on the value tensor, effectively canceling interactions between query and key. This perturbation was applied to various layers, enabling a layer-wise analysis of the contributions of query and key to CLIP's performance in text-image matching tasks.

In addition to modifying the self-attention mechanism within the CLIP vision model, this study employed a vision-encoder-decoder model to generate captions and further analyze the impact of these modifications. The approach integrated two key processes: caption evaluation using the modified CLIP model and caption generation using a decoder influenced by CLIP's outputs.

Figure 1 shows the overall idea of this study. Using a vit(Dosovitskiy et al., 2021)-based clip model, the self attention layers of the vision model was modified by canceling the query and key value.

Algorithm 1 Modifying Self-Attention in CLIP Layers

Input:
- Pre-trained CLIP model $clip\_model$
- Range of layers to modify: ls to le
- Original forward function $forward$

Output: Modified CLIP model with updated self-attention layers.

Step 1: Custom Forward Function

def $forward\_modified$:
$query, key, value \leftarrow Q(hidden\_states)$,
$K(hidden\_states), V(hidden\_states)$
$output\_modified = value$
Return $output\_modified$

Step 2: Modify Target Layers

for $i = ls$ **to** le **do**
$self\_attn\_layer = clip.self\_attn\_layer[i]$
$self\_attn\_layer = forward\_modified$
end for

Step 3: Return the Modified Model

In this work, I modified the self-attention mechanism in the CLIP vision model to investigate the impact of bypassing the interaction between query and key tensors. Specifically, the forward function of the self-attention layers was altered to directly use the value tensor as the attention output, effectively removing the dependency on query and key tensors. To implement this, I dynamically replaced the forward method of the self-attention submodule in these layers. This approach allows customizing specific layers without altering the original architecture or disrupting the pre-trained weights of the CLIP model. The implementation was designed to maintain modularity and ensure ease of reproducibility, enabling controlled experiments to analyze the contribution of self-attention components to the overall functionality of the model.

Model	Selected caption	Generated caption
Baseline	0.2525	0.2929
Layer 1-3	0.2218	0.2672
Layer 4-6	0.1906	0.2460
Layer 7-9	0.1906	0.2460
Layer 10-12	0.1906	0.2460

Table 1. Average CLIP Similarity Score for 349 images.

3. Experiments

3.1. Datasets

The MSCOCO 2017 val dataset consists of 5K images designed for large-scale object detection, segmentation, and captioning tasks. However, for this experiment, which focuses on a simple image-caption matching task, only natural domain images are utilized. To enable a specific performance comparison, 349 images containing only dogs and cats are selected.

3.2. Computing resource

The experiments were conducted on a macOS system with an M2 processor using Google Colab. The implementation utilized PyTorch, running on both L4 and CPU environments for evaluation.

3.3. Experimental Design

Using 'openai/clip-vit-base-patch16' model, two experiment were conducted. The first experiment focused on the change in quantitative score when the attention layers are modified. By canceling self attention of each layer by making the query and key an identity matrix, the results were analyzed by measuring the CLIP score of the caption selected by the model. Candidate captions were given like this: "a photo of a dog", "a photo of a cat", "a photo of a bird", and "a photo of a car". Testing on 349 images of dogs and cats, I focused on whether the modified model can match each image with each caption. Among the 12 self-attention layers in the vision model, the experiment was held on 5 different cases, modifying each different parts.

The second experiment aimed to identify the specific features each layer focuses on. Given an image, say a photo of a black cat sitting on a sink, candidate captions focusing on different features are given: "a photo of a cat", "a photo of a white cat", "a photo of a white cloud", "a photo of a black cloud". Then, by modifying each different part of the layers, I focused whether the modified model can identify the features of the image or not. Unlike the first experiment, this was tested on a single image with clear features.

Figure 2. Image used for experiment

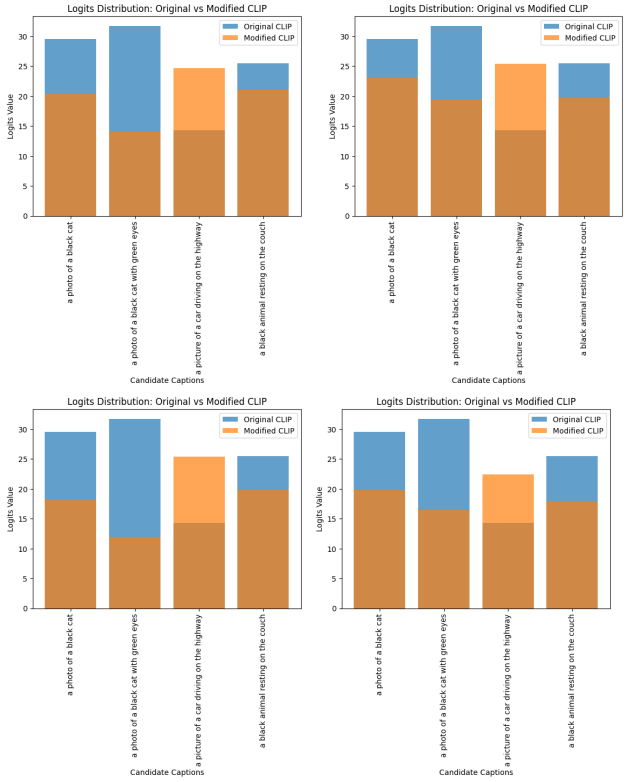

Figure 3. The height of each bar indicates the logits value output by CLIP for each candidate captions. (Top Left) Layer 1-12 modified (Top Right) Layer 1-4 modified (Bottom Left) Layer 5-8 modified (Bottom Right) Layer 9-12 modified

3.4. Quantitative results

Table 1 presents the quantitative results of the first experiment. The model with self-attention canceled in layers 1-3 is denoted as 'Layer 1-3'. The similarity scores are calculated for two types of captions: one is the selected caption among the candidate captions, and the other is a caption generated by a separate caption generation model influenced by the modified CLIP. The results indicate that canceling self-attention in lower layers has less impact on the accuracy of image-caption matching. While the accuracy for layers 1-3 decreased compared to the baseline, the overall concept of the image features was preserved. In some cases, the modified model struggled to distinguish between dogs and cats, whereas models with self-attention perturbed in higher layers often selected captions that were entirely irrelevant to the image.

However, since the candidate captions in this experiment were relatively simple—primarily distinguishing between categories such as "dog" and "cat"—an additional experiment was conducted with more specific captions. These captions included features like visual attributes (e.g., color) and behavior. Even the most irrelevant captions were augmented with additional descriptive features. The results of this experiment are shown in Figure 3. The blue bars represent the baseline. The differences in bar height highlight the inaccuracy introduced by perturbing self-attention. In all cases, models with perturbed attention exhibited an opposite trend compared to the baseline. Notably, canceling self-attention in the middle layers produced results similar to canceling all layers, emphasizing the critical role of middle layers in preserving the model's performance.

Figure 4 illustrates the logits values for each candidate caption. When the selected caption aligns closely with the ground truth, the differences in logits values are pronounced. Conversely, when the predicted caption is irrelevant, the logits values are nearly indistinguishable. This comparison highlights that CLIP with modified attention loses its ability to effectively differentiate between classes.

3.5. Qualitative results

The clarity and completeness of the generated captions reflect the qualitative results. Figure 5 displays the selected and generated captions for each layer, with the number next to each layer indicating the index of the layer where self-attention was canceled. Perturbing self-attention in the lower layers generally preserved the notion of the primary subject (e.g., a cat). However, canceling self-attention in the middle layers caused the model to predict highly irrelevant captions, significantly degrading its performance. Interestingly, perturbing self-attention in the higher layers (excluding the highest layer) had minimal impact on the results. This suggests that self-attention in the middle layers plays a critical role in maintaining the model's overall semantic understanding.

Additionally, the generated captions influenced by the selected candidate captions from Figure 3 further demonstrate the importance of self-attention when dealing with more specific and descriptive captions. The results highlight the nuanced role of self-attention in aligning textual and visual information at different levels of abstraction.

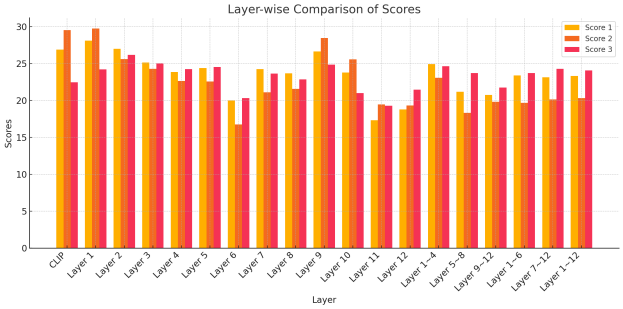

Figure 4. Layerwise Comparison of each logits score for each candidate captions. Score 1,2,3 corresponds to the score of each candidate caption. The candidate captions were given as following: "a photo of a cat", "a photo of a black cat", "a photo of a black car".

Selected and Generated Captions for Each Layer

candidate_captions = ["a photo of a cat", "a photo of a black cat", "a photo of a black car"]

Layer	*Selected Caption*	*Generated Caption*
Baseline	**a photo of a black cat**	a photo of a black cat sitting on a sink
Layer 1	**a photo of a black cat**	a photo of a black cat sitting on a sink
Layer 2	a photo of a cat	a photo of a cat sitting on a sink
Layer 3	a photo of a cat	a photo of a cat sitting on a sink
Layer 4	a photo of a black car	a photo of a black car with a cat on the floor
Layer 5	a photo of a black car	a photo of a black car with a cat on the floor
Layer 6	a photo of a black car	a photo of a black car with a cat on the floor
Layer 7	a photo of a cat	a photo of a cat sitting on a sink
Layer 8	a photo of a cat	a photo of a cat sitting on a sink
Layer 9	**a photo of a black cat**	a photo of a black cat sitting on a sink
Layer 10	**a photo of a black cat**	a photo of a black cat sitting on a sink
Layer 11	**a photo of a black cat**	a photo of a black cat sitting on a sink
Layer 12	a photo of a black car	a photo of a black car with a cat on the floor
Layer 1-4	a photo of a cat	a photo of a cat sitting on a sink
Layer 5-8	a photo of a black car	a photo of a black car with a cat on the floor
Layer 9-12	a photo of a black car	a photo of a black car with a cat on the floor
Layer 1-12	a photo of a black car	a photo of a black car with a cat on the floor

Figure 5. Selected and Generated Captions for Each Layer. The bold captions indicate the ground truth.

4. Conclusion

This study explored the role of self-attention in CLIP by systematically disabling specific layers and evaluating their impact on text-image alignment. While layers closer to the input were more critical for maintaining semantic consistency, higher layers showed limited dependence on self-attention.

While the primary intention was to identify the distinct features each self-attention layer focuses on by selectively disabling attention, practical constraints limited the depth and clarity of my findings. The relatively small subset of images (particularly those restricted to simple categories like dogs and cats) and the use of only a handful of candidate captions may have restricted the model's expressive capacity and the diversity of semantic cues available for analysis. Furthermore, the limited complexity of the provided captions reduced the opportunity to observe rich, layer-specific behaviors. For instance, distinguishing simple concepts such as "cat" versus "dog" does not fully leverage the multi-faceted nature of self-attention, which may emerge more clearly with richer textual descriptions and a broader range of image features.

5. Future Direction

To address these shortcomings, future work should consider the following improvements:

First, incorporating a more diverse and extensive image dataset spanning multiple domains, object categories, and visual attributes. More complex scenes can better highlight the subtleties in each layer's attention mechanisms. Second, using richer candidate captions and providing more nuanced, contextually rich, and feature-specific textual inputs. Captions describing detailed attributes (e.g., color patterns, poses, backgrounds) or higher-level concepts (e.g., activities, emotions) could reveal how different attention layers respond to various levels of semantic complexity. Third, perturbing the layer in an incremental manner. Rather than disabling entire layers' self-attention, applying more fine-grained perturbations at the head level or selectively masking certain tokens could yield a more granular understanding of the layer's role.

By expanding the data domain, enriching the textual prompts, and employing more sophisticated visualization and perturbation strategies, future research can overcome the current limitations and gain a more comprehensive understanding of how individual self-attention layers contribute to semantic alignment within CLIP.

References

Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H. H., Jin, K. H., and Kim, S. Self-rectifying diffusion sampling with perturbed-attention guidance, 2024. URL https://arxiv.org/abs/2403.17377.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.