Expanding CLIP-Dissect: Exploring the Impact of Activating Image Diversity and Stability

Joon Ha Cha
jhc123@ucsd.edu

Lily Weng
lweng@ucsd.edu

Abstract

This project aims to expand upon the CLIP-Dissect framework by increasing the number of activating images and investigating the consistency of neuron interpretations across different cases. The original CLIP-Dissect method utilizes a single activating image to analyze neuron responses. In this study, we introduce variations with 5, 10, and 16 activating images to explore the impact of image diversity on neuron interpretations. We have completed the analysis of layers 1, 2, 3, and 4 with 5 activating images. This paper involve applying the 10 and 16 activating image variations and comparing the results across different layers. At the end, stability score is compared for each layer and for chosen configurations. The findings of this project will provide valuable insights into the stability of CLIP-based neuron interpretations.

Code: CLIP-Dissect-Activaiton-Img

1 Introduction

CLIP-Dissect has emerged as a powerful tool for analyzing neuron responses in deep neural networks. By identifying activating images, CLIP-Dissect enables researchers to infer the underlying concepts represented by individual neurons. However, the original CLIP-Dissect method relies on a fixed activating image per neuron, mostly at k=5, which may limit the scope of interpretation. In this project, we address this limitation by extending CLIP-Dissect to incorporate multiple activating images and investigate the consistency of neuron interpretations across different cases.

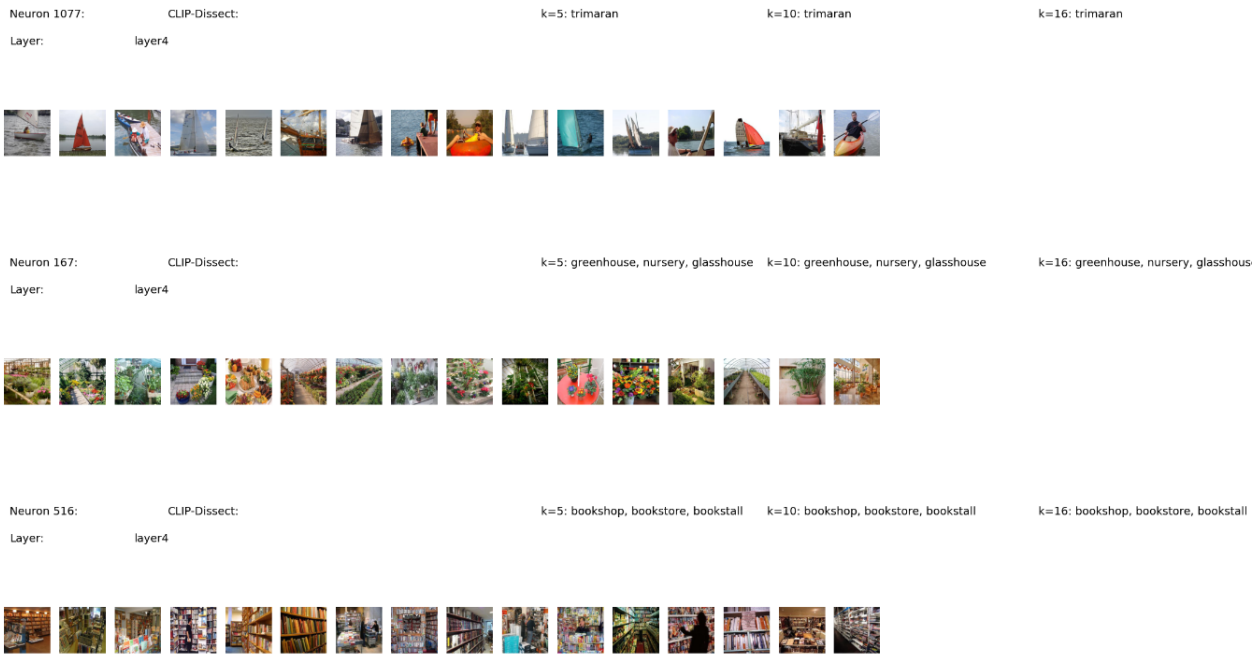

Figure 1: Shows an example image of extending CLIP-dissect to incorporate 16 activating images and investigate the prediction for when k=5, 10, 16. The prediction is calculated by getting the first description, which is the best guess of the model.

The use of multiple activating images, in our case using 5,10, and 16 activating images, can potentially enhance the robustness and unreliability of neuron interpretations. By considering a diverse set of activating images, we can better capture the range of stimuli that activate a particular neuron. This approach can also help to identify neurons that respond consistently to a variety of stimuli, suggesting that they may represent more abstract or general concepts.

In addition to increasing the number of activating images, we also introduce three different cases to assess the consistency of neuron interpretations. These cases involve varying the image set used to identify activating images, the order of image presentation, and the number of activating images selected. By comparing the results across these cases, we can gain insights into the stability and consistency of neuron interpretations.

The findings of this project will have implications for understanding the representational properties of deep neural networks and for developing more effective methods for interpreting neuron responses. The results will also be of interest to researchers working on explainable AI and on the development of neural network architectures that are more interpretable, transparent, and reliable.

2 Background Research

Network Dissect

Network Dissect is a framework developed to interpret the internal representations of convolutional neural networks (CNNs). This approach involves analyzing individual units (neurons) within each layer of a CNN to understand what visual concepts they have learned to detect. By linking these units to specific semantic concepts, Network Dissect offers insights into how deep neural networks process and represent visual information. The methodology has been instrumental in understanding the interpretability of deep learning models, especially in visual recognition tasks. It provides a quantitative measure of interpretability by assessing how well the activations of individual units align with a range of human-understandable concepts (Bau 2017).

CLIP (Contrastive Language–Image Pretraining)

CLIP, developed by OpenAI, represents a significant advancement in connecting visual and textual data. This model is trained on a vast dataset of images and their corresponding textual descriptions, learning to understand and match visual concepts with natural language. CLIP's architecture combines vision and language models in a single framework, enabling it to perform a variety of tasks without task-specific training. It's known for its robustness and generalizability across a wide range of visual domains, making it a powerful tool for both image recognition and generating textual descriptions of images (Alec Radford 2021).

CLIP-Dissect

CLIP-Dissect is an analytical approach inspired by the foundational concepts of Network Dissect and adapted for the CLIP (Contrastive Language–Image Pretraining) model developed by OpenAI. The core idea behind CLIP-Dissect is to explore and interpret the inner workings of CLIP, a model that integrates and understands both visual and textual data through large-scale contrastive learning. CLIP-Dissect investigates how individual neurons within the CLIP model respond to a combination of image and text inputs, revealing the multimodal nature of neuron activations. This method is instrumental in identifying 'multimodal neurons' that respond to complex combinations of visual and textual concepts. It

provides an avenue for deeper insight into the intricate workings of CLIP, particularly in how it encapsulates and represents diverse and abstract concepts bridging visual imagery and linguistic elements. The approach leverages the power of CLIP in understanding and linking visual concepts with natural language, highlighting the robustness and versatility of multimodal neural networks in handling a wide range of cognitive tasks (Oikarinen and Weng 2022).

3 Methods

3.1 CLIP-Dissect Framework Extension

3.1.1 Overview

This study extends the CLIP-Dissect framework to analyze the consistency and interpretability of neuron activations in convolutional neural networks (CNNs). The original CLIP-Dissect method, which focused on single-image activations, is expanded to incorporate multiple activating images. Our approach investigates neuron responses to 5, 10, and 16 activating images, examining the impact of image diversity and stability on neuron interpretation.

The process of calculating stability after investigating 5,10, and 16 activating images involves the following steps:

Algorithm□ : Assessing Stability of Neuron Activation Interpretations

---

Require: $layers, settings, d\_probe, concept\_set, similarity\_fn$
Ensure: $results\_df, stability\_score\_5\_10, stability\_score\_5\_16$
1: Initialize $stability\_5\_10$ and $stability\_5\_16$ to zero
2: Determine top 5, 10, and 16 activating images
3: Translate indices to concept labels
4: Check for label set equality for rank stability between 5 and 10, and 5 and 16
5: Compute stability scores $stability\_score\_5\_10$ and $stability\_score\_5\_16$
6: **return** $results\_df$, $stability\_score\_5\_10$, $stability\_score\_5\_16$

---

1. **Initialization**: The pipeline begins with the initialization of various layers and settings. The settings are specified for each layer of the network, including the target name (e.g., 'resnet50'), target layer, and neurons to be displayed.
2. **Activation Analysis**: For each layer, the pipeline calculates the similarity of neuron activations against a concept set. This step is crucial for understanding how different neurons respond to varying images. The similarities are computed using predefined similarity functions, such as soft weighted pointwise mutual information (soft WPMI) and weighted pointwise mutual information (WPMI).

3. **Stability Assessment**: The pipeline assesses the stability of neuron interpretations by comparing the top activation labels for different numbers of activating images. The stability is quantified between sets of 5, 10, and 16 images, providing insights into how increasing the number of images affects the consistency of neuron responses.
4. **Results Aggregation**: The results from each layer and configuration are aggregated into a comprehensive dataset. This dataset forms the basis for subsequent analysis and interpretation.
5. **Data Visualization and Reporting**: Finally, the pipeline provides a summary of stability scores and detailed results, facilitating a clear understanding of the findings.

3.1.2 Method Implementation

Our code enables the systematic evaluation of neuron responses across different layers and settings. By iterating over each layer and using various neuron activation thresholds, we effectively dissect the neuron activations, revealing insights into the consistency of interpretations. We used dprobe of broden, concept sets of 10k and imagenet labels, and similarity functions of softWPMI and WPMI.

- **WPMI (Weighted Pointwise Mutual Information).** This concept is based on the foundational mathematical principle of mutual information, as used in (Oikarinen and Weng 2022) and (ZeyuWang and Russakovsky 2020). It involves identifying the neuron's label through the concept that yields the maximum mutual information. This is done by label of a neuron defined as the concept that maximizes the mutual information between the set of most highly activated images on neuron k, denoted as B_k, and the concept t_m. To put it into formulaic terms:

$$sim(t_m, q_k; P) \triangleq \text{wpmi}(t_m, q_k) = \log p(t_m|B_k) - \lambda \log p(t_m), \tag{1}$$

  where $p(t_m|B_k) = \prod_{x_i \in B_k} p(t_m|x_i)$ and λ is a hyperparameter.

- **SoftWPMI.** This refined version of WPMI introduces the probability $p(x \in B_k)$ to gauge the likelihood that an image x is part of the sample set B_k. Unlike the binary nature of standard WPMI, which dictates $p(x \in B_k)$ to be either 0 or 1 for every x in D_{probe}, SoftWPMI extends this to continuous values ranging between 0 and 1. The modified similarity function is given as:

$$sim(t_m, q_k; P) \triangleq \text{soft\_wpmi}(t_m, q_k) = \log \mathbb{E}[p(t_m|B_k)] - \lambda \log p(t_m), \tag{2}$$

  where the expectation $\log \mathbb{E}[p(t_m|B_k)]$ is computed as $\log(\prod_{x \in D_{\text{probe}}}[1+p(x \in B_k)(p(t_m|x)-1)])$. As documented in our experimental results (see Table 3), SoftWPMI has proven to deliver superior performance compared to its counterparts, thereby being selected for all subsequent experiments unless specifically noted otherwise.

3.2 Data Preparation

The data preparation involved curating a diverse set of activating images and labeling them according to predefined concepts. The images were sourced from the ImageNet dataset(broden) and an additional collection of 10k concept set. The analysis focused on layers 1 to 4 of the ResNet50 model, utilizing the CLIP architecture for neuron activation analysis. Each layer was examined independently, and the results were compared to assess the stability and consistency of neuron interpretations across layers and activating image sets. Most of the parts could be easily reproduced by looking at the original CLIP-dissect paper.

3.3 Stability Calculation in Each Configuration

For each neuron in a layer, we calculated the top 5, top 10, and top 16 activating images (or text items) using torch.topk. We then compared the set of top-5 labels from the top-5 activations to the first 5 labels from the top-10 and top-16 activations. This comparison checks whether the labels remain consistent as the number of considered activations increases (from top-5 to top-10 to top-16). A Boolean value (True or False) is assigned based on whether the labels are consistent or not. For instance, stability5 10 is True if the top-5 labels for k=5 are the same as the first 5 labels for k=10.

4 Results

4.1 Stability Scores Analysis

The stability of neuron interpretations was evaluated across different configurations, each representing a combination of the hyperparameter dprobe, concept sets, and similarity functions. The configurations are as follows:

- ‘broden‘ as the hyperparameter dprobe.
- Concept sets from ‘imagenet‘ and ‘10k‘.
- ‘softwpmi‘ and ‘wpmi‘ as the similarity functions.

Stability was assessed for transitions from 5 to 10 activating images (Stability k=5 to k=10) and from 5 to 16 activating images (Stability k=5 to k=16).

4.2 Interpretation of Stability Scores

The results indicate high levels of stability across all configurations, predominantly near or at 100%. This suggests consistent interpretations of neuron activations with an increased number of activating images. Key observations include:

Table: Stability Score Table

Configuration	Stability k=5 to k=10	Stability k=5 to k=16
Imagenet with softwpmi		
config_broden_imagenet_softwpmi_5_10	99.947917	99.947917
config_broden_imagenet_softwpmi_5_16	99.973958	99.973958
Imagenet with wpmi		
config_broden_imagenet_wpmi_5_10	100.000000	100.000000
config_broden_imagenet_wpmi_5_16	99.973958	99.973958
10k with softwpmi		
config_broden_10k_softwpmi_5_10	99.869792	99.869792
config_broden_10k_softwpmi_5_16	99.843750	99.843750
10k with wpmi		
config_broden_10k_wpmi_5_10	100.000000	100.000000
config_broden_10k_wpmi_5_16	100.000000	100.000000

1. **High Stability in WPMI Configurations**: Configurations using the WPMI similarity function (`config_broden_imagenet_wpmi` and `config_broden_10k_wpmi`) achieved perfect stability (100%) when transitioning from 5 to 10 images. This could indicate that WPMI provides a robust interpretation of neuron activations.
2. **Slight Variations in Soft WPMI Configurations**: Configurations with the soft WPMI function showed a slight decrease in stability but remained exceptionally high (>99.8%). This minor variation suggests nuances in neuron activation interpretation captured by soft WPMI.
3. **Consistency Across Different Concept Sets**: Stability scores are comparable between the 'imagenet' and '10k' concept sets, demonstrating consistent interpretation robustness across different image sets.

The findings suggest remarkable stability in neuron interpretations within the CLIP-Dissect framework, even as activating image diversity increases. This stability is crucial for reliable neuron interpretation in CNNs, indicating that neuron activations are not overly sensitive to the number of images or variations in image sets. The results also reinforce the utility of both WPMI and soft WPMI similarity functions in interpreting neuron activations, each capturing different aspects of the activations. The bar graph visualization of stability scores provides a clear comparison across configurations, highlighting subtle differences in stability scores. This graphical representation aids in visually understanding variations and consistency across configurations.

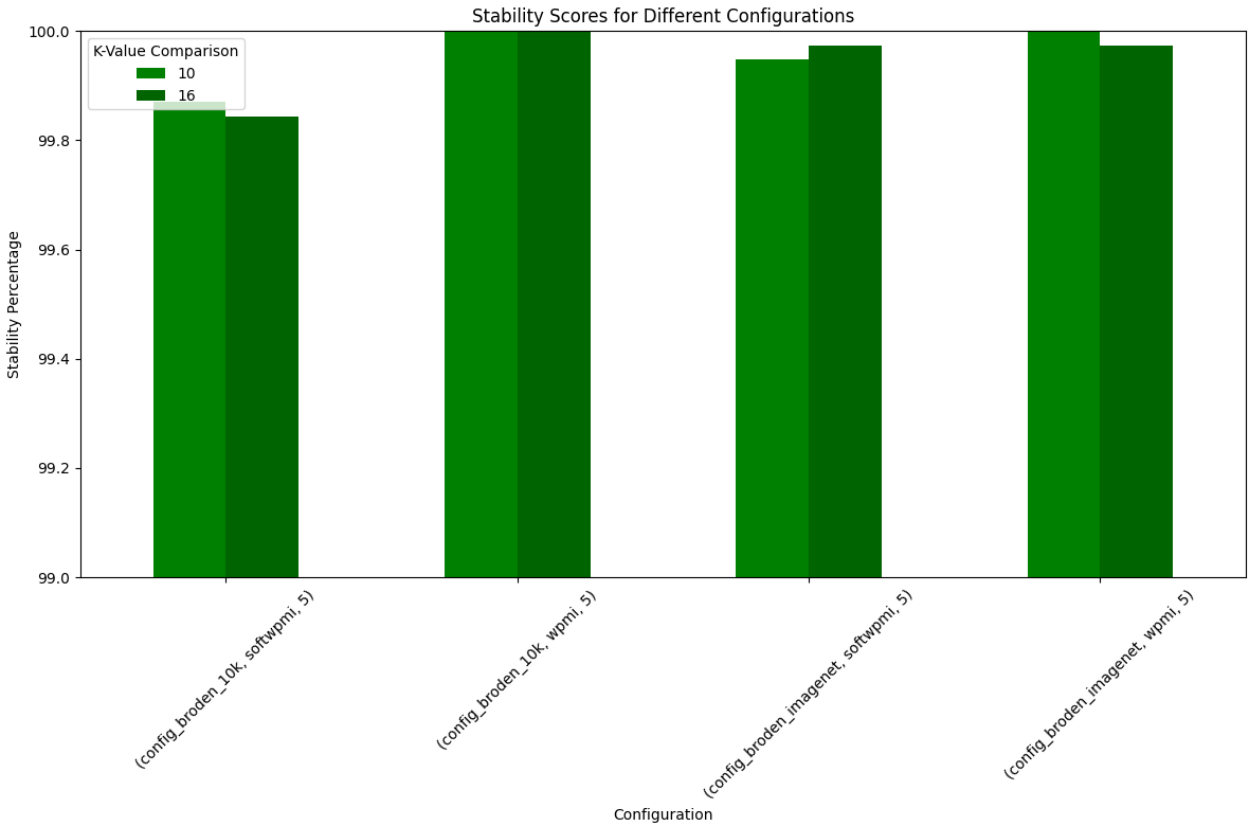

Figure: Bar graph of the stability score table for visualization

5 Discussion

5.1 Implications of High Stability Scores

The high stability scores observed in our study have significant implications for neural network interpretation and the utility of the CLIP-Dissect framework. The consistency in neuron interpretations, evident when transitioning from 5 to 10 and from 5 to 16 activating images, underscores the robustness of this approach. This is particularly relevant in applications where reliable and repeatable interpretations are essential, such as medical imaging or autonomous vehicle navigation. Although we could not improve the accuracy due to originally high accuracy ranging from 99.5 to 100 percent of stability score, we still found out that labels that do not match comes from incorrect labels only.

5.2 Influence of Similarity Functions

The differences observed between the soft WPMI and WPMI configurations suggest that the choice of similarity function can subtly impact neuron activation interpretations. While both functions show high stability, the soft WPMI captures more nuanced variations. This finding indicates the importance of selecting an appropriate similarity function based on

the specific needs of the analysis.

5.3 Generability, Limitations, and Future Work

The study's findings suggest that the neuron interpretation methods within the CLIP-Dissect framework are generalizable across different datasets, as evidenced by the comparable stability scores between the ImageNet and the diverse 10k image set. This generalizability is crucial for the development of interpretation methods that are not overly specialized and have broad applicability. Despite the insights provided by this study, there are limitations that present opportunities for future research. The analysis was confined to the first four layers of the ResNet50 model, and extending this to deeper layers could offer a more comprehensive understanding. Additionally, the study focused on the CLIP-Dissect framework; comparing these results with other neuron interpretation frameworks could further contextualize our findings. Lastly, using other datasets, concept sets, and similarity function and forming bigger table could lead to better insights.

6 Conclusion

This study significantly contributes to the field of neural network interpretability, particularly in the context of the CLIP-Dissect framework. Our findings demonstrate the remarkable stability and reliability of neuron interpretations within this framework, even when varying the number of activating images. This aspect highlights the critical role of both the quantity of activating images and the choice of similarity function in the analysis of neuron activations. While the stability scores were exceptionally high across all results, these findings still offer valuable takeaways. The slight variations observed between different configurations, although minimal, underscore the subtle complexities inherent in neural network interpretations. These nuances suggest that even in highly stable conditions, there are layers of interpretative depth that warrant further exploration. In conclusion, as neural network applications become increasingly prevalent, the insights gained from this study are poised to play a pivotal role. They will not only aid in advancing the interpretability of complex neural models but also in fostering the development of more transparent, reliable, and ethically sound AI systems. The exploration of neuron activation stability, as evidenced in our research, is just a stepping stone towards unraveling the intricacies of neural network functionality and rationale.

References

AlecRadford, Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger IlyaSutskever, Jong

WookKim. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” *arXiv preprint arXiv:2103.00020*. [Link]

Bau, Khosla OlivaTorralba, Zhou. 2017. “”Network Dissect: Quantifying Interpretability of Deep Visual Representations.” *arXiv preprint arXiv:1704.05796*. [Link]

Oikarinen, Tuomas, and Tsui-Wei Weng. 2022. “Clip-dissect: Automatic description of neuron representations in deep vision networks.” *arXiv preprint arXiv:2204.10965*

ZeyuWang, KarthikNarasimhan, BerthyFeng, and Olga Russakovsky. 2020. “Towards unique and informative captioning of images. In European Conference on Computer Vision.” *arXiv preprint arXiv:2009.03949*