CLIP-Dissect Automatic Evaluation

Hou Wan
hwan@ucsd.edu

Mentor: Lily Weng
lweng@ucsd.edu

Abstract

Network dissecting is a method used to understand the inner workings of deep neural networks by examining the functionality of individual neurons. It typically involves techniques to automatically associate neurons with specific concepts or features, providing insights into how the network processes information. CLIP-Dissect, a novel method, stands out for its capability to automatically label internal neurons with concepts, leveraging multimodal vision/language models such as CLIP. The output of CLIP-Dissect entails the labeling of individual neurons with associated concepts, elucidating the network's functioning. In the CLIP Dissect paper, quantitative evaluation for neuron labels has been done in terms of final layer neuron evaluations due to having access to the neuron ground truths. However, for hidden neurons, the inaccessible ground truths make quantitative evaluation harder. Qualitative observations have been done, including looking at small subsets of hidden neuron labels and their activating images as evaluation, however this is a manual process that requires careful observation. In this study we introduce a way to replicate this style of evaluation automatically. To do this, we utilize 3 different approaches of observing activating images and their labels. We find in this study that utilizing pretrained and benchmarked models to do this task is an effective method to conduct qualitative assessment of neuron dissections. The methods used are prompting a VQA (BLIP-2), and using OpenCLIP embedding similarities as means of evaluation. Notably, BLIP-2 demonstrated a high alignment with human evaluations, achieving an Intersection over Union (IoU) score of 0.809, surpassing the OpenCLIP methods which recorded IoUs of 0.768 and 0.776.

Website: houkinwan.github.io/CLIPdissectAutoEval
Code: github.com/houkinwan/DSC180B

1 Introduction

Neural networks are black boxes, and not interpretable. A method that is a proposed solution is network dissection to get a better understanding (David Bau).

Network dissecting is a method used to understand the inner workings of deep neural networks by examining the functionality of individual neurons. It typically involves techniques to automatically associate neurons with specific concepts or features, providing insights into how the network processes information. CLIP-Dissect, a novel method, stands out for its capability to automatically label internal neurons with concepts, leveraging multimodal vision/language models such as CLIP. Its input comprises three key components: the target DNN for dissection, a probing dataset D probe devoid of concept labels but representative of real-world data, and a set of concepts reflecting potential features or attributes the neurons might capture. The output of CLIP-Dissect entails the labeling of individual neurons with associated concepts, elucidating the network's functioning.

In the CLIP Dissect paper, quantitative evaluation for neuron labels has been done in terms of final layer neuron evaluations due to having access to the neuron ground truths. However, for hidden neurons, the inaccessible ground truths make such evaluation harder. Qualitative observations have been done, including looking at small subsets of hidden neuron labels and their activating images as evaluation, as well as using crowdsourcing such as amazon mechanical turk. This is however a manual process that requires careful observation.

Current methods for evaluation of hidden labels consist of manually observing hidden layer top activating images and the neuron label associated with them. In this study we introduce a way to replicate this style of evaluation automatically. To do this, we introduce 3 different deterministic approaches of observing activating images and their labels.

2 Background and Related Work

Network Dissection: Network dissection **?** is a comprehensive framework devised to assess the interpretability of latent representations within Convolutional Neural Networks (CNNs) by evaluating the alignment between individual hidden units and a predefined set of semantic concepts. This method leverages a diverse dataset of visual concepts to assign semantic labels to hidden units across various layers of a CNN. Prior approaches rely on labeled dataset consisting of images with associated pre-determined concepts.

CLIP-Dissect: CLIP-Dissect **?** is a recent method for understanding the roles of hidden layer neurons by leveraging the CLIP **?** multimodal model (Radford et al., 2021). It can provide a score of how close any neuron is to representing any given concept without the need of concept annotation data. This is highly advantageous for scaling to more interpretability due to not requiring prelabeled datasets.

VQA: Visual Question Answering models are a multimodal model that

BLIP-2: BLIP-2 **?** is a lightweight VQA that utilizes an efficient pretraining strategy for vision-language tasks, aiming to mitigate the computational burden associated with end-to-end training of large-scale models. It leverages off-the-shelf frozen pre-trained image encoders and large language models to bootstrap vision-language pretraining. Central to BLIP-2 is a lightweight Querying Transformer (Q-Former), pretrained in two stages: first, to facilitate vision-language representation learning from the frozen image encoder, and second, to enable vision-to-language generative learning from the frozen language model. By employing this approach, BLIP-2 achieves state-of-the-art performance on various vision-language tasks, including visual question answering, image captioning, and image-text retrieval, while requiring significantly fewer trainable parameters compared to existing methods.

OpenCLIP: OpenCLIP **?** is an open-source iteration of CLIP (Contrastive Language-Image Pretraining). It utilizes the CLIP's ideas, which harnesses natural language supervision to pre-train on a massive dataset of image-text pairs collected from the internet. OpenCLIP does what CLIP does, training image-text pairs by training an image encoder EI and a text encoder ET in parallel.

3 Methodology

3.1 Evaluation Criteria

In order to consistently evaluate the neurons, we introduce a basic criteria to identify whether a CLIP-dissect evaluation should be labeled as a correct evaluation or not. To do this, we look at the top k activating images for a singular neuron, and decide the CLIP-dissect evaluation as correct if any one of the images relate to the label according to the vision models.

3.2 BLIP-2 Prompting

The first approach utilizes a Visual Question Answering (VQA) model, which is done by crafting a suitable prompt that would allow the model to evaluate whether the top k matching images are suited to the neuron label. For this, we utilize BLIP-2, feeding it the top 5 one at a time activating images, prompting the model with a Yes or No question. This method utilizes CLIP encoding for the model, but a VQA like BLIP-2 is qualitatively shown to be able to provide more nuanced answers.

3.3 OpenCLIP

We opt to use a model trained by open sourced means to ensure that there is no redundancies by using the exact same vision transformer (CLIP) as CLIP-Dissect. For the use of OpenCLIP, we propose two different evaluation methods. The first one involves using OpenCLIP to compute image-text pair similarities for the activating image and text from

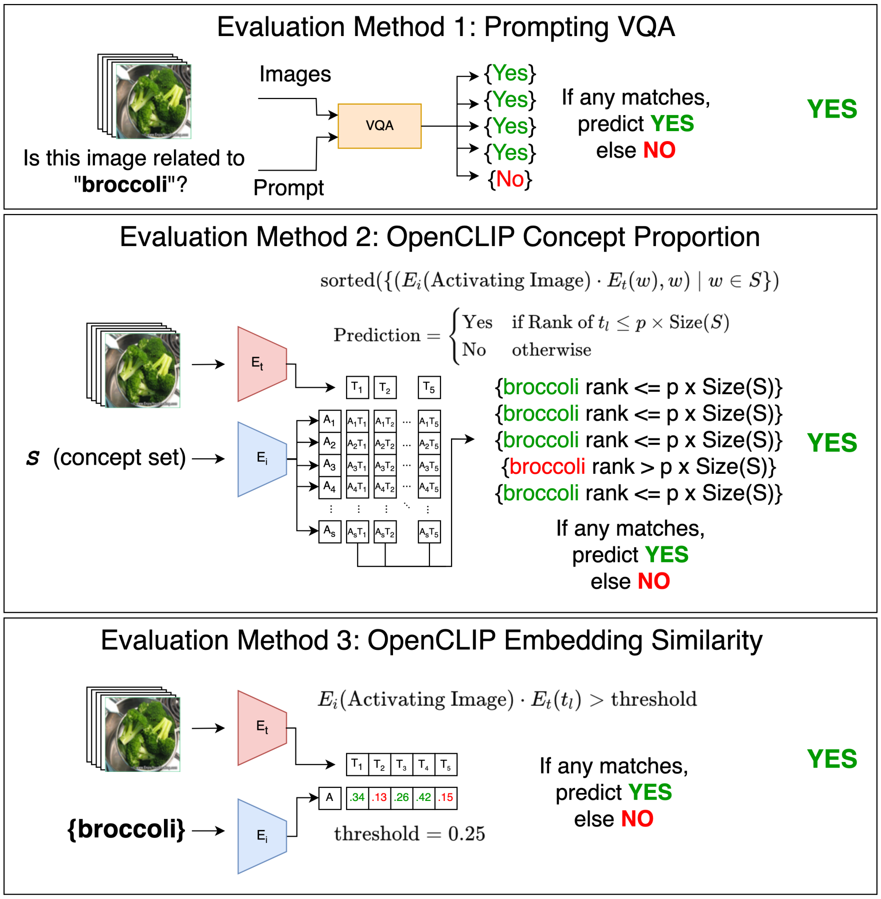
Evaluation Method 1: Prompting VQA
Images
Is this image related to "broccoli"?
Prompt
VQA
{Yes}
{Yes}
{Yes}
{Yes}
{No}
If any matches, predict YES else NO
YES
Evaluation Method 2: OpenCLIP Concept Proportion
sorted({(E_i(Activating Image) · E_t(w), w) | w ∈ S})
Prediction = Yes if Rank of t_l ≤ p × Size(S); No otherwise
E_t
E_i
S (concept set)
T1 T2 T5
A1 A2 A3 A4 As
A1T1 A1T2 A1T5
{broccoli rank <= p x Size(S)}
{broccoli rank <= p x Size(S)}
{broccoli rank <= p x Size(S)}
{broccoli rank > p x Size(S)}
{broccoli rank <= p x Size(S)}
If any matches, predict YES else NO
YES
Evaluation Method 3: OpenCLIP Embedding Similarity
E_i(Activating Image) · E_t(t_l) > threshold
E_t
E_i
{broccoli}
T1 T2 T3 T4 T5
A
.34 .13 .26 .42 .15
threshold = 0.25
If any matches, predict YES else NO
YES

Figure 1: This is a figure caption.

the CLIP-Dissect text corpus. We then look at the highest similarities and decide that an activating image is a match if the neuron label is within a chosen top proportion of the OpenCLIP similarities by rank, that is the CLIP-Dissect label t_l is within the top proportion of the size of the text corpus S. This is done for top k activating images for each neuron label with a chosen proportion p of 0.01.

$$\text{sorted}(\{(\mathbf{E}_i(\text{Activating Image}) \cdot \mathbf{E}_t(w), w) \mid w \in S\}) \tag{1}$$

$$\text{Prediction} = \begin{cases} \text{Yes} & \text{if Rank of } t_l \leq p \times \text{Size}(S) \\ \text{No} & \text{otherwise} \end{cases} \tag{2}$$

The next evaluation method uses OpenCLIP to again encode the activating image, but instead of the text corpus, we just encode a singular neuron label, then compute the image-text pair cosine distance in embedding space. We threshold this value and decide that an activating image is a match if the similarity exceeds a chosen threshold with a chosen threshold of 0.25.

$$\mathbf{E}_i(\text{Activating Image}) \cdot \mathbf{E}_t(t_l) > \text{threshold} \tag{3}$$

3.4 Compatibility

The design of our evaluation framework is inherently adaptable, enabling seamless integration with the most advanced vision-language models as they emerge. This flexibility is crucial, as it allows researchers to continuously update and refine their interpretability analyses with the best tools available, ensuring that the evaluation process remains at the cutting edge of technological progress. The use of BLIP-2 in our study, for instance, exemplifies this approach, showcasing how the framework can incorporate state-of-the-art models to benefit from the latest developments in the field.

This principle of adaptability extends to all components of the framework, including OpenCLIP and similar multimodal models. Thanks to its modular design, the framework can easily accommodate new models as they become available, without necessitating substantial modifications to the evaluation process. This ensures that the framework stays relevant and effective, providing researchers with the most up-to-date tools for a deeper understanding of image and language representations in neural networks. This is similar to CLIP-Dissect's adaptibility as well.

4 Experiments

4.1 Qualitative Results

In the qualitative analysis of the neuron labels produced by CLIP-Dissect and evaluated by our methods, we observed distinct patterns that highlight the strengths and weaknesses of

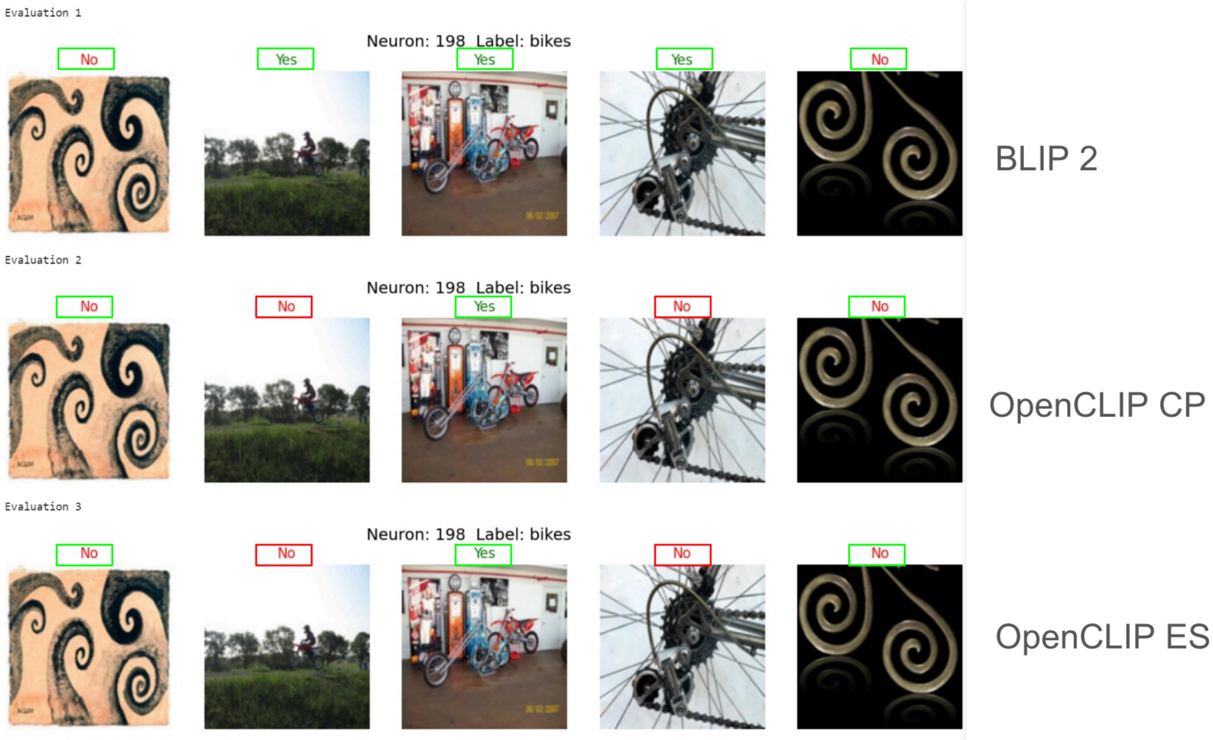

Figure 2: This is a figure caption.

each approach. For instance, when employing the BLIP-2 model, we noticed that the responses often exhibited a high level of nuance and context understanding. BLIP-2's ability to interpret complex visual scenes and relate them to the neuron labels allowed for a more human-like assessment of the images. This led to a higher agreement with human evaluations, as BLIP-2 could often capture the subtleties that a human evaluator might notice.

However, this nuanced understanding also meant that BLIP-2 sometimes produced forgiving assessments, where even if a neuron label was somewhat general or not an exact match for the activating images, the model might still consider it a reasonable label. This behavior reflects a more holistic view of image understanding but may sometimes overlook the finer details that would lead to a stricter evaluation.

On the other hand, the OpenCLIP approaches, both the concept proportion and embedding similarity methods, tended to offer a more binary perspective. The evaluation was more cut and dry, relying heavily on the direct similarity between the image embeddings and the text embeddings of the neuron labels. This often resulted in a less forgiving assessment compared to BLIP-2. For example, if a neuron label was ”dog” and the activating images were of various dog breeds, the OpenCLIP methods might not recognize the label as fitting unless the similarity metrics were particularly high, even if a human evaluator would consider ”dog” a reasonable label for those images.

4.2 Quantitative Results

To evaluate the effectiveness of our automatic neuron labeling methods, we compared the automatically generated results with those obtained through human evaluation. This comparison was conducted by examining the alignment between the top activating image for each neuron and its corresponding CLIP-Dissect label, across three automated methods. The degree of alignment was quantified using the Intersection over Union (IoU) between the automatic evaluations and human annotations, expressed as the ratio of matching elements to the total number of elements, denoted by $\frac{|A \cap H|}{|A \cup H|}$, where A represents the set of automatically labeled neurons and H denotes the set of human-annotated neurons.

Our dataset for this experiment consisted of 3840 neurons from a ResNet50 model, using the Broden D_{probe} and a set of 10,000 concepts ($S = 10k$), with each neuron undergoing five image-text evaluations. The comparison revealed that the BLIP-2 method showed a higher similarity to human-evaluated results when compared to the OpenCLIP methodologies, indicating its superior performance in accurately labeling neurons based on the predefined criteria. These results are seen in Table 1. BLIP-2 has a high IoU of 0.809, with OpenCLIP Concept Proportion (CP) and OpenCLIP Embedding Similarity (ES) having IoUs of 0.768 and 0.776 respectively.

To test the viability of this approach, we experimented with different CLIP Dissect configurations, anticipating varying levels of performance degradation based on the dataset (D_p) and concept set ($\mathscr{S}$) combinations. Contrary to expectations, the ImageNet validation set, when used with a 20k concept set, demonstrated a notable deviation from this trend, showcasing higher evaluation metrics than some Broden dataset configurations. This observation suggests that the richness and diversity of ImageNet may enhance the alignment between neuron activations and the corresponding labels, potentially due to the broader and more varied semantic content inherent in the ImageNet dataset.

When employing the Broden dataset with categories places as the concept set, the performance was notably lower across all methods. In Table 2 with BLIP-2 scoring 0.0904, and OpenCLIP embedding similarity showing a slightly better performance at 0.2510, while OpenCLIP concept proportion did not manage to align any labels successfully. However, as we change to diverse and large datsets like 10k and 20k, the performance of all methods improved significantly, with BLIP-2 consistently outperforming both OpenCLIP configurations. Interestingly, the combination of ImageNet validation and Broden datasets with a 20k concept set yielded the highest performance metrics for OpenCLIP methods, highlighting the potential of combining diverse datasets for improved neuron label alignment.

Table 1: IoU of automatic evaluation methods with human based evaluation on Broden 10k (1 denotes perfectly aligned and 0 denotes no alignment)

Evaluator	Layer 1	Layer 2	Layer 3	Layer 4	Overall
BLIP-2	0.8633	0.8047	0.7822	0.8169	0.809
OpenCLIP CP	0.8105	0.7568	0.7568	0.7520	0.768
OpenCLIP ES	0.8828	0.7910	0.7666	0.7632	0.776

4.3 Choosing Thresholds

Both the cosine similarity threshold and the top proportion criterion are pivotal in our methodological framework, serving as adjustable parameters that influence our results' granularity and specificity. The selection of a particular cosine similarity threshold, alongside the choice of a specific top percentile from the word corpus, dictates the strictness of our matching criteria. However, our qualitative findings are not dependent on the particular parameter choices: These parameters are interdependent in their impact on the analysis. Altering one while holding the other constant may shift the absolute outcomes but preserves the integrity of comparative evaluations across different CLIP dissect configurations. Essentially, the relative ordering of results remains consistent regardless of the threshold/proportion settings.

Despite this, to better align our automated evaluations with human standards, we're considering adjustments to our threshold settings to more closely match human judgment. By carefully examining how changes in these settings affect our system's agreement with human evaluations, we aim to achieve a more human-like performance in our automated processes. This effort could lead to improvements in OpenCLIP's evaluation methods. This can be seen in Figure 3. In future work, we might fine-tune our criteria using a specific subset of data, aiming to improve our system's precision to resemble human or Visual Question Answering model evaluations more closely. Adopting such tailored threshold adjustments could make OpenCLIP evaluations more interpretable and useful across various applications.

4.4 Efficiency

In the context of our study, the efficiency of the neuron evaluation methods is crucial. We observe these results in Table 3 BLIP-2, despite its longer processing time of 8 hours, offers a significant advantage in terms of scalability and consistency. This could potentially be scaled in resources to reduce runtimes. Furthermore, the automated nature of BLIP-2's evaluation negates the variability and potential biases inherent in human assessments, ensuring a more uniform and repeatable analysis process.

Conversely, OpenCLIP demonstrates exceptional efficiency, with its evaluation methods

Table 2: The accuracies based on qualitative choice of CLIP Dissect arguments (places performs worse than broden)

D_{probe}	Concept set $\mathscr{S}$	BLIP-2	OpenCLIP CP	OpenCLIP ES
Broden	places	0.0904	0.0000	0.2510
Broden	10k	0.8557	0.6263	0.6581
Broden	20k	0.8534	0.6828	0.6992
ImageNet val	20k	-	0.6268	0.6719
ImageNet val + Broden	20k	0.8480	0.6867	0.7026

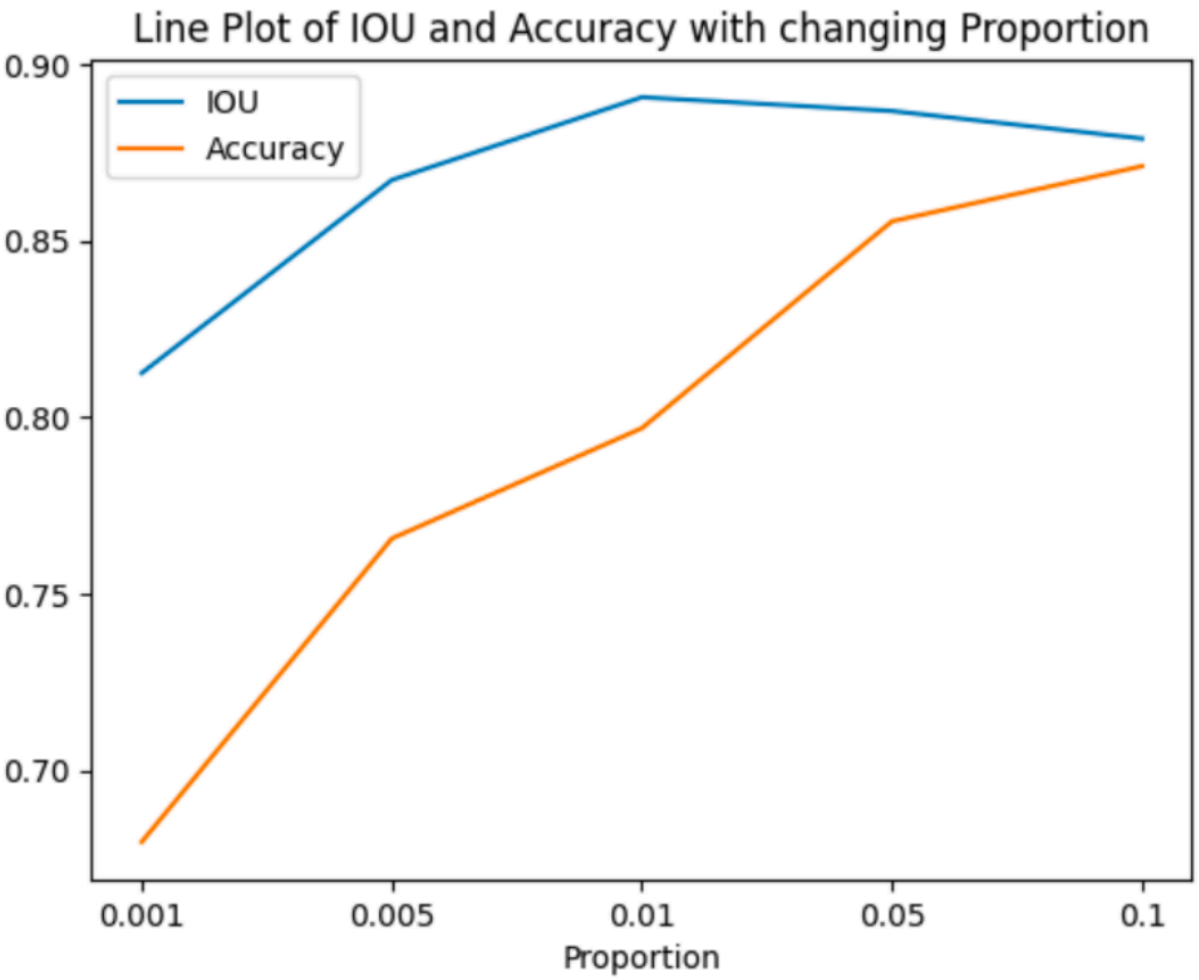

Figure 3: Open CLIP proportion accuracies and Human Set IoU with changing proportion for $D_{probe} = broden$ and $S = 10k$ across ResNet50 layers

completing tasks in just 4.5 and 2 minutes for concept proportion and embedding similarity respectively. This rapid assessment capability highlights the strength of OpenCLIP, especially for applications requiring quick insights or when operating under time constraints. The trade-off, however, lies in the balance between the depth of analysis provided by BLIP-2 and the swiftness of OpenCLIP's evaluations.

4.5 Comparisons

The kappa statistic (κ) serves as a quantifiable measure of inter-rater agreement, considering the likelihood of agreement occurring by chance. The formula is shown by $\kappa = \frac{P_o - P_e}{1 - P_e}$, where P_o represents the relative observed agreement among raters, and P_e denotes the hypothetical probability of chance agreement. The κ value ranges from -1, indicative of complete disagreement, to 1, signifying complete agreement, with a value of 0 suggesting an agreement that is no better than chance. The interpretation of κ values is typically categorized as follows: values ≤ 0 imply no agreement, 0.01–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement.

In our investigation, the kappa statistic demonstrated a moderate concordance between human assessments and the BLIP-2 methodology, with a κ value of 0.4714 as seen in Table 4, implying that BLIP-2 reasonably corresponds with human judgments albeit with certain discrepancies. Furthermore, substantial agreement was observed between the two OpenCLIP methodologies—concept proportion and embedding similarity—with a κ of 0.8358. However, their concordance with BLIP-2 was comparatively lower, with κ values ranging from 0.3829 to 0.3953, indicating divergent evaluation strategies.

This disparity highlights the inherent complexities of aligning automated neuron evaluation methods with human evaluations, especially when dealing with binary data represented by 1s and 0s for neuron activations. Despite the good internal consistency observed in BLIP-2 and OpenCLIP methods, the evident gap with human evaluative standards underscores the necessity for refinement in these automated methodologies. Such enhancements aim to more closely replicate human evaluative patterns, potentially improving the interpretability of neural networks within binary data frameworks.

Table 3: Comparing automatic evaluation methods with human based evaluation on Broden 10k

Evaluator	Time Taken
Human	4 hours
BLIP-2	8 hours
OpenCLIP CP	4.5 minutes
OpenCLIP ES	2 minutes

Table 4: Kappa agreement values among the evaluators with $D_{probe} = broden$ and $S = 10k$. The matrix is symmetric, so values below the diagonal are omitted for clarity.

Evaluator \ Evaluator	Human	BLIP 2	OpenCLIP CP	OpenCLIP ES
Human	-	0.4714	0.4863	0.4884
BLIP 2	-	-	0.3829	0.3953
OpenCLIP CP	-	-	-	0.8358
OpenCLIP ES	-	-	-	-

5 Use Case

These evaluation method's first primary use would be to evaluate whether CLIP-Dissect's predictions on neuron labels are reasonable. It aims to replace the manual overhead of having to evaluate the entire network personally.

However, this may also potentially be extended to have a quantitative but visual outlook on whether neurons are interpretable. The original CLIP-Dissect paper introduced a novel approach to evaluate the interpretability by their similarity score, as defined by the function sim$(t, q_k; P)$, exceeded a certain threshold τ. This threshold was determined empirically, aiming to ensure that 'interpretable' neurons had an average description score significantly higher than the general average, setting a cutoff at $\tau = 0.16$ with the SoftWPMI similarity function.

Building upon this foundation, we can effectively perform the same evaluation, for each individual neuron based on OpenCLIP embedding similarity. Given CLIP-Dissect is a perfect labelling method, the binary yes or no can also indicate whether a label is able to be associated with a picture, and this inability of association by definition makes a neuron uninterpretable.

6 Limitations and Conclusion

Limitations: Evaluating is inherently hard, due to also requiring human labels. A method of evaluating this in a label-free setting will always require human annotation.Once again, it suffers from the original problem that it is hard to get human annotated at scale without having to rely on methods like crowdsourcing. Also, our evaluation framework relies heavily on the performance of external vision models, such as BLIP-2 and OpenCLIP. This dependency means that the effectiveness of our evaluation is contingent upon these models' abilities to accurately process and understand the probing datasets and associated text concepts. This could potentially introduce biases or inaccuracies if these models have limitations or perform poorly on specific types of data. For example, the need for pre-processing or augmenting datasets to suit the evaluation models (as seen with the places and Broden datasets) can add another layer of complexity and potential error.

Conclusion: In conclusion, our proposed method serves as an effective tool for performing sanity checks on the CLIP-Dissect process, ensuring its functionality aligns with intended outcomes. It introduces a structured framework that leverages the power of activating images to evaluate the labeling of neurons, enriching our understanding of neural network interpretability. The innovative aspect of our approach lies in its use of vision models distinct from those employed in CLIP-Dissect, broadening the scope of evaluation and providing a more versatile analysis.

While the ultimate evaluation of interpretability inevitably draws upon human judgment, our method offers a quantifiable, automated metric that reflects the effectiveness of CLIP-Dissect's parameters. This is evidenced by the correlation between the evaluators' improved accuracies and the optimization of CLIP-Dissect configurations. Therefore, our framework not only aids in verifying the operational integrity of CLIP-Dissect but also contributes to refining its parameters for enhanced performance, marking a significant advancement in the automated assessment of neural network interpretability.

7 Future Work

For future work, addressing the challenge of interpretability in the context of complex datasets like ImageNet is crucial. Our initial attempts to test against the ground truth of ImageNet labels highlighted a significant issue: the specificity of ImageNet classes can make them difficult for human evaluators to interpret meaningfully. This specificity poses a challenge for evaluating the interpretability of neural networks, as the direct correlation between neuron activations and such detailed classes may not readily translate into intuitive, human-understandable concepts.

To overcome this challenge, one promising direction is to leverage a dataset that focuses on more broadly interpretable concepts, such as basic shapes, colors, and simple, universally recognizable objects or animals. Training a neural network, potentially on a ResNet50 backbone, with this type of dataset could facilitate a more intuitive understanding of what each neuron is capturing, making the process of interpretability more aligned with human cognition.