Profiling Vision Neural Networks with CLIP-DISSECT - DSC Capstone Quarter 1 Project Report

Yongce Li
yol013@ucsd.edu

Lily Weng
lweng@ucsd.edu

Abstract

In this report, we propose a layer-wise network profiling method. Traditional approaches often focus on neuron-wise interpretations, which, while valuable, may not provide a comprehensive understanding of the network's functionality as a whole. Our project emphasizes a layer-wise analytical framework that offers a macro-examination of network architectures. With the strong capabilities of CLIP-DISSECT and GPT-4, our method enables the automatic profiling of networks on a layer-by-layer basis. This profiling allows researchers to conduct comparative analyses across different DNNs and to track changes in network profiles throughout the training or fine-tuning phases.

Code: https://github.com/YongceLi/Profiling-Vision-Neural-Networks-with-CLIP-DISSECT

1 Introduction

Deep Neural Networks (DNNs) have succeeded across various domains such as computer vision, natural language processing, and multimodal tasking. Yet, as DNNs become increasingly capable, the need for interpretability intensifies to develop more robust and explainable models. One way to understanding DNNs is to inspect the functionalities of individual neurons. Manual inspection was first presented to describe neuron functionality by categorizing the set of highly activated images of each neuron. To automate the human inspection process, Network Dissection (Bau et al. 2017) proposed an auxiliary densely-labeled dataset named Broden. By selecting the concept with the highest IoU score between the activated region and the semantically segmented region, Network Dissect was able to automatically detect neuron functionalities. Despite its efficiency, the creation of intricately labeled datasets like Broden requires a lot of human labor. To get rid of such limitations, CLIP-DISSECT (Oikarinen and Weng 2023) leverages recent advances in multimodal vision/language models to label internal neurons with open-ended concepts without the need for any labeled data or human examples. Instead of calculating semantic overlaps, CLIP-DISSECT calculates similarity scores using internal CLIP representation and the activation vectors. It summarizes highly activated images by selecting the text concept with the highest similarity score. While these methods have advanced our understanding of neuron-level concepts, researchers still do not have a comprehensive, overarching view of the network's profile. In this project, we leverage the strong capabilities of CLIP-DISSECT and GPT-4 to automatically classifies the detected neuron functionalities into its super-category, enabling a thorough layer-wise network profile. With such profiles, different interesting directions of research are proposed. One line of research is to compare different networks' profiles to analyze the correlation between the profile distribution and the models' performance. The other one is to monitor the evolution of network's profile during its training or finetuning process.

2 Methods

In this section, we illustrate our main method to get neural networks' profile. Our methodology involves multiple models:

- The subject model is the vision model that we are attempting to profile.
- The explainer model generates neuron explanations for MLP neurons in subject models. In our project, the explainer model is set to be the automated neuron concepts detector CLIP-DISSECT.
- The summarizer model summarizes neurons of subject models into categories based on the detected neuron explanations. The summarizer model should be a large language model with strong summarization capabilities. In our project, the summarizer model is set to be OpenAI's GPT-4 model.

Our methodology contains two main parts. The first one is to label each neuron with a concept in MLP layers we want to analyze. In this step, we apply CLIP-DISSECT (Oikarinen

and Weng 2023) to dissect the subject network. The second one is to summarize the detected neuron concepts into different categories for each layer, creating a layer distribution for users to monitor.

2.1 Neuron Explanations with CLIP-DISSECT

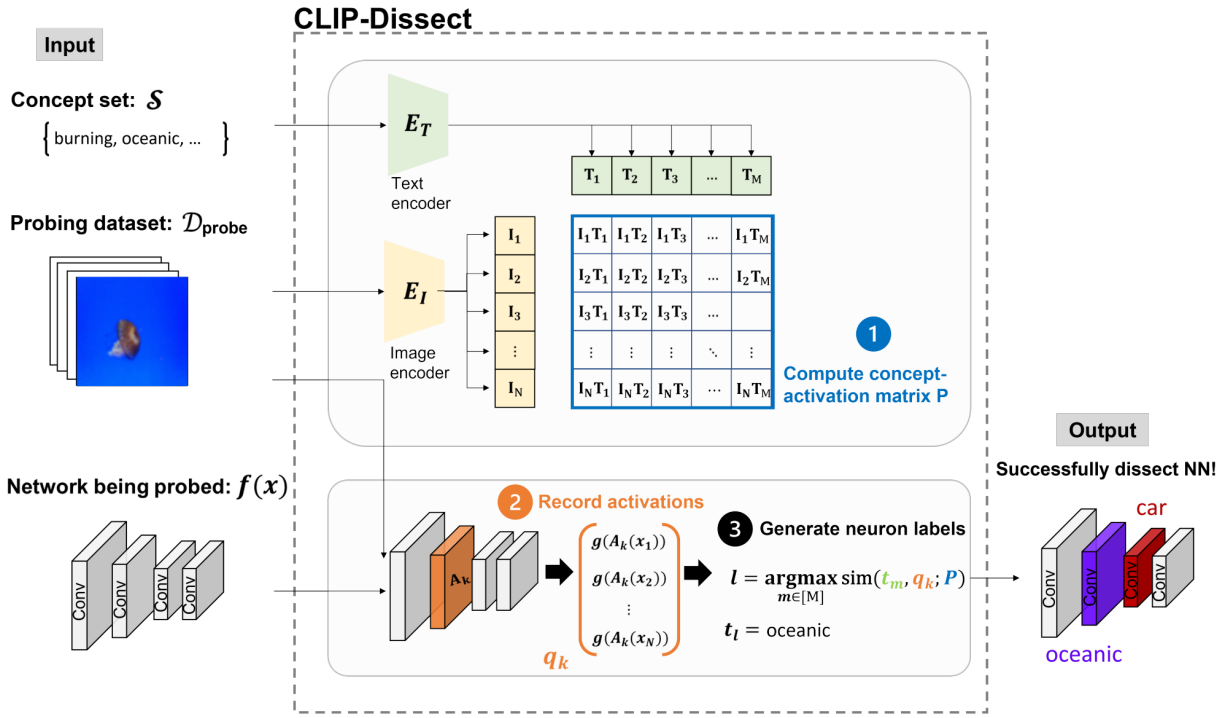

Figure 1: Overview of CLIP-Dissect: a 3-step algorithm to dissect neural network of interest

In this section, we reviewed the main algorithm of CLIP-DISSECT. CLIP-DISSECT takes advantage of advanced vision-language multimodal model CLIP to connect potential text concepts with the highly-activated probing image sets, allowing efficient automatic neuron explanations for vision models.

CLIP-DISSECT contains the following inputs:

1. a deep vision neuron network to be probed and dissected, denoted as $f(x)$
2. a set of probing images, denoted as $\mathscr{D}_{\text{probe}}$ where $|\mathscr{D}_{\text{probe}}| = N$
3. a set of concepts, denoted as $\mathscr{S}$, $|\mathscr{S}| = M$

The output of CLIP-Dissect is the neuron labels, which identify the concept associated with each individual neuron.

It consists of the following three key steps:

1. *Compute the concept-activation matrix P.* Using the image encoder E_I and text encoder E_T of a CLIP model, we first compute the text embedding T_i of the concepts t_i in the concept set $\mathscr{S}$ and the image embedding I_i of the probing images x_i in the

probing dataset $\mathscr{D}_{\text{probe}}$ Next, we calculate the concept-activation matrix $P \in \mathbb{R}^{N \times M}$ whose (i, j)-th element is the inner product $I_i \cdot T_j$, i.e. $P_{i,j} = I_i \cdot T_j$.

2. *Record activations of target neurons.* Given a neuron unit k, compute the activation $A_k(x_i)$ of the k-th neuron for every image $x_i \in \mathscr{D}_{\text{probe}}$. Define a summary function g, which takes the activation map $A_k(x_i)$ as input and returns a real number. Here we let g be the mean function that computes the mean of the activation map over spatial dimensions, but g can be any general scalar function. We record $g(A_k(x_i))$, for all i, k.
3. *Determine the neuron labels.* Given a neuron unit k, the concept label for k is determined by calculating the most similar concept t_m with respect to its activation vector $q_k = [g(A_k(x_1)), \ldots, g(A_k(x_N))]^\top, q_k \in \mathbb{R}^N$. The similarity function sim is defined as $\text{sim}(t_m, q_k; P)$. In other words, the label of neuron k is t_l, where $l = \arg\max_m \text{sim}(t_m, q_k; P)$. In this project, we choose similarity function to be Soft Weighted Pointwise Mutual Information (WPMI) as defined below.

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

where we compute $\log \mathbb{E}[p(t_m \mid B_k)] = \log\left(\Pi_{x \in \mathscr{D}_{\text{probe}}} [1 + p(x \in B_k)(p(t_m \mid x) - 1)]\right)$.

2.2 Neuron Categorizations with GPT-4

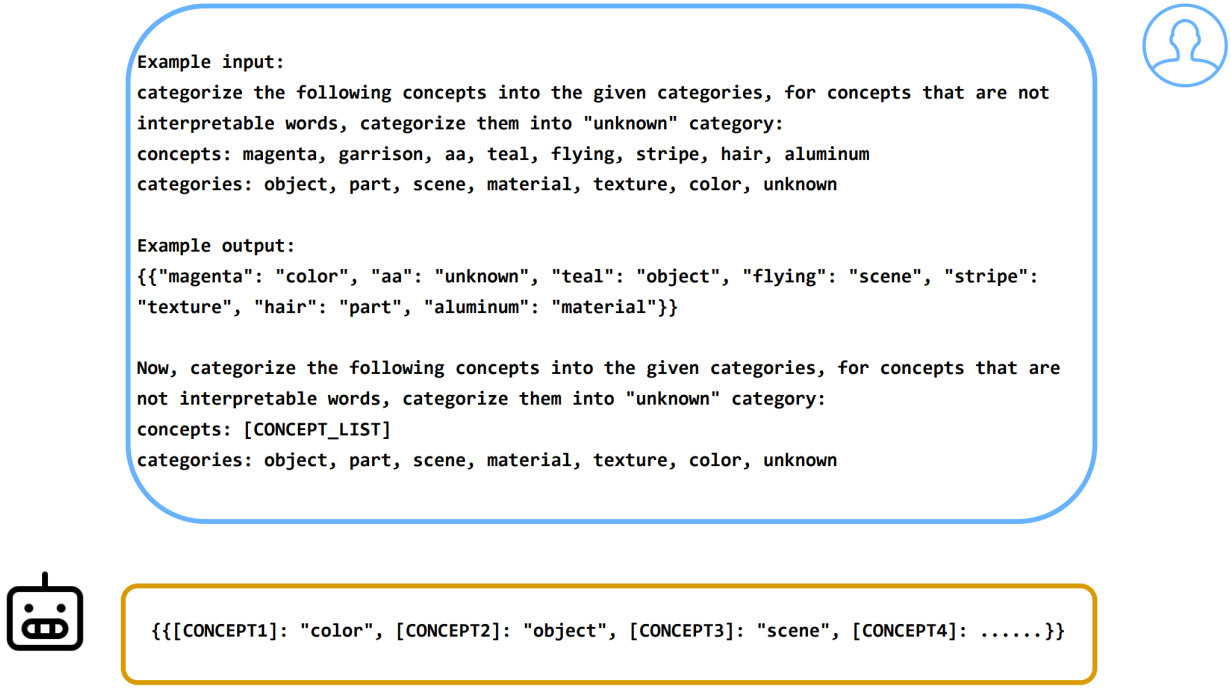

Figure 2: Overview of One-shot Prompting for GPT-4 Summarization

In this section, we propose the method to use large language model to summarize detected neuron concepts into different categories, thus generating neuron distributions for each layer of the subject model.

In this step, we aim to build a Concept-to-Category mapping to help us count the frequency of each concept in each layer. Due to the strong capabilities provided by GPT-4, we were able to provide a one-shot example to make GPT-4 generate such mappings from our input concept lists. Specific one-shot prompting example is shown above in Figure 2. After getting the mapping, we were able to draw a neuron distribution for each layer and compare the distributions across different layers to see how neuron concepts change throughout the whole model, eventually providing a concept profiling for the neural network. In our experiments, we set 7 categories: ”object”, ”part”, ”scene”, ”material”, ”texture”, ”color”, ”unknown” to align with concept categories presented in Network Dissection. However, users can input different super-categories for their specific research purposes.

3 Results

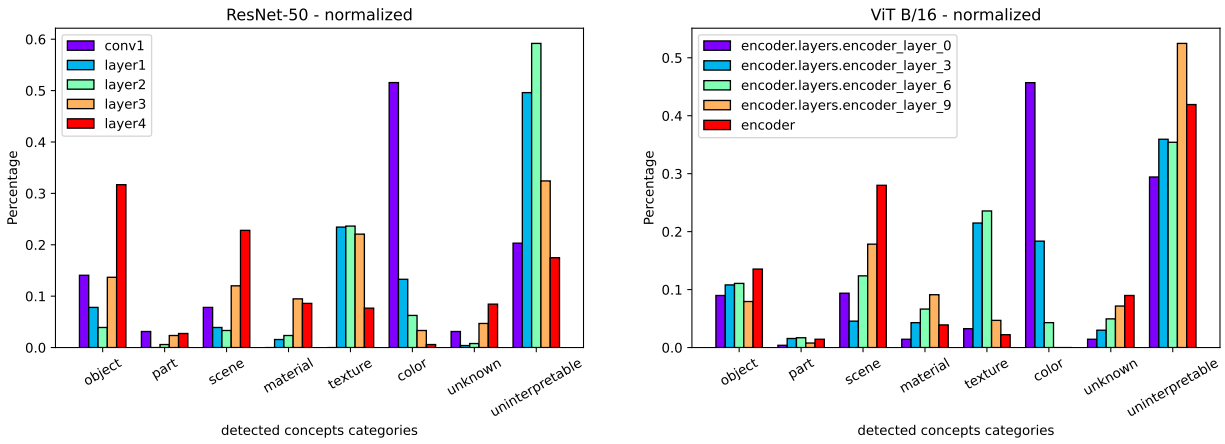

Figure 3: Example network profiling for ResNet-50 and ViT B/16 models, detected interpretable concepts are categorized into seven super-categories: object, part, scene, material, texture, color, unknown

In this project, we tested our network profiling techniques on two vision models with different scales: ResNet-50 and ViT B/16. For ResNet-50, we created profiling on all 5 MLP layers, and for ViT B/16, we created profiling for layer 0, 3, 6, 9, and the last MLP layer for the ease of qualitative comparison. Resulting Network Profiles are shown in Figure 3. The y-axis represents the percentage of neurons that belong to certain categories, and the x-axis represents different concept categories.

3.1 Definition of Interpretability

In the above plots, we have one additional category - uninterpretable. In our project, interpretability of a certain neuron is defined similarly as in CLIP-DISSECT paper.

After we apply CLIP-DISSECT on the subject model, each description for a neuron is associated with a similarity score, and the higher the similarity the more accurate the description. we consider a neuron k with description t as interpretable if $SoftWPMI(t, q_k; P) > \tau$.

To choose the best cutoff τ , we leverage human evaluation to give a description score on the results of CLIP-DISSECT. Specifically, we evaluated the description quality of 100 randomly selected neurons for each of the 5 layers and 2 models studied, for a total of 1000 evaluations. Human evaluator was presented with 10 most highly activating images, and answered the question: *"Does the description: [DESCRIPTION] match this set of images?"* Each evaluation had three options which we used with the following guidelines:

- Yes (score 1) - Most of the 10 images are well described by this description
- Maybe (score 0.5) - Around half (i.e. 3-6) of the images are well described, or most images are described relatively well (accurate but too generic, or slightly inaccurate)
- No - (score 0) Most images are poorly described by this caption

In particular, for ResNet-50, we choose the lowest τ such that interpretable neurons will have an average human description score of 0.75 or higher (compared to 0.655 of all neurons), and for ViT B/16, we choose the lowest τ such that interpretable neurons will have an average human description score of 0.75 or higher (compared to 0.51 of all neurons). This gives us a cutoff threshold $\tau = 0.16$ with the proposed SoftWPMI similarity function. In contrast, the neurons with $SoftWPMI \leq \tau$ have an average description score of 0.5257 for ResNet-50, and an average description score of 0.35 for ViT B/16. Using this cutoff τ, we find 77.8% of neurons in ResNet-50 and 61% of neurons in ViT B/16 to be interpretable, indicating that around 20-40% of neurons do not have a simple explanation for their functionality, i.e. are 'uninterpretable'. Detailed interpretability information is shown in Figure 4. On average, we see a high portion of neurons to be uninterpretable for ViT B/16 model than ResNet-50, indicating that neurons of larger-scaled models would be harder to label by CLIP-DISSECT.

3.2 Network Profile for ResNet-50 and ViT B/16

We have the following observations for both ResNet-50 and ViT B/16 from Figure 3.

1. intermediate layers (layer 1, 2, 3 in ResNet-50 and layer 3, 6, 9 in ViT B/16) have a higher proportion of uninterpretable neurons, indicating they are harder to interpret, which conforms with the results in CLIP-DISSECT.
2. Neurons in shallower layers tend to have more abstract concepts than neurons in deeper layers. E.g. first several layers have more neurons in category "color", "texture", while deeper layers have more concepts in "object", "scene", etc.

model	Interpretability threshold for similarity score	Average score above threshold (human eval)	Average score below threshold (human eval)	Average score overall	% of Interpretable neurons
ViT B/16	0.16	0.75	0.35	0.51	37.8%
ViT B/16	0.12	0.65	0.27		61.0%
ResNet-18 (Places-365)	0.16	0.75	0.53	0.655	69.7%
ResNet-50 (ImageNet)	0.16				77.8%

Figure 4: Interpretability for the two models given different τ value

4 Discussion

In this section, we analyze Pros and Cons of our method, and introduce potential future works based on our method.

1. *Pros* Our method, though based on the neuron-wise explainers, has a very different function than neuron explanations. Traditional neuron explainers only give us information about single neurons, but not integrate the information to give researchers a wholistic view of network functions. Our Network Profiles provide a more macro view of the network's function, allowing researchers to monitor how network behaves across different layers. With GPT-4, the whole process is automatic without any human effort involved. Additionally, our network profiling method is very efficient, taking less than 10 minutes to dissect large vision model such as ViT B/16.
2. *Cons* Our method based heavily on the performance of CLIP-DISSECT and GPT-4, so the network profile may not be useful when CLIP-DISSECT perform bad on a certain model. As in Figure 4, many neurons are labeled as uninterpretable, especially for large-scaled model. When more capable neuron explainer comes out, we may replace CLIP-DISSECT with them to improve the profiling effectiveness and accuracy.
3. *Future works* With our network profiling method, one line of research is to compare different networks' profiles to analyze the correlation between the profile distribution and the models' performance, which helps researchers to understand why different models perform differently on various tasks. The other one is to monitor the evolution of network's profile during its training or finetuning process, which would help us understand how models are trained to perform better and potentially create more effective training or finetuning strategies.

5 Conclusion

In this project, we propose a layer-wise network profiling method. Based on the state-of-the-art CLIP-DISSECT model and GPT-4, our method first label every single neurons in the subject vision network, and then make use of the strong categorization capabilities of GPT-4 to produce concept distribution for every single MLP layer in the subject model. Our network profile focuses on a macro view than traditional neuron-wise explanations, providing insights for researchers about how neural network behaves through different layers.

References

Bau, David, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. "Network Dissection: Quantifying Interpretability of Deep Visual Representations." In *Computer Vision and Pattern Recognition*.

Oikarinen, Tuomas, and Tsui-Wei Weng. 2023. "CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks." *International Conference on Learning Representations*

Appendices

A.1 Additional Figures

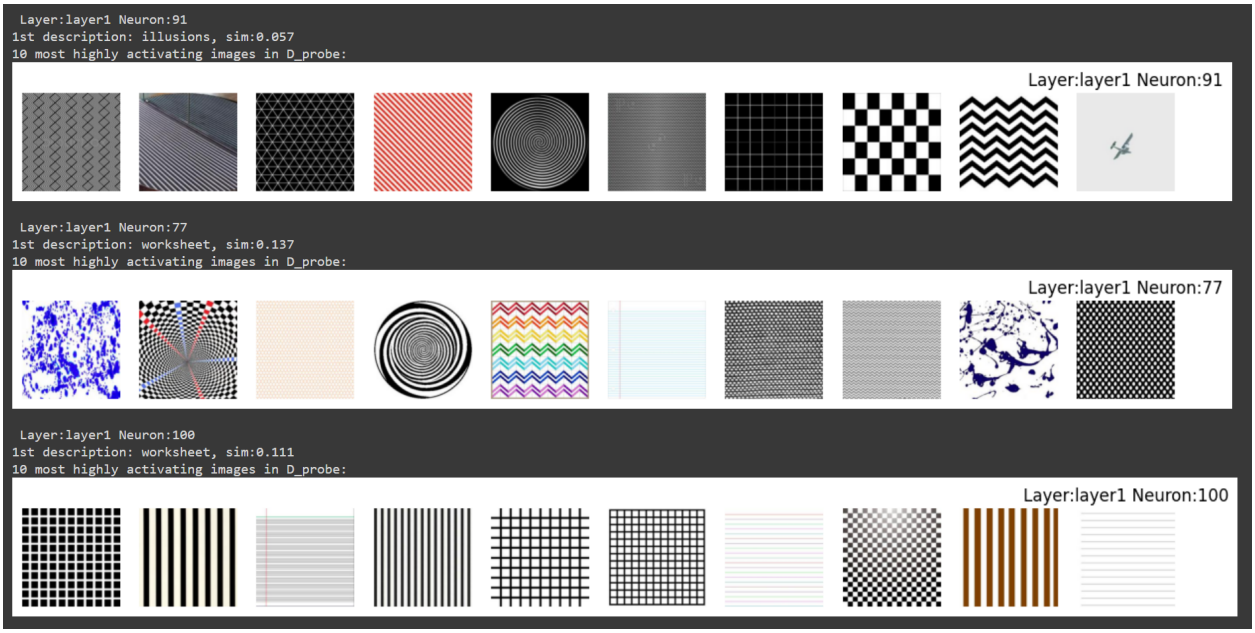

Figure A 1: Examples for human evaluation process to determine interpretable threshold