DnR: Describe and Refine

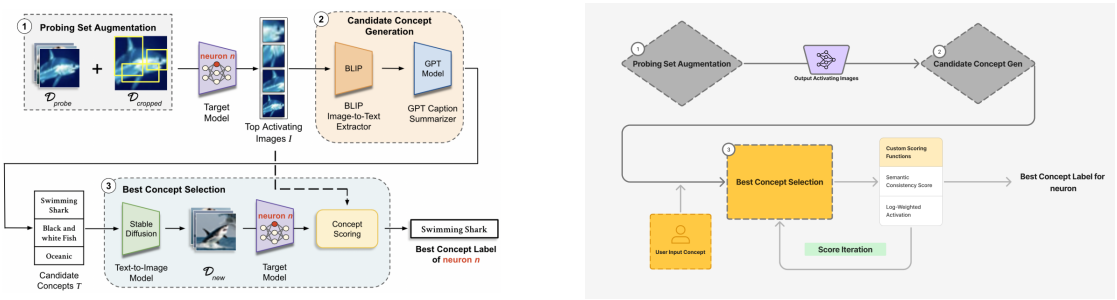
Chia Lee
chl147@ucsd.edu

Minchan Kim
mik042@ucsd.edu

Abstract

Understanding hidden neurons in vision models is key to improving interpretability. Describe-and-Dissect **(DnD)** (Bai et al. 2025) is a training-free method that generates highly rated neuron explanations without labeled data or retraining. However, we believe that we could build on and even improve model performance by incorporating more machine learning techniques. We introduce Describe-and-Refine **(DnR)**, an effort to enhance **DnD** by adding several learning techniques to improve neuron interpretation. We introduce reinforcement learning by adding the option for users to input concepts, create custom scoring functions for measuring which concept best fits a neuron, and an iterative process for stable diffusion image generation and rating candidate concept accuracy to find the best concept fit for the highest activating generated images.

Code: https://github.com/m1nce/Describe-and-Refine



(a) DnD (b) DnR

Figure 1: Describe-and-Dissect Versus Describe-and-Refine

1 Introduction

1.1 Contextualizing the Problem

The ability to interpret the behavior of hidden neurons in Deep Neural Networks (DNNs) is critical for improving the transparency and reliability of these models, especially in high-stakes applications like medical imaging or autonomous systems. While vision networks achieve remarkable performance, they are often seen as a ”black box” model as their interpretability often lags behind, making it challenging to understand how individual neurons in the entire landscape of the network contribute to the model's overall functionality. Existing methods for neuron interpretation either rely heavily on labeled datasets, which are time-intensive to curate, or require retraining models, which can be computationally intensive.

1.2 Conceptualizing the Gap

To address these limitations, the Describe-and-Dissect framework was introduced. **DnD** offers a training-free method to interpret hidden neurons in vision networks, leveraging multi-modal models to generate meaningful concept descriptions without requiring labeled training data or predefined concept sets. **DnD's** iterative feedback loop ensures that the generated labels are both relevant and accurate, outperforming traditional neuron labeling techniques in terms of interpretability and usability. Reproducing **DnD's** results on different platforms and environments will be put to the test, particularly in scalable, cloud-based setups like UCSD's Data Science Machine Learning Platform (DSMLP) on Datahub.

1.3 Resolving the Problem

In this paper, we aim to reproduce the results demonstrated in the original **DnD** paper and improve on them with our own conceived methods **(DnR)** using the DSMLP environment on Datahub. By improving on established scoring methods and building on the pipline's

concept selection capabilities, we seek to validate **DnD's** effectiveness and create further insight on how specific image types activate more highly in certain neurons than others. Our study contributes to the growing effort to make neuron interpretability more accessible and effective for broader research applications.

2 Objectives

- Output more robust neuron concept explanations
- Improve model performance and runtime efficiency
- Reveal insight on black box neural network functionality in vision networks
- Explore potential for reinforcement learning and user improvement in a label-free neuron descriptor

3 Methods

3.1 How DnD Works

This section details the methodology used in **DnD's** approach for neuron interpretability. This process includes selecting an initial probing set of images, associating these neuron activations with textual concepts for human understanding, and finally refining these associations through iterative feedback using image generation and ranking.

3.1.1 Step 1: Activation Identification

In order to determine what specific visual features a neuron responds to, a diverse set of probing images is fed into a arbitrary vision network (can be passed through into the **DnD** architecture). Alongside these full images, smaller sub-crops of the images are also used to isolate regions that trigger strong activations in the target neuron. This mimics the localized feature detection performed by convolutional layers in Convolutional Neural Networks, which identify spatially relevant features in images.

For each image, the network computes activation values for the target neuron. Images (or image regions) that produce high activation values are selected as "highly activating images." These images serve as the foundation for understanding the neuron's function, as they capture the "stimuli" the neuron is most sensitive to.

3.1.2 Step 2: Concept Extraction

The selected highly activating images are passed through an image-to-text extraction model (e.g., Contrastive Language-Image Pre-Training (Hafner et al. 2021) by OpenAI or a sim-

ilar multi-modal model. This step generates descriptive textual phrases for each image, encapsulating the visual features that might correspond to the neuron's behavior.

For instance, if a neuron activates strongly for images containing "striped patterns," the image-to-text model might output phrases like "striped texture" or "zebra pattern." These initial textual descriptions are then summarized into a set of concise, high-level concepts using GPT, ensuring that the descriptions are generalizable and interpretable.

3.1.3 Step 3: Image Generation and Feedback

To validate the relationship between the neuron and extracted concepts, the summarized concepts are fed into Stable Diffusion, a powerful text-to-image generation model. Stable Diffusion produces synthetic images based on the textual descriptions.

These synthetic images are then reintroduced to the neuron in the vision network, and the activations are re-evaluated. This feedback process ensures that the extracted concepts are not only representative of the neuron's behavior but are specific enough to avoid irrelevant associations. For example, if a neuron is labeled as a "striped texture," the regenerated images should consistently trigger high activations in the neuron. If they do not, the process highlights possible mismatches for further refinement.

3.1.4 Step 4: Concept Ranking and Labeling

The final synthetic images that are produced by the aforementioned feedback loop are once again passed through the image-to-text extractor, producing a refined set of candidate concepts for the neuron. These concepts are then ranked based on their alignment with the neuron's activations and other criteria such as specificity and consistency.

The top-ranked concept, which best explains the neuron's role across multiple iterations, is selected as the neuron's label. This label serves as a human-understandable explanation of what the neuron "does," in the context of the entire neural architecture, making the network more interpretable.

3.2 Experimental Methods

This section explores the methods used for **DnR** in order to improve neuron interpretability. This process includes building and improving on top-k scoring methods, refining the concepts through optimizing the iterative step for image generation and scoring, and finally exploring incorporating user input concepts to reinforce the model's concept set.

Our testing for **DnR** was performed on the ImageNet public dataset. We used the University of California San Diego's Data Science Machine Learning Platform, a Kubernetes cluster for running docker containers, to run the experiments with a GPU.

3.2.1 Top-K Scoring

In deep neural networks, not all neurons contribute equally to meaningful feature representations. Some neurons activate strongly for specific concepts, while others may fire due to random noise or irrelevant patterns. Traditional methods, such as mean or median activation, fail to capture the most relevant neurons because they treat all neurons equally, including weak and noisy ones. To address this limitation, we added additional Top-K scoring methods based on contemporary studies (Sander et al. 2023) to prioritize the strongest activations of a neuron, ensuring that we have efficient and important responses, rather than averaging all inputs.

We implemented Top-K Log-Weighted Activation and Top-K Semantic Consistency because they offer complementary advantages in neuron ranking.

$$S_{\text{log-k}} = \frac{1}{K} \sum_{i=1}^{K} \log(1 + a_{(i)})$$

Top-K Log-Weighted Activation highlights neurons with some strong spikes in activity, applying a logarithmic transformation to prevent extreme values from dominating while still highlighting their importance. It makes it ideal for filtering out weak activations while ensuring neurons with strong responses are ranked highly.

$$S_{\text{semantic-k}} = \alpha \cdot \left(\frac{1}{K} \sum_{i=1}^{K} a_{(i)} \right) + (1 - \alpha) \cdot \text{CLIP}(n)^2$$

Top-K Semantic Consistency, on the other hand, goes beyond just activation strength by incorporating semantic similarity (CLIP scores), ensuring that neurons are not just highly active, but also meaningful in relation to human-understandable concepts.

By combining these two approaches, we can more effectively rank neurons that are both functionally important conceptually relevant, improving the interpretability and analysis of the network.

3.2.2 Iterative Scoring

We initially devised a more complex methods of iterative scoring: to create a formula for gradient descent to minimize the score of the stable diffusion generated concepts. However after a long period of testing, we realized that this was not feasible given the randomness of the images generated by stable diffusion. Instead, to test our idea, we implemented an iterative loop into the pipeline that compared the average scores of the top concepts from the prior iteration to the current one. All concepts are saved in a master list, which is called upon for comparison in the final iteration. We initially tested this loop with the original scoring functions, then proceeded with our custom-made scoring functions.

3.2.3 User Input Reinforcement Learning

To introduce a larger variety of concepts and increase robustness in the model, we create a step in the **DnD** pipeline to allow for a user-chosen concept to be inputted into the overall concept set. This step is done before stable diffusion to allow for greater coverage of potential concepts and wider choice of generated imagery. This step is largely left to the discretion of the user based on what they deem to be an appropriate concept description similar to or better than the initial candidate concept set generated in step two.

4 Results

Table 1: Table of Scoring Methods and Neuron Metrics

	Scoring Method	**Neuron ID**	**Rank**	**Score**
1023	Top-K Squared Mean	927	1024	0.565151
1605	Mean	927	582	0.037649
2233	Median	927	186	0.011822
3981	Squared Mean	927	910	0.011951
4125	Top-K Log-Weighted Activation	927	30	0.580204
5150	Top-K Semantic Consistency	927	31	0.551183

Our results demonstrate the advantages of using Top-K scoring methods over traditional activation-based ranking techniques in identifying the most functionally and semantically meaningful neurons. Specifically, Top-K Log-Weighted Activation and Top-K Semantic Consistency ranked neuron 927 among the top 30-31 neurons, whereas traditional methods like Mean (Rank 582), Median (Rank 186), and Squared Mean (Rank 910) ranked it significantly lower. This discrepancy suggests that neuron 927 exhibits a few strong activations, rather than maintaining consistently high responses across all inputs. The high scores produced by Top-K methods ($\sim$0.58 and $\sim$0.55) indicate that these activations are not only strong but also semantically aligned with meaningful concepts, reinforcing the importance of considering peak activations rather than averaging across all responses.

Conversely, ranking methods based on Mean and Median activations distribute importance evenly across all activations, causing neurons with sporadic but highly significant activations to be down-ranked. Additionally, the Top-K Squared Mean method ranks neuron 927 at 1024, despite a high score of 0.565, suggesting that while neuron 927 exhibits strong activations, it is outperformed by others with even greater cumulative squared activations. These results highlight the limitations of global averaging techniques, which fail to differentiate between neurons that are consistently active and those that play a more specialized, high-impact role. By selectively considering only the most relevant activations, our Top-K methods provide a more precise and interpretable ranking, enabling better identification of neurons that contribute meaningfully to feature representations in deep networks.

Iterative scoring is fairly inconsistent compared with baseline **DnD**, with some iterations of the same neuron ending in two or three iterations while others continue for more. This results in mostly the same concepts being generated in some iterations, while others result in very marginal score differences. The randomness of the Stable Diffusion also makes it difficult to quantify why this could be the case without a better method of comparing images to concept sets during the generation process, as all figures are generated before and after step 3. Overall, it is difficult to quantify whether performance is improved outside of comparing scoring results. The same can be said with user input concepts, which are often scored lower than the gpt generated concepts and fail to generate better output in stable-diffusion concept sets.

5 Conclusion

Although Top-K scoring consistently shows most accurate results for neuron interpretability, it does not explain why each neuron activates highly to certain images. This lack of explanation for the efficiency of Top-K scoring in finding the best activating concepts makes it difficult to determine why we achieve the results that we have. In a way, there is still much to the "black box" of neural networks that we still do not understand.

Our method of incorporating user input concepts as well as our iterative method also needs refinement due to the redundancy of using user concepts in a AI generation pipeline. We also have yet to incorporate methods of reducing stopping at local minima for the iteration threshold, resulting in non-optimal concept scoring. This could be improved with further exploration of optimization algorithms. User input could be better implemented through feature tuning systems and instant scoring functionality. We may plan on exploring these systems in a future experiment.

6 Appendix

6.1 Broad Problem Statement

Our proposal for our Quarter 2 project presents CLEAR: Contextual Latent Exploration for Adaptive Representation. We plan on exploring and building on the foundation laid by contemporary methods of understanding neural networks (Oikarinen et al. 2023), (**?**). We believe that we can further improve on these established methods for visually guided neural network explanation by improving on each model's context adaptability, allowing for models that are more robust and perform better in a greater variety of contexts. We plan on testing a variety of image sets in our research, ranging from strategically blurred imagery to applying image corruptions and filters. By tuning the image sets already used in the prior experiments, we will discover new insights to build upon for feature tuning and model performance.

6.2 Domain Problem Statement

Unlike traditional concept bottleneck models, we believe that applying CLEAR will improve model adaptability environmental changes, maintaining robust interpretability by adjusting concept representations based on context. Through statistical regularization, CLEAR methods will ensure distinct, interpretable concepts. This adaptability and transparency will make CLEAR well-suited for complex, real-world applications where data variability and interpretability are crucial. For image blurring, we plan on applying convolution of image pixels within a bound with a Gaussian kernel, similar to methods detailed in (**?**). Previous studies found success in isolating foreground and background through image blurring to improve context recognition, but only through the use of mixed extractors and image classifiers. Other methods (**?**) make use of significant training and image transformers, whereas the methods we will explore have no training and are generative. Our research will hopefully reveal a crucial flaw, or solidify the performance of aforementioned neuron explainer models.

6.3 Primary Output Statement

We plan on communicating our results through a research paper/report. All findings, data analyses or additional reflections, will be provided through a Latex document. We may also consider creating a website to display our findings in a digestible format. All extra analyses and experimental documentation will be included in the appendix of the research paper. Figures will be created using jupyter notebooks in the source code.

7 Contribution

Chia: Created iterative concept selection and user input concept in pipeline. Wrote final report, worked on poster.

Minchan: Wrote concept scoring functions and worked on the website. Worked on final report and poster.

References

Bai, Nicholas, Rahul A. Iyer, Tuomas Oikarinen, Akshay Kulkarni, and Tsui-Wei Weng. 2025. "Interpreting Neurons in Deep Vision Networks with Language Models." [Link]

Hafner, Markus, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. "CLIP and complementary methods." *Nature Reviews Methods Primers* 1 (1): 1–23

Oikarinen, Tuomas, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. 2023. "Label-free concept bottleneck models." *arXiv preprint arXiv:2304.06129*

Sander, Michael E., Joan Puigcerver, Josip Djolonga, Gabriel Peyré, and Mathieu Blondel. 2023. “Fast, Differentiable and Sparse Top-k: a Convex Analysis Perspective.” [Link]