Contextual Latent Exploration for Adaptive Representation

Chia Lee
chl147@ucsd.edu

Minchan Kim
mik042@ucsd.edu

Abstract

Contextual Latent Exploration for Adaptive Representation **(CLEAR)** is a research experiment that builds on Describe-and-Dissect **(DnD)**, a new method of inference on hidden neuron behavior in vision networks. **DnD** is training free and requires no model training, nor does it need labeled training data or concept sets. Compared to more contemporary methods, **DnD's** labels are more highly rated as the correct explanation when compared to a neuron's baseline definition. We 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 test a variety of image sets in our research, ranging from strategically blurred imagery to applying image corruptions and filters, as well as different blurring techniques that isolate an image's foreground and background. By tuning the image sets already used in the prior experiments, we discover new insights to build upon for feature tuning and model performance.

Code: https://github.com/m1nce/CLEAR.git

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 impossible.

In addition, neural networks may run into the issue of classifying images with perturbations such as blurring, occlusions, noise, or other distortions. While standard deep learning models may perform well on clean datasets, their ability to generalize across different contexts remain questionable. In real-world applications, models must handle a variety of inputs that may not align with these models.

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 using the DSMLP environment on Datahub. By analyzing and labeling highly activating neurons, we seek to validate DnD's effectiveness and investigate its potential integration with Label-Free Concept Bottleneck Models (Oikarinen et al. 2023). This experiment will also serve as a foundation for our subsequent project, Contextual Latent Exploration for Adaptive Representation (CLEAR), where we plan to leverage DnD for feature tuning and image data manipulation. Our study contributes to the growing effort to make neuron interpretability more accessible and effective for broader research applications.

Furthermore, by contextualizing models in this way, we can better understand how they contribute to decision-making across different scenarios, leading to improved model trans-

parency and robustness. Through this exploration, we aim to uncover patterns in feature activation that may reveal previously unknown biases or vulnerabilities in deep vision models.

2 Methods

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.

2.1 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.

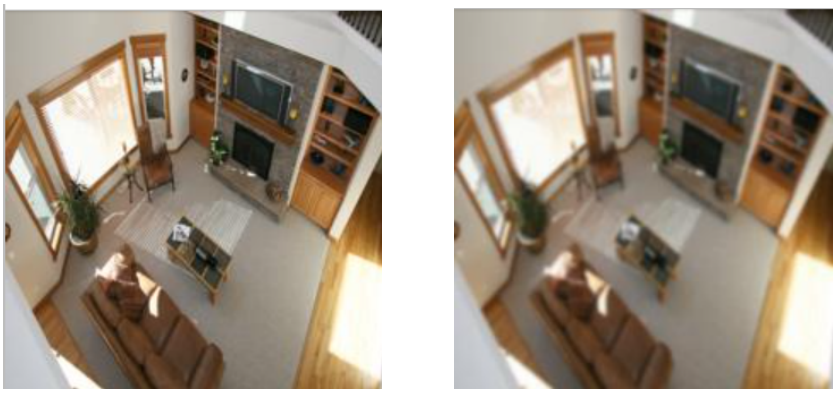
2.2 Step 2: Concept Extraction

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 similar 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.

2.3 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.



(a) Clear (b) Blurred

Figure 1: Applied Gaussian Kernel to Image

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.

2.4 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 beset explainst he 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.

2.5 Experimental Step: Gaussian Blur

To make use of the methods above, we write scripts to apply Gaussian blur to the image set data. These images are then fed back into the network to produce new results that we can use to do feature tuning and achieve better understanding of the model. We have a blurring method for general blur and are working on a more isolated blurring method. This should in theory be more powerful in determining the image contexts and interpretation of the dataset. If "blur" can also be isolated as a feature, it may lead to new insights that help identify less-than-perfect images.

3 Appendix

3.1 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.

3.2 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.

3.3 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.

4 Contribution

Chia: Ran baseline experiments, ran experiments with blurred images, documented figures for experiment results, created slides for weekly check-ins.

Minchan: Wrote script to blur image sets, writing script/ neural network for selective image blur, wrote weekly PA updates, compiled issues for experiment Github

References

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*