Automated Natural Language Explanation of Deep Visual Neurons with Large Models (Student Abstract)

Chenxu Zhao[1], Wei Qian[1], Yucheng Shi[2], Mengdi Huai[1], Ninghao Liu[*2]

[1]Department of Computer Science, Iowa State University
[2]School of Computing, University of Georgia
616 Boyd Graduate Studies Research Center Athens, GA, 30602
{cxzhao, wqi, mdhuai}@iastate.edu, {yucheng.shi, ninghao.liu}@uga.edu, Phone: 4047191028

Abstract

Interpreting deep neural networks through examining neurons offers distinct advantages when it comes to exploring the inner workings of Deep Neural Networks. Previous research has indicated that specific neurons within deep vision networks possess semantic meaning and play pivotal roles in model performance. Nonetheless, the current methods for generating neuron semantics heavily rely on human intervention, which hampers their scalability and applicability. To address this limitation, this paper proposes a novel post-hoc framework for generating semantic explanations of neurons with large foundation models, without requiring human intervention or prior knowledge. Experiments are conducted with both qualitative and quantitative analysis to verify the effectiveness of our proposed approach.

Introduction

Comprehending the behavior of modern machine learning models, particularly deep neural networks (DNNs) remains a significant challenge. Interpreting DNNs from the neuron perspective unravels the roles of individual neurons, which has proven to be effective in exploring the inner workings of deep models (Bau et al. 2020). Existing techniques for understanding deep neuron behaviors are still limited. Approaches based on visualization (Zhou et al. 2014) could only identify the relevant features (image regions) of the target neuron, but cannot explain the meaning of the features. This leaves much of the explanation and analysis work to humans, which can lead to a burdensome workload, especially given the fast growing scale of modern models.

To address the above challenges, in this paper, we present a novel interpretation approach, which can automatically generate semantic explanations for neurons of DNNs trained on different datasets. Our approach is model-agnostic, simple and flexible, eliminating the need for additional training or manual data collection steps. To validate the efficacy of our proposed method, we conduct a series of experiments. We delve further into exploring the semantic information contained within neurons. Additionally, we carry out a neuron ablation to quantitatively ascertain the significance of these semantic neurons.

[*]indicates corresponding author.

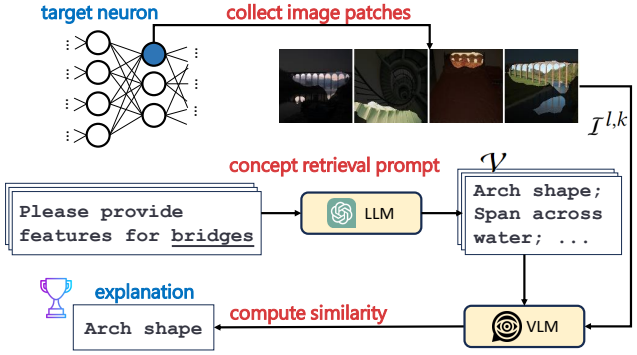

Figure 1: An illustration of the proposed method for explaining a target deep neuron with an LLM and a VLM.

Methodology

Provided with a target classification model and a designated subset of neurons within it, our objective is to understand and describe the semantics of these neurons using natural language tokens. Our proposed methodology comprises three phases, including: 1) image patches collection, 2) explanation vocabulary construction, and 3) explanation generation. Specifically, we first extract a set of activated image patches that are associated with a given neuron of the target model. In this way, we can identify the specific regions of the input images that contribute to the neuron's activation. Second, we construct a vocabulary of semantic concepts that is tailored to the application domain (e.g., the predicted class label). To obtain a comprehensive vocabulary, we propose prompting large language models (LLMs) to leverage their common-sense knowledge. The vocabulary could constrain the generated description to reduce randomness in explanation results. Third, we establish the connection between image patches and the matching concepts to produce explanations. Instead of constructing a dataset to train the annotation model (Hernandez et al. 2022), we adopt off-the-shelf vision language models (VLMs) such as CLIP (Radford et al. 2019) to align appropriate concepts with the image patches. By going through the concept vocabulary, we are able to generate a comprehensive and meaningful explanation to understand the target neuron's behavior. *More method details can be found in the supplementary (Zhao et al. 2023).*

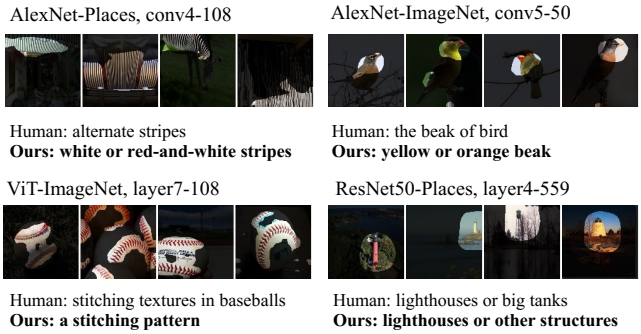

Figure 2: Illustrations of neuron explanations.

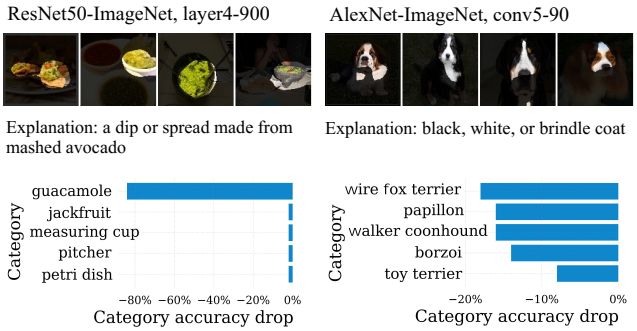

Figure 3: Test accuracy drop after ablating a unit.

Experiments

Experimental Settings

In experiments, we adopt ImageNet and Places365 datasets. We analyze and compare the interpretability of neurons within the convolutional layers of ResNet50 and AlexNet, as well as the neurons within the MLP layers of Vision Transformers (ViT). To compare the interpretability of neurons in DNNs, we have five human observers to look at each top-activating image patch and ask them to describe common features or patterns in phrases.

Performance of Neuron Explanations

We begin by evaluating the performance of neuron explanations in various layers in deep neural networks. We query the GPT-3 for the feature descriptions from categories in ImageNet and Places365. We utilize the ViT-B/32 model in CLIP to label the neurons' descriptions for the activated image patches. Figure 2 shows the experimental results on different pre-trained models. The patch generated by each unit is shown on four maximally activating images. Firstly, our neuron explanations demonstrate a high level of agreement with human annotations. For instance, both descriptions for unit 559 in layer 4 of ResNet-Places depict "lighthouse" objects. Secondly, by comparing different layers within the models, we observe that the units in different convolutional layers detect different levels of patterns or features. Specifically, units in earlier layers usually capture low-level features like edges and simple textures, such as the "stripes" in conv4 of AlexNet-Places. As the network progresses deeper, the units tend to capture more abstract and generic visual features, such as the "beak" in conv5 of AlexNet-ImageNet. Overall, our proposed method demonstrates the capability to generate comprehensive descriptions without relying on human annotations, thereby providing efficient and effective explanations for neurons in deep neural networks.

Feature Importance Analysis

In this section, we explore the impact of neurons on model predictions. We conduct experiments to ablate a unit in the last convolutional layer, where semantic information emerges most, and we measure the resulting decrease in category accuracy. To ablate a unit, we set the weights and biases of its feature maps to 0, thereby eliminating its contribution to predictions for any input image. Figure 3 presents examples of units with valuable semantic information and illustrates their impacts on category accuracy. We can find that ablating a single unit leads to a significant category accuracy drop for specific categories. For instance, in the first example, when we ablate a unit representing "a dip or spread made from mashed avocado", the test accuracy of the "guacamole" category suffers a significant drop of 84%. These findings highlight the relative importance of neurons that capture attributes in influencing model behaviors. Particularly, they play a crucial role in recognizing specific subsets of categories within the dataset.

Conclusion

This paper introduces a novel post-hoc method for generating semantic explanations for neurons in deep models. The proposed method eliminates the need for human intervention and can be applied to any deep learning architectures and datasets without limitations. Extensive experiments have been conducted to verify the effectiveness of the proposed method. It is believed that this approach will serve as a valuable tool for the research community, enabling investigations into individual neurons in deep learning models.

References

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *In PNAS*.

Hernandez, E.; Schwettmann, S.; Bau, D.; Bagashvili, T.; Torralba, A.; and Andreas, J. 2022. Natural language descriptions of deep visual features. In *In ICLR*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Zhao, C.; Qian, W.; Shi, Y.; Huai, M.; and Liu, N. 2023. Automated Natural Language Explanation of Deep Visual Neurons with Large Models. *arXiv preprint arXiv:2310.10708*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2014. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.