Describe-and-Dissect: Interpreting Neurons in Vision Networks with Language Models

Nicholas Bai*, Rahul Ajay Iyer*, Tuomas Oikarinen , and Tsui-Wei Weng
University of California San Diego
nicholaszybai@gmail.com, rahul.ajay.iyer@gmail.com, toikarinen@ucsd.edu, lweng@ucsd.edu
*equal contributions

Abstract

*In this paper, we propose Describe-and-Dissect (**DnD**), a novel method to describe the roles of hidden neurons in vision networks. **DnD** utilizes recent advancements in multimodal deep learning to produce complex natural language descriptions, without the need for labeled training data or a predefined set of concepts to choose from. Additionally, **DnD** is* training-free*, meaning we don't train any new models and can easily leverage more capable general purpose models in the future. We have conducted extensive qualitative and quantitative analysis to show that **DnD** outperforms prior work by providing higher quality neuron descriptions. Specifically, our method on average provides the highest quality labels and is more than 2× as likely to be selected as the best explanation for a neuron than the best baseline.*

1. Introduction

Recent advancements in Deep Neural Networks (DNNs) within machine learning have enabled unparalleled development in multimodal artificial intelligence. While these models have revolutionized domains across image recognition and natural language processing, they haven't seen much use in various safety-critical applications, such as healthcare or ethical decision-making. This is in part due to their cryptic "black box" nature, where the internal workings of complex neural networks have remained beyond human comprehension. This makes it hard to place appropriate trust in the models and additional insight in their workings is needed to reach wider adoption.

Previous methods have gained a deeper understanding of DNNs by examining the functionality (also known as *concepts*) of individual neurons[1]. This includes works based on manual inspection [3, 4, 9, 13], which can provide high quality description at the cost of being very labor intensive. Alternatively, Network Dissection [1] automated this labeling process by creating the pixelwise labeled dataset, *Broden*, where fixed concept labels serve as ground truth binary masks for corresponding image pixels. While earlier works, such as Network Dissection, were restricted to an annotated dataset, CLIP-Dissect [8] offered a solution by no longer requiring labeled concept data, but still requiring a predetermined concept set as input. MILAN [6] sought to enhance the quality of neuron labels by providing generative descriptions, but their method requires training a new descriptions model from scratch to match human explanations on a dataset of neurons.

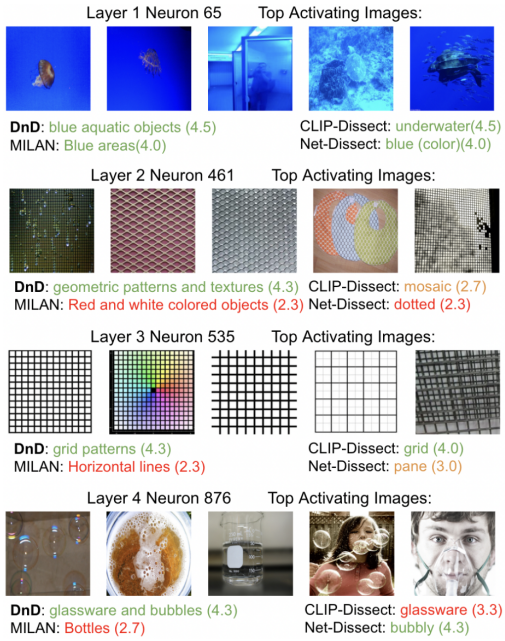

Figure 1. Neuron descriptions provided by our method (**DnD**) and baselines CLIP-Dissect [8], MILAN [6], and Network Dissection [1] for random neurons from ResNet-50 trained on ImageNet. We include the average quality rating from our Amazon Mechanical Turk experiment next to each label and color-coded the descriptions by whether we believed they are accurate, somewhat correct or vague/imprecise.

[1]We conform to prior works' notation and use "neuron" to describe a channel in CNNs.

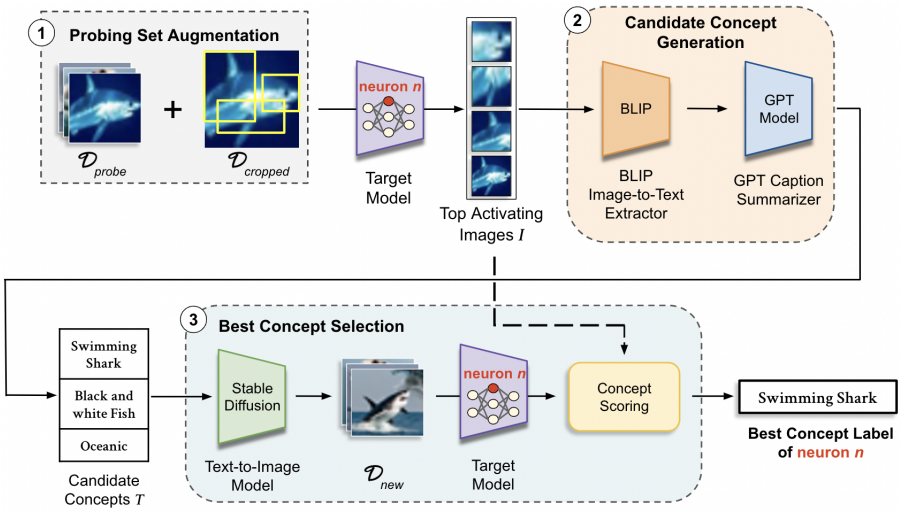

Figure 2. Overview of Describe-and-Dissect (**DnD**) algorithm. Given a Target model, it consists three important steps to identify the neuron concepts (e.g. 'Swimming Shark' for neuron n).

2. Proposed Method

We present Describe-and-Dissect (**DnD**), a comprehensive method to produce generative neuron descriptions in deep vision networks. Our method is training-free, model-agnostic, and can be easily adapted to utilize advancements in multimodal deep learning. **DnD** consists of three steps shown in Fig. 2:

- **Step 1. Probing Set Augmentation:** Augment the probing dataset with attention cropping to include both global and local concepts;
- **Step 2. Candidate Concept Generation:** Generate initial concepts by describing highly activating images [7] and subsequently summarize them into candidate concepts using GPT 3.5 [2];
- **Step 3. Best Concept Selection:** Generate new images based on candidate concepts and select the best concept based on neuron activations on these synthetic images [11] with a scoring function.

Scoring Function. For a given neuron, we use a scoring function to rate candidate concept accuracy. Simple metrics such as mean are heavily prone to outliers that result in skewed predictions so we propose a scoring function that weights the average rank of top activating images mapping to a candidate concept.

$$score(R_j, I, \mathcal{D}_j^t) = (N - \text{Rank}(R_j)) \cdot E(I, \mathcal{D}_j^t)$$

Here, the average rank of images for candidate concept j, $\forall j \in \{1, ..., N\}$, is denoted R_j and $\text{Rank}(R_j)$ sorts R_j in increasing order. $E(I, \mathcal{D}_j^t)$ computes the average cosine similarity between image embeddings of $\mathcal{D}_j^t$ and I using CLIP-ViT-B/16 [10], with $\mathcal{D}_j^t \subset \mathcal{D}_j$ for t highest activating images. In practice, R_j is computed as the square of the ranks in top β ranking images for better differentiation between scores, $R_j = \{(R_j^i)^2; i \leq \beta\}$.

3. Results

Table 1. **Averaged AMT results across layers in ResNet-50**. Our descriptions are consistently rated the highest and chosen as the best more than twice as often as the best baseline.

Metric / Method	NetDissect	MILAN	CLIP-Dissect	**DnD (Ours)**
Mean Rating	3.14	3.21	3.67	**4.15**
selected as best	12.71%	13.29%	23.11%	**50.89%**

Our crowdsourcing experiment compares the quality of labels produced by **DnD** against 3 baselines: CLIP-Dissect [8], MILAN [6], and Network Dissection [1]. We evaluate 800 randomly chosen neurons across 4 intermediate layers of ResNet50 on ImageNet [12]. Each neurons description is evaluated by 3 different workers. Shown in Table 1, **DnD** performs over 2× better than all baseline methods when dissecting ResNet-50 [5]. Results for an identical experiment using 200 randomly chosen neurons for ResNet18 [5] trained on ImageNet and Places365 [14] yielded similar results, with a mean rating of 4.16 and selected 63.21% of the time.

We also follow CLIP-Dissect [8] to quantitatively analyze description quality on last layer neurons, which have known ground truth labels. Results show **DnD** outperforms MILAN with a greater average CLIP cosine similarity by 0.0518, average mpnet cosine similarity by 0.18.

Additional qualitative figures, ablation studies, use case examples, and evaluation on MILANNOTATIONS will be publicly available at our website after publication.

Acknowlegements

This work is supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gbps networks. This work used Expanse CPU, GPU and Storage at SDSC through allocation CIS230152 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296. The authors thank REHS program (Research Experience for High School students) in San Diego Supercomputer Center. This work is also partially supported by National Science Foundation under Grant No. 2107189 and 2313105, and Hellman Fellowship.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *Computer Vision and Pattern Recognition*, 2017. 1, 2

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. 2

[3] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. 2009. 1

[4] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. *International Conference on Learning Representations*, 2022. 1, 2

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2

[8] Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *International Conference on Learning Representations*, 2023. 1, 2

[9] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020. 1

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[13] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 1

[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding, 2016. 2