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

Nicholas Bai*[1], Rahul Iyer *[1], Tuomas Oikarinen[1], Tsui-Wei (Lily) Weng[1]

*Equal contribution, work done during internship at UC San Diego [1]UC San Diego

ΨArxiv: https://arxiv.org/abs/2403.13771

ΨCode: https://github.com/Trustworthy-ML-Lab/Describe-and-Dissect

Motivation

- The internal workings of complex Deep Neural Networks (DNNs) have remained beyond human comprehension, stifling their use in various safety-critical applications.
- Due to this "black-box" nature, we cannot place appropriate trust in such models.
- **Our goal is to gain a deeper understanding of DNNs by examining the functionality of individual neurons.**

Related work

- Though previous works aiming to accomplish our goal have been based on manual inspection [3, 4, 8, 10], which can provide high quality description at the cost of being very labor intensive, other methods have automated this labeling process:

1) Network Dissection [1], creates the pixelwise labeled dataset, Broden, where fixed concept set labels serve as ground truth binary masks for corresponding image pixels, to match neurons to a label from the concept set. This causes the method to be greatly limited to an annotated dataset.

2) CLIP-Dissect [7] matches neurons to concepts based on their activations in response to images. It does not require labeled concept data, but still requires a predetermined concept set.

3) MILAN [5] provides generative descriptions, but requires training a new descriptions model from scratch to match human explanations on a dataset of neurons.

Reference

[1] Bau et al, Network dissection: Quantifying interpretability of deep visual representations. CVPR, 2017

[2] Brown et al, Language models are few-shot learners. CoRR, abs/2005.14165, 2020

[3] Erhan et al, Visualizing higher-layer features of a deep network. 2009

[4] Goh et al. Multimodal neurons in artificial neural networks. Distill, 2021

[5] Hernandez et al, Natural language descriptions of deep visual features. ICLR, 2022

[6] Li et al, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. 2022

[7] Oikarinen, T. and Weng, T.-W. Clip-dissect: Automatic description of neuron representations in deep vision networks. ICLR, 2023

[8] Olah et al, Zoom in: An introduction to circuits. Distill, 2020

[9] Rombach et al, High-resolution image synthesis with latent diffusion models. 2022

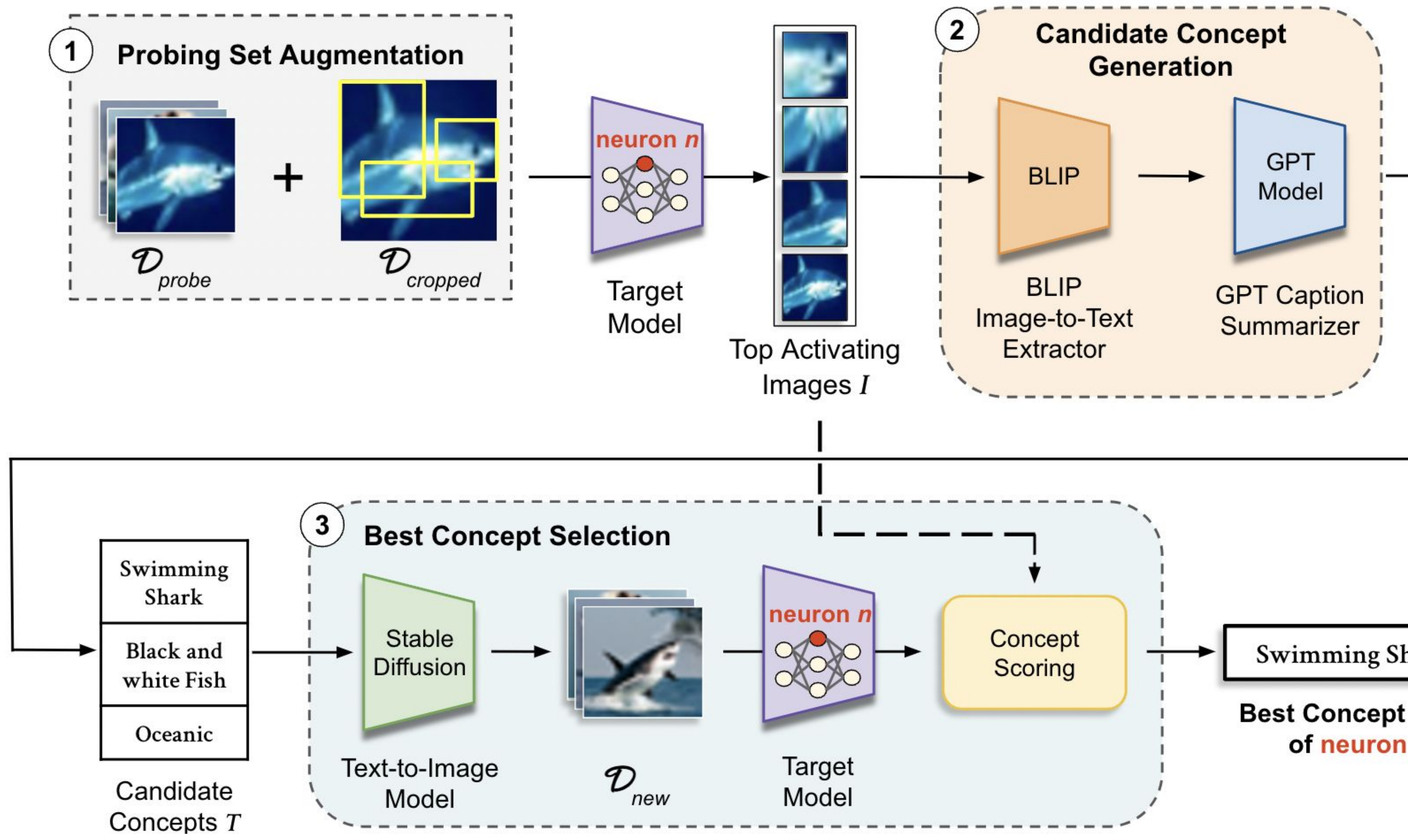
[10] Zhou et al, Object detectors emerge in deep scene cnns. arXiv:1412.6856, 2014

Method

We propose a comprehensive, training-free, and model-agnostic method that can be easily adapted to utilize advancements in multimodal deep learning.

Describe-and-Dissect consists of 3 steps:

1. ***Probing Set Augmentation:*** Augment the probing dataset with attention cropping to include both global and local concepts.
2. ***Candidate Concept Generation:*** Generate initial concepts by describing highly activating images [6] and subsequently summarize them into candidate concepts using GPT 3.5 [2].
3. ***Best Concept Selection:*** Generate new images based on candidate concepts and select the best concept based on neuron activations on these synthetic images [9] with a proposed scoring function, TopK Squared + Image Products.



Results

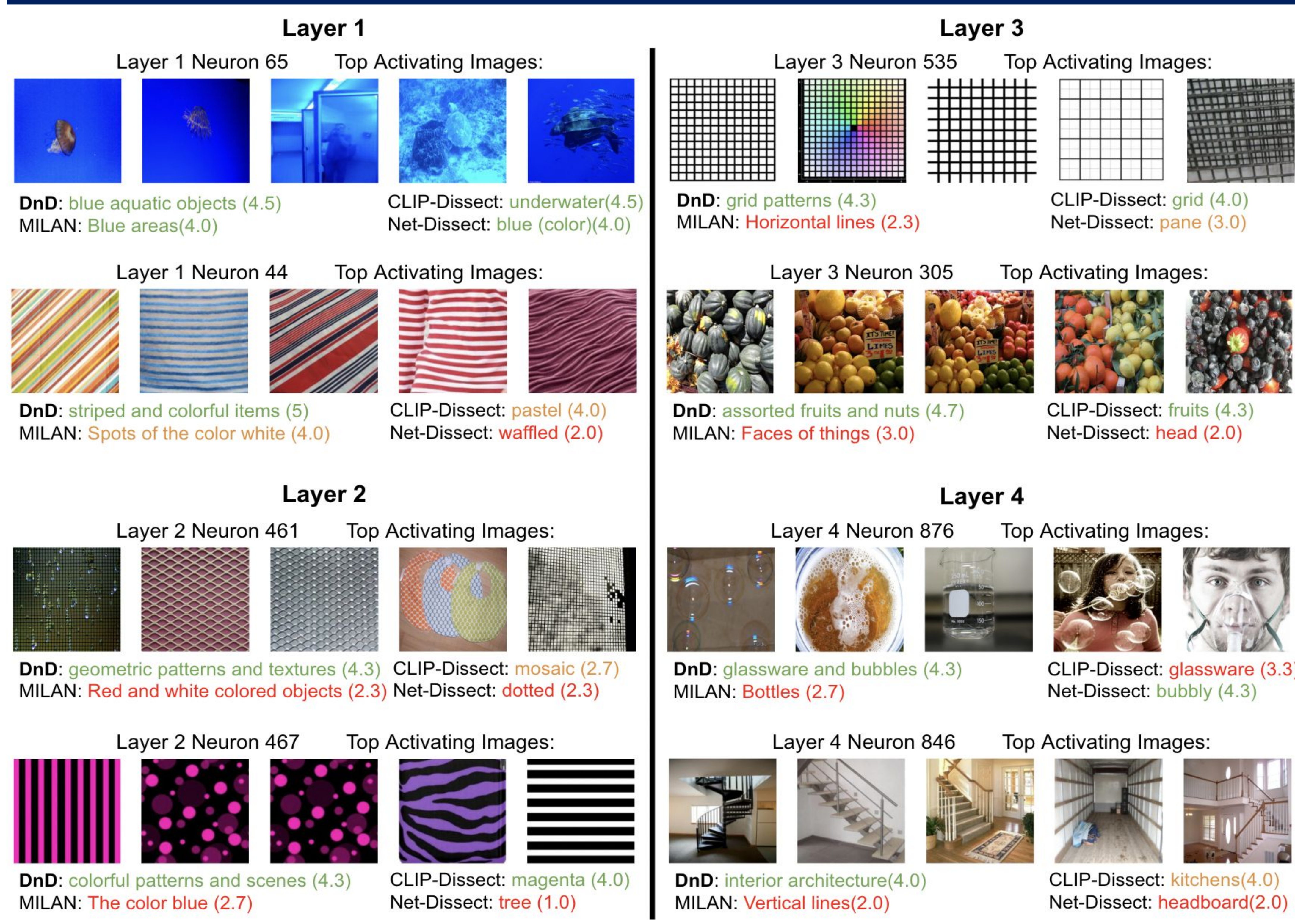


Fig 1: Neuron descriptions provided by our method (DnD) and baselines CLIP-Dissect, MILAN, and Network Dissection for random neurons from ResNet-50 trained on ImageNet.

		Method			
Metric	Layer	Network Dissection	MILAN	CLIP-Dissect	DnD (Ours)
Mean Rating	Layer 1	3.41 ± 0.058	3.41 ± 0.060	3.63 ± 0.057	$\mathbf{4.16 \pm 0.041}$
	Layer 2	3.14 ± 0.067	3.12 ± 0.064	3.55 ± 0.057	$\mathbf{4.07 \pm 0.048}$
	Layer 3	3.04 ± 0.066	2.96 ± 0.066	3.66 ± 0.055	$\mathbf{4.14 \pm 0.042}$
	Layer 4	2.97 ± 0.066	3.34 ± 0.061	3.82 ± 0.054	$\mathbf{4.21 \pm 0.044}$
% time selected as best answer	Layer 1	13.18%	14.32%	22.50%	**50.00%**
	Layer 2	15.27%	12.41%	19.33%	**52.98%**
	Layer 3	11.82%	12.73%	25.00%	**50.45%**
	Layer 4	10.56%	13.71%	25.62%	**50.11%**

Tab 1: AMT results for individual layers of ResNet-50. Our descriptions are consistently rated the highest and chosen as the best more than twice as often as the best baseline.

Metric / Methods	MILAN	DnD (Ours)
CLIP cos	0.7080	**0.7598**
mpnet cos	0.2788	**0.4588**
BERTScore	0.8206	**0.8286**

Tab 2: Textual similarity between predicted labels and ground truths on the fully-connected layer of ResNet-50 trained on ImageNet. We can see DnD outperforms MILAN