

SAND: Enhancing Open-Set Neuron Descriptions through Spatial Awareness

Anvita A. Srinivas¹, Tuomas Oikarinen¹, Divyansh Srivastava¹, Wei-Hung Weng², Tsui-Wei Weng¹
¹University of California San Diego, ²Google Research

{a5sriniv, toikarinen, ddivyansh, lweng}@ucsd.edu, ckbjimmy@google.com

Abstract

We propose *Spatially-Aware open-set Network Dissection (SAND)*, a technique to identify and label the learned representation of the neurons of deep vision networks. Contrary to earlier open-vocabulary neuron explanation methods, we also leverage a neuron’s spatial pattern of activation to guide our predictions towards more accurate and relevant concepts, while avoiding being misled by confounding visual information. We highlight important regions for a neuron through image masking, which has the advantage of being able to block out irrelevant concepts from an image, handling irregularly shaped activation regions, and revealing the visual concepts that a neuron learns in order to identify objects. We use CLIP to connect highly activating image regions with descriptive concepts, and measure the quality of our results through human evaluation. Further, since such manual evaluation can be highly time consuming, costly, and unscalable, we also propose an automated approach which uses image generation to get quantitative feedback on the generated concepts. Finally, as an application of our interpretability method, we demonstrate how it can be tuned to the medical domain. Our code is available at <https://github.com/Trustworthy-ML-Lab/SAND>

1. Introduction

With the rise of AI systems in real-world applications, it is increasingly crucial to understand their decision processes. Safety-critical domains such as healthcare, autonomous driving, and security – where models’ predictions can have a significant and lasting impact on a population – are especially in need of mechanisms that reveal models’ failure cases and biases, in order to improve their trustworthiness.

Several feature attribution methods [26] [30] [32] aim to understand what input features influence vision models’ final predictions. However, while these approaches help understand what a model is looking at, they do not immediately reveal how the model interpreted that visual information, or why it was of interest. With neuron-level expla-

nations of a model’s behavior, one can potentially uncover how the model *ties together* the diverse concepts it identifies, and how it subsequently uses that understanding to reach a final prediction.

Recently, there have been a number of neuron-level explanation methods [2, 3, 10, 24] which attempt to describe a neuron’s learned representation by examining its highly activating images. However, images in public datasets often contain confounding information and spurious correlations between classes, as previous works have shown (e.g. ‘dog’ and ‘frisbee’ [15], ‘water’ and ‘waterbirds’ [29], or ‘pacifier’ and ‘baby’ [36]). In the context of network dissection [2, 3], looking at image-level information alone can lead to inaccurate concepts in such cases, such as labelling a ‘frisbee’ detecting neuron as a ‘dog’ detector. Thus it is imperative for a model to be able to separate such distinct concepts.

To address this issue, methods such as [2] [3] [10] ameliorate this challenge to an extent by explicitly focusing only on the relevant parts of an image that activate a neuron. Yet, because of their dependence on manually annotated visual concepts, their vocabulary is limited [2], and so is their capacity to generalize to out-of-domain tasks without retraining [10]. On the other hand, [24] supports an open-set vocabulary by using CLIP to connect a neuron’s highly activating images to a concept that describes them. However, since it operates at the image-level, this comes at the cost of losing local spatial information, and leaves the approach open to misdirection by spurious correlations in an image. [13] addresses this by cropping a tight bounding box around key regions of an image, but because of the mismatch between the irregularity of an activation region’s shape and the rigidity of a crop, it cannot prevent confounding visual information from going through.

We thus motivate the need for spatially aware open-set neuron descriptions that can address both these challenges. In our approach, as illustrated in Figure 1, we highlight activation regions of a neuron through image masking, because we find that this has two main advantages. First, it supports non-rectangular, irregularly shaped image regions, without needing to include spurious information from background

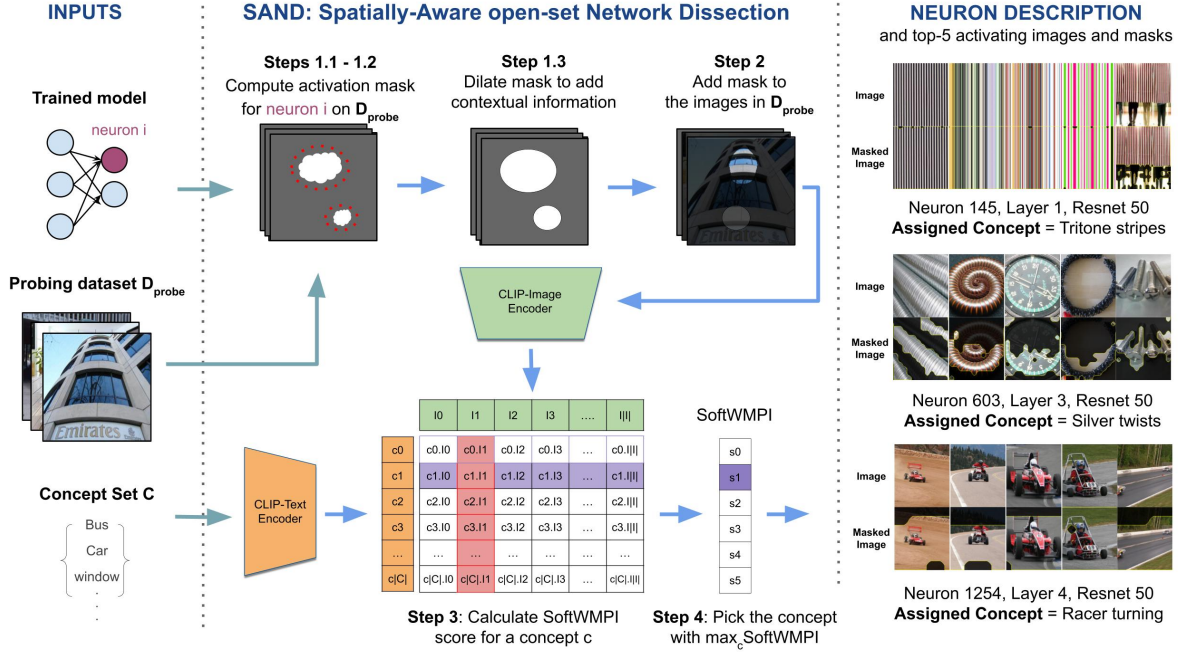


Figure 1. Pipeline for automated generation of a neuron’s description with SAND algorithm.

pixels. This is something cropping fails to do. As seen in Figure 2, it also has the added ability to *block out* regions that are irrelevant for a neuron, which is especially useful when a neuron focuses on the background of an image rather than its foreground. The second advantage emerged from our study of the activation patterns of lower layer neurons – masking helps identify the visual patterns neurons learn. For example, we show the activation pattern for the neuron corresponding to ‘large animal cage’ (Figure 2, Layer 3, unit 941), and notice how it places more importance on the animal’s fur than its facial features. In our experiments, we find this spatial information to be especially useful for more localized visual concepts, such as those in the early layers of neural networks, as well as in medical images.

Besides the proposed new generation pipeline (SAND), in this paper we also propose to improve the evaluation pipeline. This is because while human evaluation can provide a strong signal about the relevance of a neuron’s assigned descriptions, such manual inspection of its activation regions is a time consuming and tedious process; and with the steady increase in the number of parameters in present day models, dismally unscalable. To this end, we propose an automated method to evaluate neuron descriptions that leverages feedback from a neuron itself, combined with image generation, which we describe in detail in Section 3.3.

We summarize our main contributions below:

- We propose Spatially Aware open-set Network Dis-

section (SAND), a technique to automatically describe neuron representations without the need for manually labeled datasets, which utilizes local spatial information from images to remove confounding visual cues while generating concepts.

- We demonstrate the applicability of our method in two diverse domains – the natural image domain as well as the safety-critical medical domain.
- We propose an automated approach to evaluate neuron descriptions for the intermediate neurons of a model, that can significantly reduce the human effort needed for this task.

2. Related Work

Neuron explanation methods. This class of methods aims to explain the learned representation of every neuron of a DNN. For example, network dissection [2] [3] approach this by assigning a concept to a neuron based on the overlap between the highly activated regions of images from a probing dataset and a ground truth segmentation map. However, due to the dependence on pre-computed segmentation maps, its vocabulary is limited to the classes for which these maps are available. In contrast, our method leverages the massively pre-trained vision-language model CLIP [27] to connect image regions and neuron concepts, and is thus able to assign open-vocabulary concepts to neurons.

Some contemporary methods address this limitation in vocabulary by generating natural language expressions instead. For example, MILAN [10] generates explanations of neuron activations by maximizing the pointwise mutual information between image regions and human annotations during training time. While this can lead to a richer and more descriptive set of explanations, it is still dependent on the availability of a curated dataset, which can make generalization to out-of-domain contexts less accurate.

Our approach is similar to CLIP-Dissect [24] and FALCON [13], with a few key differences. CLIP-Dissect [24] generates concepts by using CLIP to pair open-vocabulary concepts with a neuron’s highly activating images. However, it does not incorporate critical spatial information from the local activation regions of images that highly activate a neuron, unlike our approach. FALCON [13] addresses this by cropping images to their activation regions before passing them through CLIP. However, cropping may include spurious and confounding information that can mislead a model, especially when the region of interest surrounds an irrelevant object. Our approach instead uses image masks to highlight important regions of an image, which can better cater to irregularly shaped activation regions.

Another category of methods, such as [18], approach neuron interpretability by focusing on a model’s erroneous decisions and analysing which neurons were responsible for it. Although such a method can measure the extent to which a neuron influences a prediction outcome, it does not indicate what learned representation the neuron itself embodies. To understand *why* the misclassification occurred semantically, such a method would need to be used in conjunction with a method like ours to discern such information.

In a similar vein, methods have been recently proposed to better describe the entire activation range of a neuron [17], [25]. While this is not our focus, these improvements could be combined with our approach. Other recent works such as [1], [31], also propose improving generative neuron descriptions by leveraging stronger pretrained models.

Feature attribution methods. This class focuses on instance-level interpretability rather than neuron-level interpretability, by examining how a model’s inputs influence its decisions. For example, saliency map methods such as [26], [30], [32], harness a model’s activations and/or gradients to highlight the image regions that were the most important for the model making its prediction. Similar to neuron-level explanations, TCAV [14] produces concept-based explanations, but it is focused on finding directions in activation space that are interpretable, instead of individual neurons, and requires labeled concept data.

Automated Description Evaluation. Traditionally, the quality of generated neuron descriptions has mostly been evaluated manually, by assessing how well the highly ac-

tivating images match the assigned description. However, this is labor intensive and prone to noise. While [4] propose automatically evaluating neuron descriptions via *simulation* in the language model setting, which [25] extend for vision models, they only measure behavior on pre-existing test data. Alternatively, the idea of using image-generation models as the basis of evaluation has been independently proposed by contemporary work [31], [16], but they differ from ours in how the explanations are scored, and do not utilize a prompt ensemble to generate more diverse images.

3. Methodology

In this section, we describe our approach to explain Deep Neural Networks (DNNs) on a global level by assigning natural language descriptions to an *individual neuron*. We call our approach **Spatially-Aware open-set Network Dissection**, or **SAND**, as it leverages spatial information from neuron activation region to assign open-vocabulary concepts to neurons. Our approach is agnostic to the vision-language model used, and can be applied to explain Convolutional Neural Network (CNN) layers in a DNN.

The rest of the section is organized as follows: Section 3.1 describes our algorithm to assign concepts to neurons, Section 3.2 describes the similarity function used to measure the relevance of a concept to a neuron, and finally, Section 3.3 describes our novel approach to automatically evaluate the quality of concepts assigned to the neurons by Neuron Explanation Methods (NEMs).

Notations. We define $X = \{x \mid x \in \mathbb{R}^{m \times n \times 3}\}$ as a set of images, Y as a set of labels, and $f : X \rightarrow Y$ to be a Convolution Neural Network (CNN) with L layers. i represents a neuron unit in f , and $f_i(x)$ denotes the i -th neuron’s activation map for input image $x \in \mathbb{R}^{m \times n \times 3}$, which is a three dimensional tensor. Further, let $\mathcal{D}_{probe}$ represent the probing dataset used for computing the activations of a neuron, and $\mathcal{C}$ represent the set of open-vocabulary concepts (i.e. descriptions) that can be assigned to a neuron.

Input and Output. Our method takes as input a probing dataset $\mathcal{D}_{probe}$, a set of open-vocabulary concepts $\mathcal{C}$, and a target model f . It outputs the concept $c_i^* \in \mathcal{C}$ that describes the neuron i in the target model f .

3.1. Automated neuron-level explanation

Figure 1 outlines our proposed algorithm to describe the visual concepts learnt by a DNN’s neuron. At a high level, we use images from $\mathcal{D}_{probe}$ to find the top activating images for a neuron, and process their activation maps to generate masks that can help direct the explanation method’s attention to relevant spatial regions for the neuron. Later, using a vision-language model such as CLIP, we extract the common concept seen across these regions. Our overall pipeline is described in detail below:

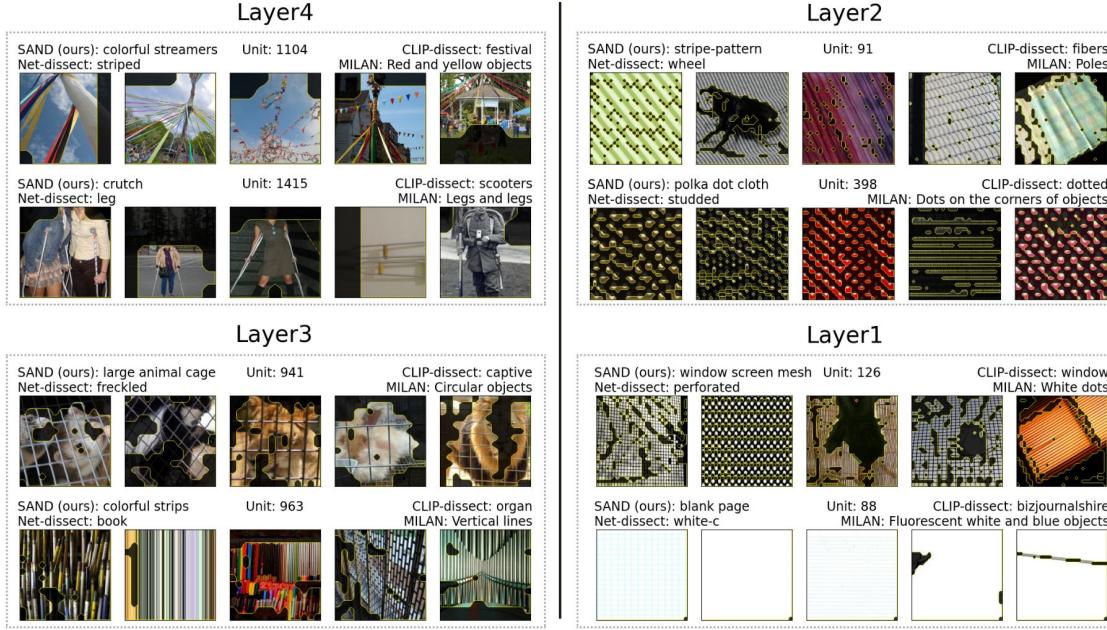


Figure 2. Visualization of the top activating images for neurons of ResNet50, and the concepts assigned to them by various methods. SAND not only helps localize and highlight visual concepts of interest while generating neuron descriptions (e.g. crutches in Layer 4, unit 1415), it also prevents confounding information from passing through by blocking out irrelevant regions for a neuron (e.g. the foreground objects in Layer 2, unit 91, and Layer 1, units 126 and 88).

- **Step 1.** Generate image masks for spatially relevant regions for a neuron.

1.1 Compute activation threshold $T_i \in R$ as the top η quantile level for probability distribution $\mathbb{P}$ of activation maps $f_i(x)$, $\forall x \in \mathcal{D}_{probe}$. Here, using the η quantile level allows us to adjust for the variability of neurons' activation scales.

$$T_i = \left[\max_t \mathbb{P}_x [f_i(x) > t] \geq \eta \right] \quad (1)$$

1.2 Create binary activation map $m_i(x) \in \{0, 1\}^{m \times n}$ by upsampling $f_i(x)$ to match the size of input image $x \in \mathbb{R}^{m \times n \times 3}$, using a bilinear interpolation function $\mathcal{B}$, and thresholding with T_i for all images in $\mathcal{D}_{probe}$. While we experimented with various levels of transparency for the mask, we found that using a binary mask worked best.

$$m_i(x) = \mathcal{B}(f_i(x)) \geq T_i, x \in \mathcal{D}_{probe} \quad (2)$$

1.3 Dilate the mask to add contextual information from the image while maintaining the shape of the activation region, as we find that this improves the quality of the final descriptions. The dilation operation is a convolution with a 15×15 kernel which takes the maximum value of the pixels in the kernel.

$$d_i(x) = \text{dilate}(m_i(x)), x \in \mathcal{D}_{probe} \quad (3)$$

- **Step 2.** Generate the masked image set $D_i^{\text{mask}} = \{x'_i(x) \mid x \in \mathcal{D}_{probe}\}$ which contains k images with highlighted regions containing information relevant to describe working of neuron i .

2.1 Define the subset of top- k activating images $D_i \subset \mathcal{D}_{probe}$ for neuron i based on their mean activation value.

$$D_i = [x \mid x \in \text{top-}k_{x \in \mathcal{D}_{probe}} \text{mean}(f_i(x))] \quad (4)$$

2.2 Compute the dot product between $d_i(x)$ and input image x to obtain masked images $x'_i, \forall x \in D_i$.

$$x'_i(x) = [d_i(x) \odot x, \forall x \in D_i] \quad (5)$$

- **Step 3.** Obtain CLIP embeddings for the masked images $x'_i(x) \in D_i^{\text{mask}}$ and concepts $c \in \mathcal{C}$ using the CLIP vision and text encoders respectively:

$$e_{x'_i(x)} = \text{CLIP-IMAGE}(x'_i(x)) \quad (6)$$

$$e_c = \text{CLIP-TEXT}(c) \quad (7)$$

- **Step 4.** Compute the similarity score (as described in detail in Section 3.2) between a concept c and neuron

i , and assign neuron i the concept with the highest similarity score:

$$c_i^* = \arg \max_{c \in \mathcal{C}} \text{sim}(c, i) \quad (8)$$

3.2. Similarity function

Our approach is based on the SoftWPMI similarity function introduced in [24], which measures the similarity between a set of images and a natural language description. However, unlike the original SoftWPMI, we use a simplified version of the similarity function, as shown below.

$$\text{sim}(c, i) = \mathbb{E}_{x \in D_i^{\text{mask}}} \log(1 + p(x)(p(c | x) - 1)) \quad (9)$$

We estimate $p(x \in D_i^{\text{mask}})$ based on the rank of the image x in the set D_i^{mask} , starting with 0.998 of the most highly activating image for neuron k to 0.99 for the 10th most highly activating image. We estimate $p(c | x)$ as the softmax of the dot product between the CLIP embeddings of the masked image $x \in D_i^{\text{mask}}$ and the concept $c \in \mathcal{C}$, using a temperature parameter T .

$$p(c | x) = \frac{\exp(e_x \cdot e_c * T)}{\sum_{c' \in \mathcal{C}} \exp(e_x \cdot e_{c'} * T)} \quad (10)$$

3.3. Automated evaluation of neuron descriptions

This section describes our novel approach to automatically evaluate the quality of the generated neuron descriptions, which is important to be able to accurately and swiftly assess which neuron descriptions are trustworthy. A good description of a neuron implies the neuron should activate if and only if the concept is present in the input image. To identify whether the neuron always activates on the assigned concept, we propose creating new images using a generative text-to-image model G , and rating how highly these activate the neuron compared to images in $\mathcal{D}_{\text{probe}}$.

We score each description based on the likelihood that an image generated by G highly activates the given neuron. Using generative models has the benefit of being able to test whether the description holds on unseen images of that concept, and can help distinguish between concepts that have spurious correlations in the probing data.

We define a highly activating image as an image that is in top-1%, or top-0.1% of all activations for that neuron. This allows us to easily compare activations in different neurons and is not sensitive to changes in the activation scale. The higher this fraction, the more likely the neuron is to activate on inputs with that concept, indicating a better description. We call this metric the *Activation-Fraction*. Mathematically, the procedure goes as follows:

1. Given a concept c_i^* assigned to a neuron i , generate a set of M images S_i that represent the concept using a generative model G :

$$S_i = [G(c_i^*)]_{j=1}^M \quad (11)$$

2. Run $\mathcal{D}_{\text{combined}} = \mathcal{D}_{\text{probe}} \cup S_i$ through the model and record the activations of each (neuron, image) pair denoted as $A_i(x)$. For CNN layers we average the activations over the spatial dimensions.

$$A_i(x) = \frac{1}{|L|} \sum_{(j,k) \in L} f_i(x)_{j,k} \quad (12)$$

where f_i is the activation of channel i and L is the set of all valid spatial coordinates positions in the channel.

3. Compute threshold T_i as the top τ quantile of neuron i activations

$$T_i = \left[\max_t \mathbb{P}_{x \sim \mathcal{D}_{\text{combined}}} [A_i(x) > t] \geq \tau \right] \quad (13)$$

We use both $\tau = 0.99$ and $\tau = 0.999$ in our experiments.

4. Compute the *Activation-Fraction* for the concept c_i^* assigned to neuron i as the fraction of images in S_i that are in the τ quantile of the $\mathcal{D}_{\text{combined}}$:

$$\text{Activation-Fraction}(c_i^*) = \frac{1}{|S_i|} \sum_{x \in S_i} \mathbb{I}[A_i(x) \geq T_i] \quad (14)$$

where $\mathbb{I}$ is the indicator function.

4. Experiments

We apply and evaluate our approach on pretrained vision models from two diverse domains – ResNet50 [9], trained on ImageNet [28], and medical ResNet50 [6], which was trained on multiple chest X-ray datasets.

4.1. Qualitative and Human evaluation

4.1.1 Experiment settings

In these experiments, we dissect ResNet50 [9] trained on ImageNet-1K [28] using CLIP ViT-B/16 [27] as our vision-language model. Further, our $\mathcal{D}_{\text{probe}}$ consists of images from both the Broden dataset [2] and the ImageNet validation set [28]. For image masking, we dilate the mask using a uniform kernel of size 15 after resizing the mask to match the size of the input image.

We compile our concept set from the Amazon-Berkeley Objects dataset [7] and the MS COCO dataset’s captions [20] as described in Appendix A, resulting in 190k+ concepts overall. We set $k=10$ and restrict D_i to top 10 activating images, as we observe that this is sufficient to generate meaningful concepts. To create more robust descriptions, we employ prompt ensembling, creating the text embedding for a concept by averaging over several different prompts as described in Appendix B. Lastly, we set η in Eq 1 to 0.1, and the temperature parameter in Eq. 10 to 10.0.

4.1.2 Human evaluation of neuron descriptions

As manual inspection of neuron’s activations can provide a strong signal about what its learned representations potentially embody, we conducted an experiment on Amazon Mechanical Turk (AMT) to evaluate our method against existing approaches. In the study, we presented participants with the top 10 images that activate a neuron, highlighting their activation maps. We then provided them anonymized options corresponding to the outputs of various neuron explanation methods, and asked them to rate each one.

Specifically, for each neuron description, we asked participants to rate their agreement with the statement “The description describes most of the *Highlighted* portions in the above images well”, on a 5-point scale ranging from “1-Strongly Disagree” to “5-Strongly Agree”. Further, we asked them to pick the “best” description from the provided list of options. To handle ties and repeated descriptions, we allowed participants to pick more than one option for this question. In total, we evaluated 200 neurons across layers 1 to 4 of ResNet50, with 3 evaluators per neuron.

Our results for this evaluation are summarized in Tables 1 and 2. On average, SAND scores the highest overall across the 4 layers of ResNet50, achieving a score of 3.93 for concept quality and relevance, while being picked as the best concept 43.17% of the time. In particular, we see the greatest improvement for lower-layer neurons from layers 1 and 2, whose activation maps focus on finer-grained spatial information as compared to later layers, and thus stand to benefit the most from masking. For example, in Table 1 we see that our method scores 4.01 for concept quality for layer 1, which is significantly greater than to the second highest score of 3.67 achieved by MILAN. Similarly, in Table 2, SAND is picked as the best concept 49.33% of the time for layer 1 as compared to the second highest score of 36.67% for MILAN, and 47.33% of the time for layer 2 as compared to the second highest score of 41.33% for CLIP-dissect.

While masking helps boost visual information from the background, we see that deeper neurons typically focus on foreground objects (possibly due to the lower-resolution spatial information available to them at that stage) and thus do not seem to benefit as much from it. However, for lower-layer neurons from layers 1 and 2, the combination of highlighting relevant visual regions and using an open-set vocabulary appears to lead to promising improvements in the quality of the neuron descriptions.

4.1.3 Qualitative results

In Figure 2, we see that SAND not only helps localize and highlight visual concepts of interest while generating neuron descriptions (e.g. crutches in Layer 4, unit 1415), it also prevents confounding information from passing through by blocking out irrelevant regions for a neuron (e.g. the fore-

ground objects in Layer 2, unit 91, and Layer 1, units 126 and 88). Moreover, its ability to use open-set vocabulary often enables it to pick more descriptive and specific neuron descriptions compared to its counterparts like Net-Dissect and MILAN that also focus on a neuron’s activation region, as seen in the concepts “colorful streamers” for Layer 4, unit 1104, and “polka-dot cloth” for Layer 2, unit 398. It is also interesting to note from the activation map of Layer 3, unit 941 – which appears to be detecting large animal cages by focusing on the fur and bars of a cage – how localized spatial information can provide key signals about the visual features a model uses to understand concepts, thus highlighting the importance of integrating such information into the neuron explanation process.

4.2. Automated evaluation using image generation

4.2.1 Experiment settings

For our automated evaluation experiments, we generate 20 images per concept $c \in \mathcal{C}$, using Stable Diffusion v2-1 to generate 2 images for each of the 10 prompts listed in Appendix B. Using multiple prompts maximize the diversity of generated images, giving us a more robust measure of description quality. Our D_{probe} consists of images from Broden [2] and the ImageNet validation set [28]. The descriptions are then scored using the Activation-Fraction metric described in Section 3.3. We only evaluate the descriptions of the ImageNet trained model, as current generative models cannot produce sufficiently accurate medical images.

4.2.2 Evaluating the evaluation metric

We display qualitative example results in Figure 3. We see that both Layer 4, neuron 282, and Layer 1, neuron 0 have good descriptions, and also reach high evaluation scores. However, in some cases we notice that relevant descriptions fail to reach a high score, potentially due to two reasons – first, as seen in Layer 3, neuron 537, G might fail to generate appropriate images for generic neuron descriptions. In this case, the description lacks specific features about the tray, leading G to generate visually different images that do not activate the neuron. Secondly, as seen in neuron 323 from Layer 2, failure on the part of the image generator itself to generate patterns in the appropriate style could lead to a lower than expected score. On the other hand, we observe that poor explanations consistently receive low scores. Overall, despite some noise we believe our method could be a reliable metric to use when comparing averages between explanation methods.

4.2.3 Evaluating SAND descriptions

We present our quantitative results for ResNet50 neurons in Table 3, where we report the Activation-Fraction for 50 ran-

Table 1. Average human evaluator scores from Amazon Mechanical Turk, ranging from 1 (bad) to 5 (good), for the neuron concepts of ResNet50. SAND scores the highest overall. In particular, we see the greatest improvement for lower layer neurons, whose activation maps contain more fine-grained spatial information than later layers, and thus benefit the most from masking.

Layers	Net-Dissect [2]	MILAN [10]	CLIP-dissect [24]	SAND (Ours)
Layer 1	3.59 ± 0.082	3.67 ± 0.095	3.59 ± 0.094	4.01 ± 0.081
Layer 2	3.55 ± 0.096	3.58 ± 0.086	3.93 ± 0.077	4.00 ± 0.069
Layer 3	3.70 ± 0.088	3.69 ± 0.094	3.85 ± 0.073	3.93 ± 0.083
Layer 4	3.58 ± 0.092	3.97 ± 0.079	3.99 ± 0.074	3.79 ± 0.081
All layers	3.61 ± 0.045	3.73 ± 0.045	3.84 ± 0.040	3.93 ± 0.040

Table 2. Percentage of instances a method’s concept was picked as the ‘best’ option describing a neuron’s top activating images (can sum to greater than 100% for a layer, since we allow multiple selections for this prompt). Similar to Table 1, SAND performs the best overall and shows the largest benefit for lower layer neurons, which focus on finer visual concepts that might need to be boosted from the background.

Layers	Net-Dissect [2]	MILAN [10]	CLIP-dissect [24]	SAND (Ours)
Layer 1	22.67	36.67	30.00	49.33
Layer 2	23.33	34.67	41.33	47.33
Layer 3	34.00	30.00	46.67	38.00
Layer 4	34.67	37.33	42.00	38.00
All layers	28.67	34.67	40.00	43.17

domly selected neurons in each layer. We see that SAND performs the best overall, with especially good results in layer 1 and layer 4. Interestingly, different from our human evaluation results, descriptions for lower layers receive worse scores than those of higher layers. We believe this to be partially caused by limitations of the image generator, as it is better at generating images of object-centric concepts which are more common in higher layers. In addition, the two metrics measure slightly different things, where the human evaluation favors more generic concepts, while our automated evaluation is biased towards more specific concepts. Due to this, it is important to consider the two evaluations together for a more complete picture.

4.3. Applying SAND in the Medical Domain

4.3.1 Experiment settings

In this section, we demonstrate SAND’s applicability in the medical domain by analyzing Torchxrayvision’s ResNet50 [6] (henceforth referred to as medical ResNet50), trained on multiple chest X-ray datasets [33] [12], [5] [11]. We use MedCLIP [34], the NIH Chest X-ray dataset [33] as D_{probe} , and a concept set formed by class labels and short descriptions of their indicators from datasets [33] [21] [22] (Appendix A). Prompts are provided in Appendix B.

4.3.2 Analysing medical ResNet50

Here, we analyse how medical ResNet50 connects concepts in Layer 4 to make its final prediction in the fully con-

nected layer. Specifically, for every final layer neuron corresponding to a class label, we look at its top-5 connections by weight in the previous layer, as these neurons have the largest contribution towards predicting that corresponding class. This selection of neurons is also likely to be interpretable, making it easier to get expert feedback.

To evaluate SAND’s neuron descriptions, we ask a medical expert to rate whether the activation map of a neuron corresponds with the appropriate region for the concept assigned to it on a scale of 1 to 3, and note that SAND receives an average score of 2.18 ± 0.12 for this evaluation. Figure 4 shows an example of a description assigned by our method.

In Figure 5 we display the class-wise distribution of neuron description ratings, where we see that our neuron descriptions for some output classes are appropriate and useful, but require improvement for others. Overall, this shows promise in applying neuron-level interpretation in medical settings, and demonstrates the importance of localized information in understanding what a model is basing its decision on. It also highlights the need for improvement before neuron-level interpretation can be reliably useful for practitioners. Paths towards better explanations include training medical models that are more interpretable by design, and using stronger medical vision-language models to assign neuron descriptions.

5. Conclusion

In this work, we proposed Spatially Aware open-set Network Dissection (SAND), which improves neuron explana-

Table 3. Automatic evaluation of neuron concepts for ResNet50 [9] trained on ImageNet [28], using images generated by Stable Diffusion v2-1. These scores represent the average fraction of the generated images that are in the top-1% most highly activating inputs for that neuron, as well as standard error of the mean. SAND achieves the highest overall score, and descriptions for higher layer neurons receive higher scores than lower layer neurons.

Layers	Net-Dissect [2]	MILAN [10]	CLIP-dissect [24]	SAND (Ours)
Layer 1	0.301 ± 0.040	0.222 ± 0.035	0.264 ± 0.041	0.355 ± 0.046
Layer 2	0.236 ± 0.030	0.193 ± 0.031	0.287 ± 0.033	0.293 ± 0.035
Layer 3	0.213 ± 0.036	0.210 ± 0.034	0.296 ± 0.035	0.294 ± 0.043
Layer 4	0.370 ± 0.043	0.216 ± 0.039	0.579 ± 0.041	0.681 ± 0.043
All layers	0.280	0.210	0.357	0.406

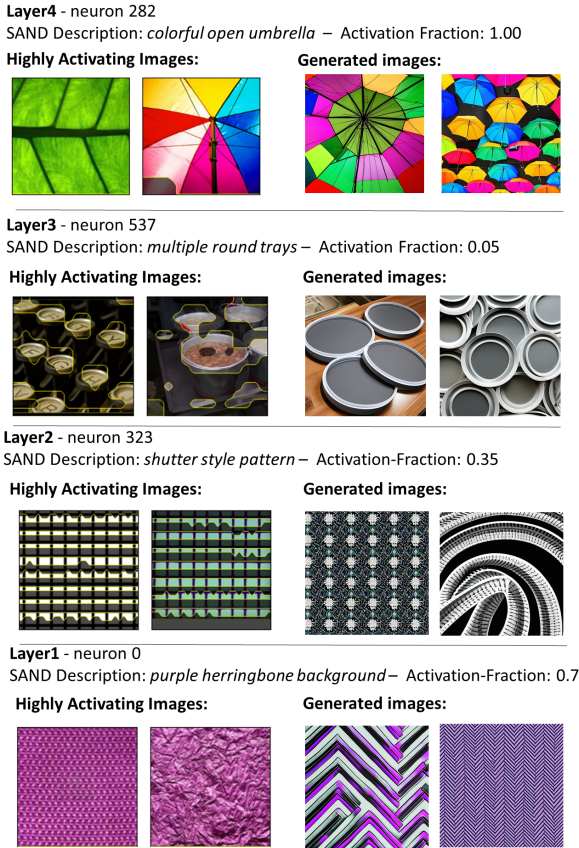


Figure 3. Example neurons in Resnet50 with their description assigned by our method along with automated evaluation scores. Additional results are present in the Appendix.

tions by explicitly focusing on the spatial region activated for a neuron. This helps us better describe neurons with highly localized concepts, such as the early layers of a CNN or the neurons of a model trained on medical X-ray images. Extensive experiments and human evaluations demonstrate that our approach assigns better descriptions than existing methods. In addition, we proposed an automated metric to

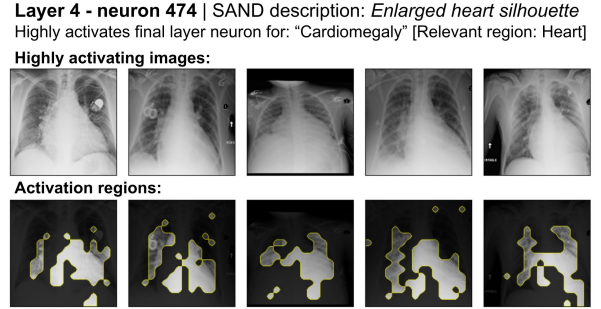


Figure 4. Example of a Layer 4 neuron concept from medical ResNet50 that matches the final layer class it highly influences. Additional examples are discussed in Appendix G.

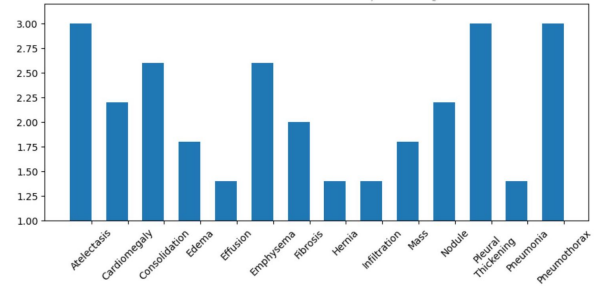


Figure 5. Class-wise average ratings of Layer 4 neuron descriptions from medical ResNet50, that measure whether the assigned descriptions corresponds with the activation map regions of the neuron, on a scale of 1-“Does not match” to 3-“Matches”.

evaluate the quality of neuron explanations, which can help accelerate the development of neuron explanation methods.

Acknowledgements

The authors thank the anonymous reviewers for valuable feedback on the manuscript. The authors are partially supported by National Science Foundation awards CCF-2107189, IIS-2313105, IIS-2430539, Hellman Fellowship, and Intel Rising Star Faculty Award.

References

- [1] Nicholas Bai, Rahul A. Iyer, Tuomas Oikarinen, and Tsui-Wei Weng. Describe-and-dissect: Interpreting neurons in vision networks with language models, 2024. [3](#)
- [2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017. [1](#), [2](#), [5](#), [6](#), [7](#), [8](#), [13](#)
- [3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020. [1](#), [2](#)
- [4] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>, 2023. [3](#)
- [5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, Dec. 2020. [7](#)
- [6] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, 2020. [5](#), [7](#)
- [7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. [5](#), [11](#)
- [8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. [11](#)
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. [5](#), [8](#), [13](#)
- [10] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2022. [1](#), [3](#), [7](#), [8](#), [13](#)
- [11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghighi, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. [7](#)
- [12] Pollard T.J. Berkowitz S.J. Greenbaum N.R. Lungren M.P. Deng C. Mark R. Horng S. Johnson, A.E.W. A de-identified publicly available database of chest radiographs with free-text reports. 2019. [7](#)
- [13] Neha Kalibhat, Shweta Bhardwaj, C. Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15623–15638. PMLR, 23–29 Jul 2023. [1](#), [3](#)
- [14] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. [3](#)
- [15] Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2584–2594, 2023. [1](#)
- [16] Laura Kopf, Philine Lou Bommer, Anna Hedström, Sebastian Lapuschkin, Marina M-C Höhne, and Kirill Bykov. Cosy: Evaluating textual explanations of neurons. *arXiv preprint arXiv:2405.20331*, 2024. [3](#)
- [17] Biagio La Rosa, Leilani Gilpin, and Roberto Capobianco. Towards a fuller understanding of neurons with clustered compositional explanations. *Advances in Neural Information Processing Systems*, 36, 2023. [3](#)
- [18] Roman Levin, Manli Shu, Eitan Borgnia, Furong Huang, Micah Goldblum, and Tom Goldstein. Where do models go wrong? parameter-space saliency maps for explainability. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15602–15615. Curran Associates, Inc., 2022. [3](#)
- [19] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. [11](#)
- [20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. [5](#), [11](#)
- [21] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations, 2020. [7](#), [11](#)
- [22] Ngoc H. Nguyen, Hieu H. Pham, Thanh T. Tran, Tuan N.M. Nguyen, and Ha Q. Nguyen. Vindr-pcxr: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *medRxiv*, 2022. [7](#), [11](#)

- [23] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023. [11](#)
- [24] Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *International Conference on Learning Representations*, 2023. [1](#), [3](#), [5](#), [7](#), [8](#), [13](#)
- [25] Tuomas Oikarinen and Tsui-Wei Weng. Linear explanations for individual neurons. In *International Conference on Machine Learning*, 2024. [3](#)
- [26] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–694, 2015. [1](#), [3](#)
- [27] 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. *CoRR*, abs/2103.00020, 2021. [2](#), [5](#)
- [28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. [5](#), [6](#), [8](#), [13](#)
- [29] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *CoRR*, abs/2005.04345, 2020. [1](#)
- [30] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. [1](#), [3](#)
- [31] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024. [3](#)
- [32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. [1](#), [3](#)
- [33] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017. [7](#), [11](#)
- [34] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. [7](#)
- [35] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, and Julian McAuley. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023. [11](#)
- [36] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23. JMLR.org*, 2023. [1](#)