Interpreting Neurons in Deep Vision Networks with Language Models

Anonymous authors
Paper under double-blind review

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 $2\times$ as likely to be selected as the best explanation for a neuron than the best baseline. Finally, we present a use case providing critical insights into land cover prediction models for sustainability applications.

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 (Erhan et al., 2009; Zhou et al., 2015; Olah et al., 2020; Goh et al., 2021), which can provide high quality description at the cost of being very labor intensive. Alternatively, Network Dissection (Bau et al., 2017) automated this labeling process by creating the pixelwise labeled dataset, *Broden*, where fixed concept set labels serve as ground truth binary masks for corresponding image pixels. The dataset was then used to match neurons to a label from the concept set based on how similar their activation patterns and the concept maps were. While earlier works, such as Network Dissection, were restricted to an annotated dataset and a predetermined concept set, CLIP-Dissect (Oikarinen & Weng, 2023) offered a solution by no longer requiring labeled concept data, but still requires a predetermined concept set as input. By utilizing OpenAI's CLIP model, CLIP-Dissect matches neurons to concepts based on their activations in response to images, allowing for a more flexible probing dataset and concept set compared to previous works.

However, these methods still share a major limitation: Concepts detected by certain neurons, especially in intermediate layers, prove to be difficult to encapsulate using the simple, often single-word descriptions provided in a fixed concept set. MILAN (Hernandez et al., 2022) sought to enhance the quality of these 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. This leads to their proposed method being more brittle and often performs poorly outside its training data.

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

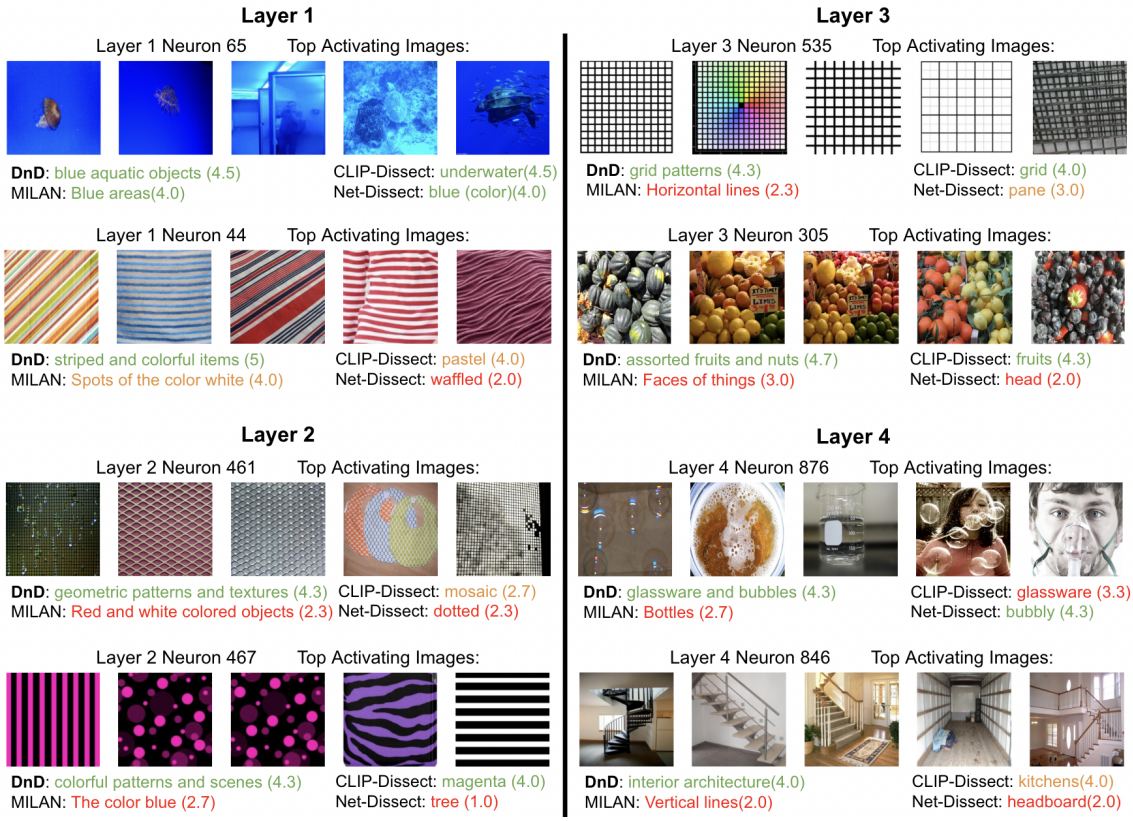


Figure 1: Neuron descriptions provided by our method (**DnD**) and baselines CLIP-Dissect (Oikarinen & Weng, 2023), MILAN (Hernandez et al., 2022)), and Network Dissection (Bau et al., 2017) for random neurons from ResNet-50 trained on ImageNet. We have added the average quality rating from our Amazon Mechanical Turk experiment described in section 4.3 next to each label and color-coded the neuron descriptions by whether we believed they were accurate, somewhat correct or vague/imprecise.

To overcome these limitations, we propose Describe-and-Dissect (abbreviated as **DnD**) in Section 3, a pipeline to *dissect* DNN by utilizing an image-to-text model to *describe* highly activating images for corresponding neurons. The descriptions are then semantically combined by a large language model, and finally refined with synthetic images to generate the final concept of a neuron. We conduct extensive qualitative and quantitative analysis in Section 4 and show that Describe-and-Dissect outperforms prior work by providing high quality neuron descriptions. Specifically, we show that Describe-and-Dissect provides more complex and higher-quality descriptions (up to 2-4× better) of intermediate layer neurons than other contemporary methods in a large scale user study. Example descriptions from our method are displayed in Figure 1. Additionally, we present a use-case study demonstrating **DnD**'s ability to interpret and improve upon current sustainability models in Section 5.

2 Background and related work

2.1 Neuron Interpretability Methods

Network Dissection (Bau et al., 2017) is the first method developed to automatically describe individual neurons' functionalities. The authors first defined the densely-annotated dataset *Broden*, denoted as $\mathcal{D}_{\text{Broden}}$, as a ground-truth concept mask. The dataset is composed of various images x_i, each labeled with concepts c at the pixel-level. This forms a ground truth binary mask $L_c(x_i)$ which is used to calculate the intersection over union (IoU) score between $L_c(x_i)$ and the binary mask from the activations of the neuron k over all

Table 1: Comparison of existing automated neuron labeling methods and our Describe-and-Dissect (**DnD**). Green and boldfaced **Yes** or **No** indicates the **desired property** for a column. **DnD** has all the **desired properties** while existing work has some limitations.

Method \ property	Requires Concept Annotations	Training Free	Generative Natural Language Descriptions	Uses Spatial Activation Information	Can easily leverage better future models
Network Dissection (Bau et al., 2017)	Yes	**Yes**	No	**Yes**	No
MILAN (Hernandez et al., 2022)	Training only	No	**Yes**	**Yes**	No
CLIP-Dissect (Oikarinen & Weng, 2023)	**No**	**Yes**	No	No	**Yes**
FALCON (Kalibhat et al., 2023)	**No**	**Yes**	No	**Yes**	**Yes**
DnD (This work)	**No**	**Yes**	**Yes**	**Yes**	**Yes**

images $x_i \in \mathcal{D}_{\text{Broden}}$, denoted $M_k(x_i)$: $\text{IoU}_{k,c} = \frac{\sum_{x_i \in \mathcal{D}_{\text{Broden}}} M_k(x_i) \cap L_c(x_i)}{\sum_{x_i \in \mathcal{D}_{\text{Broden}}} M_k(x_i) \cup L_c(x_i)}$. The concept c is assigned to a neuron k if $\text{IoU}_{k,c} > \eta$, where the threshold η was set to 0.04. Intuitively, this method finds the labeled concept whose presence in the image is most closely correlated with the neuron having high activation. Extensions of Network Dissection were proposed by (Bau et al., 2020) and (Mu & Andreas, 2020).

However, Network Dissection is limited by the need of concept annotation and the concept set is a closed set that may be hard to expand. To address these limitations, a recent work CLIP-Dissect (Oikarinen & Weng, 2023) utilizes OpenAI's multimodal CLIP (Radford et al., 2021) model to describe neurons automatically without requiring annotated concept data. They leverage CLIP to score how similar each image in the probing dataset $\mathcal{D}_{probe}$ is to the concepts in a user-specified concept set to generate a concept activation matrix. To describe a neuron, they compare the activation pattern of said neuron to activations of different concepts on the probing data, and find the concept that is the closest match using a similarity function, such as softWPMI. Another very recent work FALCON (Kalibhat et al., 2023) uses a method similar to CLIP-Dissect but augments it via counterfactual images by finding inputs similar to highly activating images with low activation for the target neuron, and utilizing spatial information of activations via cropping. However, they solely rely on cropping the most salient regions within a probing image to filter spurious concepts that are loosely related to the ground truth functionality labels of neurons. This approach largely restrict their method to local concepts while overlooking holistic concepts within images, as also noted in (Kalibhat et al., 2023). Their approach is also limited to single word / set of words description that is unable to reach the complexity of natural language.

MILAN (Hernandez et al., 2022) is a different approach to describe neurons using natural language descriptions in a generative fashion. Note that despite the concept sets in CLIP-Dissect and FALCON being flexible and open, they cannot provide generative natural language descriptions like MILAN. The central idea of MILAN is to train an images-to-text model from scratch to describe the neuron's role based on 15 most highly activating images. Specifically, it was trained on crowdsourced descriptions for 20,000 neurons from selected networks. MILAN can then generate natural language descriptions to new neurons by outputting descriptions that maximize the weighted pointwise mutual information (WPMI) between the description and the active image regions. One major limitation of MILAN is that the method require training a model to imitate human descriptions of image regions on relatively small training dataset, which may cause inconsistency and poor explanations further from training data. In contrast, our **DnD** is *training-free*, *generative*, and produces a *higher quality of neuron descriptions* as supported by our extensive experiments in Figure 1, Table 3, and Table 4. A detailed comparison between our method and the baseline methods is shown in Table 1.

2.2 Leveraging Large Pretrained models

In our **DnD** pipeline, we are able to leverage recent advances in the large pre-trained models to provide high quality and generative neuron descriptions for DNNs in a *training-free* manner. Below we briefly introduce the Image-to-Text Model, Large Language Models and Text-to-Image Model used in our pipeline implementation. The first model is Bootstrapping Language-Image Pretraining (BLIP) (Li et al., 2022),

which is an image-to-text model for vision-language tasks that generates synthetic captions and filters noisy ones, employing bootstrapping for the captions to utilize noisy web data. While our method can use any image-to-text model, we use BLIP in this paper for our step 2 in the pipeline due to BLIP's high performance, speed, and relatively low computational cost. However, we note that our method can be easily adapted to leverage more advanced models in the future.

The second model is GPT-3.5 Turbo, which is a transformer model developed by OpenAI for understanding and generating natural language. It provides increased performance from other contemporary models due to its vast training dataset and immense network size. We utilize GPT-3.5 Turbo for natural language processing and semantic summarization in the step 2 of our **DnD**. We use GPT-3.5 Turbo in this work as it's one of the SOTAs in LLMs and cheap to use, but our method is compatible with other future and more advanced LLMs. We provide a quantitative comparison between GPT-3.5 Turbo, GPT-4.0, and LLaMA2 effect on **DnD**'s label quality in Appendix A.4.4 as well as evaluation on cost and usage limitations for each model.

The third model is Stable Diffusion (Rombach et al., 2022), which is a text-to-image latent diffusion model (LDM) trained on a subset from the LAION-5B database (Schuhmann et al., 2022). By performing the diffusion process over the low dimensional latent space, Stable Diffusion is significantly more computationally efficient than other diffusion models, such as DALLE (Ramesh et al., 2021). Due to its open availability, lower computational cost, and high performance, we employ Stable Diffusion for our image generation needs in the step 3 of **DnD**.

3 Describe-and-Dissect: Methods

Overview. In this section, 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:

- **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 and subsequently summarize them into candidate concepts using GPT;
- **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 with a scoring function.

An overview of Describe-and-Dissect (**DnD**) and these 3 steps are illustrated in Fig. 2.

3.1 Step 1: Probing Set Augmentations

Probing dataset $\mathcal{D}_{probe}$ is the set of images we record neuron activations on before generating a description. As described in Section 2.1, one major limitation of (Kalibhat et al., 2023) is the restriction to local concepts while overlooking holistic concepts within images, while one limitation of (Oikarinen & Weng, 2023) is not incorporating the spatial activation information. Motivated by these limitations, **DnD** resolves these problems by *augmenting* the original probing dataset with a set of attention crops of the highest activating images from the original probing dataset. The attention crops can capture the spatial information of the activations and we name this set as $\mathcal{D}_{cropped}$, shown in Fig. 2. We discuss the implementation details of our attention cropping procedure in Appendix A.1.1 and an perform an ablation study of its effects in Appendix A.4.1.

3.2 Step 2: Candidate Concept Generation

The top K most highly activating images for a neuron n are collected in set $I, |I| = K$, by selecting K images $x_i \in \mathcal{D}_{probe} \cup \mathcal{D}_{cropped}$ with the largest $g(A_k(x_i))$. Here g is a summary function (for the purposes of our experiments we define g as the spatial mean) and $A_k(x_i)$ is the activation map of neuron k on input x_i. We then generate a set of candidate concepts for the neuron with the following two part process:

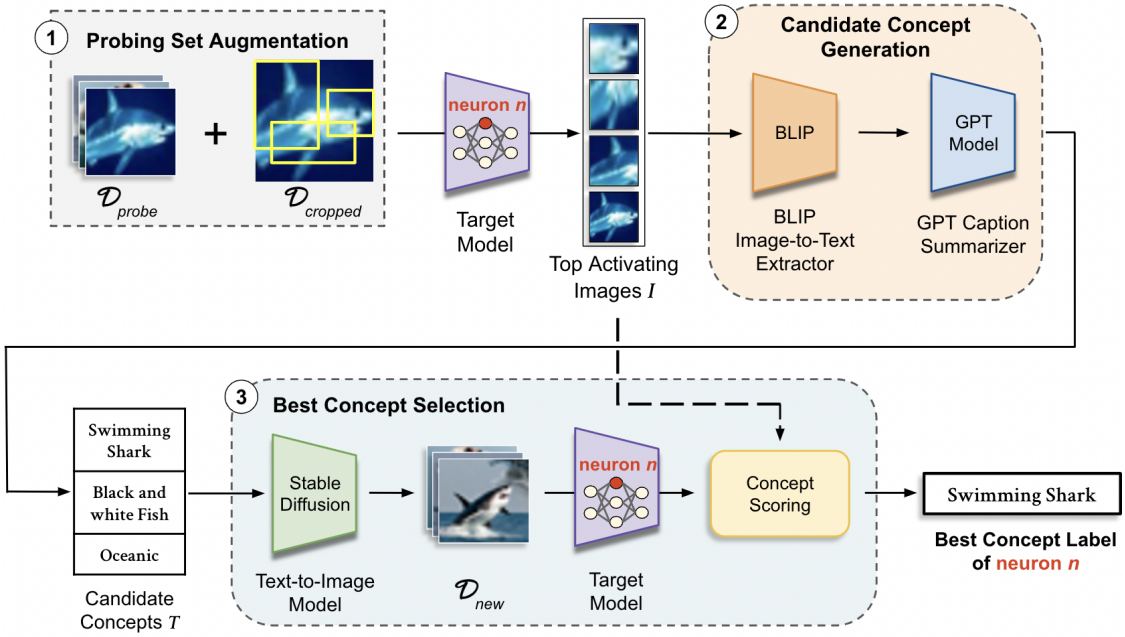


Figure 2: Overview of Describe-and-Dissect (**DnD**) algorithm. Given a Target model, it consists three important steps to identify the neuron concepts (e.g. 'Swimming Shark' for neuron n).

- **Step 2A - Generate descriptions for highly activating images:** We utilize BLIP image-to-text model to generatively produce an image caption for each image in I. For an image $I_{j\in[K]}$, we feed I_j into the base BLIP model to obtain an image caption.
- **Step 2B - Summarize similarities between image descriptions:** Next we utilize OpenAI's GPT-3.5 Turbo model to summarize similarities between the K image captions for each neuron being checked. GPT is prompted to generate N descriptions which identify and summarize the conceptual similarities between most of the BLIP-generated captions.

The output of **Step 2B** is a set of N descriptions which we call "candidate concepts". We denote this set as $T = \{T_1, ..., T_N\}$. For the purposes of our experiments, we generate $N = 5$ candidate concepts unless otherwise mentioned. The exact prompt used for GPT summarization is shown in Appendix A.1.2.

3.3 Step 3: Best Concept Selection

The last crucial component of **DnD** is concept selection, which selects the concept from the set of candidate concepts T that is most correlated to the activating images of a neuron. We first use the Stable Diffusion model (Rombach et al., 2022) from Hugging Face to generate images for each concept $T_{j\in[N]}$. Generating new images is important as it allows us to differentiate between neurons truly detecting a concept or just spurious correlations in the probing data. The resulting set of images is then fed through the target model again to record the activations of a target neuron on the new images. Finally, the candidate concepts are ranked using a concept scoring model, as discussed in section 3.4.

Concept Selection Algorithm The algorithm consists of 4 substeps. For each neuron n, we start by:

1. *Generate supplementary images.* Generate Q synthetic images using a text-to-image model for each label $T_{j\in[N]}$. The set of images from each concept is denoted as $\mathcal{D}_j$, $|\mathcal{D}_j| = Q$. The total new dataset is then $\mathcal{D}_{new} = \bigcup_{j=1}^{N} \mathcal{D}_j = \{x_1^{new}, ..., x_{N\cdot Q}^{new}\}$, which represents the full set of generated images. For the purposes of the experiments in this paper, we set $Q = 10$.

2. *Feed new dataset $\mathcal{D}_{new}$, back into the target model and rank the images based on activation.* We then evaluate the activations of target neuron n on images in $\mathcal{D}_{new}$ and compute the rank of each image in terms of target neuron activation. Given neuron activations $A_n(x_i^{new})$, we define $\mathcal{G}_n = \{g(A_n(x_1^{new})), ..., g(A_n(x_{N.Q}^{new}))\}$ as the set of scalar neuron activations.

3. *Gather the ranks of images corresponding to concept T_j.* Let $\text{Rank}(x; \mathcal{G})$ be a function that returns the rank of an element x in set $\mathcal{G}$, such that $\text{Rank}(x'; \mathcal{G}) = 1$ if x' is the largest element in $\mathcal{G}$. For every concept T_j, we record the ranks of images generated from the concept in $\mathcal{H}_j$, where $\mathcal{H}_j = \{\text{Rank}(g(A_n(x)); \mathcal{G}_n) \; \forall \; x \in \mathcal{D}_j\}$, and $\mathcal{H}_j$ is sorted in increasing order, so $\mathcal{H}_{j1}$ is the rank of the lowest ranking element.

4. *Assign scores to each concept.* The scoring function $score(\mathcal{H}_j)$ assigns a score to a concept using the rankings of the concept's generated images, and potential additional information. The concept with the best (highest) score in T is selected as the concept label for the neuron. Concept scoring functions are discussed below in Section 3.4.

In simpler terms, the intuition behind this algorithm is that if a neuron n encodes for a concept c, then the images generated to encapsulate that concept c should cause the neuron n to activate highly. While we only experiment with Best Concept selection within the **DnD** framework, it can be independently applied with other methods like (Bau et al., 2017; Hernandez et al., 2022; Oikarinen & Weng, 2023) to select the best concept out of their top-k best descriptions, which is another benefit of our proposed method. **DnD** can also be used without Best Concept selection to reduce computational costs.

3.4 Scoring Function

For a given neuron, we use a scoring function to rate candidate concept accuracy during Best Concept Selection (step 3). 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 (Radford et al., 2021), 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, $R_j = \{(R_j^i)^2; i \leq \beta\}$. Sections A.1.3 the details specifics behind the function. In section A.1.4, we compare between various functions and show our algorithm works robustly with different options.

4 Experiments

In this section, we present extensive qualitative and quantitative analysis to show that **DnD** outperforms prior works by providing higher quality neuron descriptions. For fair comparison, we follow the setup in prior works to run our algorithm on the following two networks: ResNet-50 and ResNet-18 (He et al., 2016) trained on ImageNet (Russakovsky et al., 2015) and Place365 (Zhou et al., 2016) respectively. In section 4.1, we qualitatively analyze **DnD** along with other methods on random neurons and show that our method provides good descriptions on these examples. Next in section 4.2 we quantitatively show that **DnD** yields superior results to comparable methods. In section 4.3, we show that our method outperforms existing neuron description methods in large scale crowdsourced studies. Finally in section 4.4 we study the importance of critical steps in our pipeline by ablating away Generative Image Captioning (step 2A) and Concept Selection (step 3). Supplementary results are presented in the appendix, including method details in section A.1, additional qualitative examples in section A.2, extensive ablation studies on each step of the **DnD** framework in section A.4, an additional use case of **DnD** as an OOD classifier in section A.5, and the capability to describe polysemantic neurons by producing multiple labels in section A.6.

4.1 Qualitative evaluation

We qualitatively analyze results of randomly selected neurons from various layers of ResNet-50, ResNet-18, and ViT-B-16. Sample results are displayed in Figure 1 and Figures 8, 9, 10, 11, 12, 13, and 14 in the Appendix. We use the union of the ImageNet validation dataset and Broden as $\mathcal{D}_{probe}$ and compare to Network Dissection (Bau et al., 2017), MILAN (Hernandez et al., 2022), and CLIP-dissect (Oikarinen & Weng, 2023) as baselines. Labels for each method are color coded by whether we believe they are accurate, somewhat correct, or vague/imprecise. Compared to baseline models, we observe that **DnD** captures higher level concepts in a more semantically coherent manner. Specifically, methods such as CLIP-dissect and Network Dissection have limited expressability due to the use of restricted concept sets while MILAN produces labels confined to lower level concepts. Additionally, we find that **DnD** can express multiple concepts within in a single label owing to its generative nature.

4.2 Quantitative evaluation

4.2.1 Final layer evaluation

Here we follow (Oikarinen & Weng, 2023) to quantitatively analyze description quality on the last layer neurons, which have known ground truth labels (i.e. class name) to allow us to evaluate the quality of neuron descriptions automatically. In this evaluation, we focus on comparison with MILAN (Hernandez et al., 2022), as it is the other generative contemporary work in the baselines. Network Dissection (Bau et al., 2017) and CLIP-Dissect (Oikarinen & Weng, 2023) are not included in this comparison because these methods have concept sets where the "ground truth" class or other similar concepts can be included, giving them an unfair advantage to the methods without concept sets like MILAN and **DnD**. We reported the results for all of the neurons of ResNet-50's final fully-connected layer in Table 2. Our results show that **DnD** outperforms MILAN, producing labels are significantly closer to the ground truths than MILAN's.

Table 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.

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

4.3 Crowdsourced experiment

Table 3: **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%**

Table 4: **Averaged AMT results across layers in ResNet-18**. We can see **DnD** outperforms existing methods on ResNet-18 trained on Places365. Our model was selected the best out of the three methods for more than 54% of time time, almost 3× as often as the second best method.

Metric / Methods	NetDissect	MILAN	CLIP-Dissect	**DnD (Ours)**
Mean Rating	3.33	3.14	3.52	**4.14**
selected as best	12.62	13.32%	19.39%	**54.67%**

Setup. Our experiment compares the quality of labels produced by **DnD** against 3 baselines: CLIP-Dissect, MILAN, and Network Dissection. For MILAN we used their most powerful *base* model in our experiments.

We dissected both a ResNet-50 network pretrained on Imagenet-1K and ResNet-18 trained on Places365, using the union of ImageNet validation dataset and Broden (Bau et al., 2017) as our probing dataset. For both models we evaluated 4 of the intermediate layers (end of each residual block), with 200 randomly chosen neurons per layer for ResNet50 and 50 per layer for ResNet-18. Each neurons description was evaluated by 3 different workers. In total, 3000 human ratings were conducted, 2400 evaluations on ResNet-50 and 600 evaluations on ResNet-18.

The full task interface and additional experiment details are available in Appendix A.1.5. Workers were presented with the top 10 highest activating images of a neuron followed by four separate descriptions; each description corresponds to a label produced by one of the four methods compared. The descriptions are rated on a 1-5 scale, where a rating of 1 represents that the user "strongly disagrees" with the given description, and a rating of 5 represents that the user "strongly agrees" with the given description. Additionally, we ask workers to select the description that best represents the 10 highly activating images presented. For these highly activating images, we used the images calculated by our method. As our probing dataset is a superset of the image sets used by prior methods, we believe our model is the most accurate for determining images to visualize since the probing dataset encapsulates the most concepts.

Results. Table 3 and Table 4 shows the results of a large scale human evaluation study conducted on Amazon Mechanical Turk (AMT). Looking at "% time selected as best" as the comparison metric, our results show that **DnD** performs over 2× better than all baseline methods when dissecting ResNet-50 or ResNet-18, being selected the best of the four up to 54.67% of the time. In terms of mean rating, our method achieves an average label rating over 4.1 for both dissected models, whereas the average rating for the second best method, CLIP-Dissect, is only 3.67 on ResNet-50 and 3.52 on ResNet-18. Our method also significantly outperforms MILAN's *generative* labels, which averaged below 3.3 for both target models. In conclusion we have shown that our method significantly outperforms existing methods in crowdsourced evaluation, and does this consistently across different models and layers.

4.3.1 MILANNOTATIONS evaluation

Though evaluation on hidden layers of deep vision networks can prove quite challenging as they lack "ground truth" labels, one resource to perform such task is the MILANNOTATIONS dataset (Hernandez et al., 2022), which collects annotated labels to serve as ground truth neuron explanations. We perform quantitative evaluation by calculating the textual similarity between a method's label and the corresponding MILANNOTATIONS. Our analysis in section A.3 found that if every neuron is described with the same constant concept 'depictions', it will achieve better results than any explanation on the dataset, but this is not a useful nor meaningful description. We hypothesize this is due to high levels on noise and interannotator disagreement, leading to low textual similarity between descriptions and generic descriptions scoring highly. We conclude that this dataset is unreliable to serve as ground truths for comparing different methods.

4.4 Ablation Studies

4.4.1 DnD with fixed concept set

To analyze the importance of using a generative image-to-text model, we explore instead utilizing fixed concept sets with CLIP (Radford et al., 2021) to generate descriptions for each image instead of BLIP, while the rest of the pipeline is kept the same (i.e. using GPT to summarize etc). For the experiment, we use CLIP-ViT-B/16, where we define $L(\cdot)$ and $E(\cdot)$ as text and image encoders respectively. From the initial concept set $\mathcal{S} = \{t_1, t_2, ...\}$, the best concept for image I_m is defined as t_l, where $l = \text{argmax}_i(L(t_i) \cdot E(I_m)^\top)$. Following CLIP-dissect (Oikarinen & Weng, 2023), we use $\mathcal{S} = 20\text{k}^2(20,000$ most common English words) and $\mathcal{D}_{probe} = \text{ImageNet} \cup \text{Broden}$.

[2]Source: https://github.com/first20hours/google-10000-english/blob/master/20k.txt

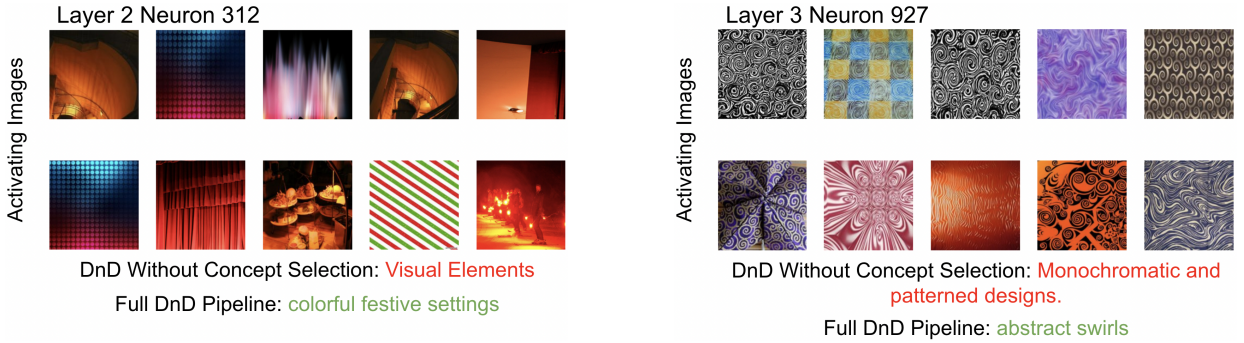
To compare the performance, following (Oikarinen & Weng, 2023), we use our model to describe the final layer neurons of ResNet-50 (where we know their ground truth role) and compare descriptions similarity to the class name that neuron is detecting, as discussed in Section 4.2.1. Results in Table 5 show that both methods perform similarly on the FC layer. In intermediate layers, we notice that single word concept captions from 20k significantly limit the expressiveness of **DnD**, suggesting having generative image descriptions is important for our overall performance. Qualitative examples and notable failure cases of CLIP descriptions can be found under Appendix A.4.2.

Table 5: **Mean FC Layer Similarity of CLIP Captioning.** Utilizing a fixed concept set to caption activating images via CLIP (Radford et al., 2021), we compute the mean cosine similarity across fully connected layers of RN50. We find the performance of **DnD** w/ CLIP Captioning is slightly worse than BLIP generative caption.

Metric / Methods	**DnD (Ours)**	DnD w/ CLIP Captioning	% Decline
CLIP cos	**0.7598**	0.7583	0.197%
mpnet cos	**0.4588**	0.4465	2.681%
BERTScore	**0.8286**	0.8262	0.290%

4.4.2 Effects of Concept Selection

We use 50 randomly chosen neurons from each of the 4 layers of ResNet-50 to conducted an ablation study on the impact of Best Concept Selection (step 3) on the pipeline. Each neuron was evaluated twice yielding a total of 400 human ratings. Table 6 shows the effect of Best Concept Selection on the overall accuracy of **DnD**. We can see DnD performance is already high without Best Concept Selection, but Concept Selection further improves the quality of selected labels in Layer 2 through Layer 4, while having the same performance on Layer 1. One potential explanation is due to Layer 1 detecting more limited lower level concepts – there is less variance in candidate descriptions identified in Concept Generation (step 2), resulting in similar ratings across the set of candidate concepts T. We can see some individual examples of the improvement Concept Selection provides in Figure 3, with the new labels yielding more specific and accurate descriptions of the neuron. For example Layer 2 Neuron 312 becomes more specific *colorful festive settings* instead of generic *Visual Elements*.



(a) Layer 2 Neuron 312

(b) Layer 3 Neuron 927

Figure 3: **Concept Selection (Step 3) supplements Concept Generation (Step 2) accuracy**. We show that concept selection improves Concept Generation by validating candidate concepts.

5 Use Case: Land-Cover Prediction

One important role of interpretability tools is the ability to create real-world impacts. In this section, we study applications in sustainability and climate change by applying **DnD** to the task of land cover prediction–a

Table 6: Human evaluation results for **DnD** (w/o Best Concept Selection) versus full Describe-and-Dissect. Full pipeline improves or maintains performance on every layer in ResNet-50.

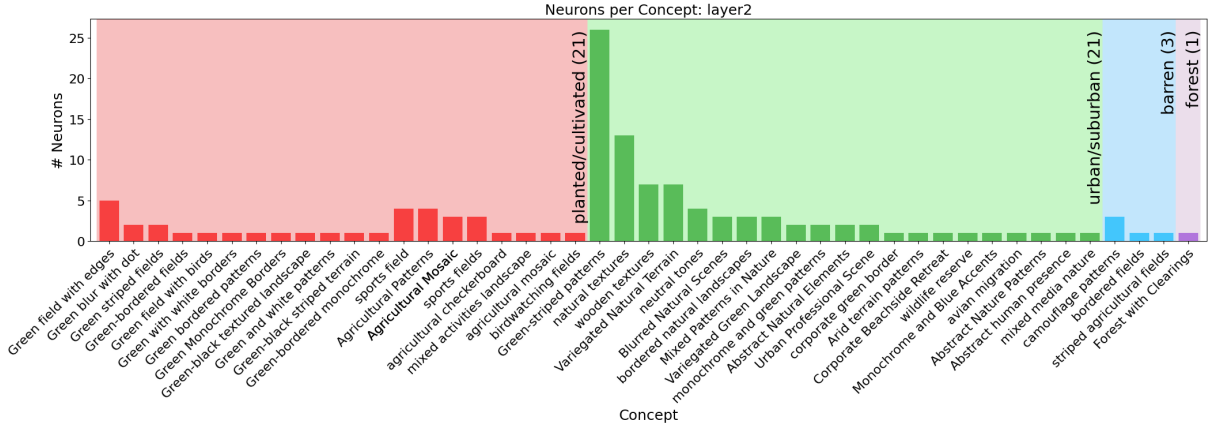
Method / Layer	Layer 1	Layer 2	Layer 3	Layer 4	**All Layers**
DnD (w/o Best Concept Selection)	**3.54**	3.77	4.00	4.02	3.84
DnD (full pipeline)	**3.54**	**4.00**	**4.24**	**4.13**	**3.97**

critical factor for managing water resources, conserving biodiversity, planning sustainable urban development, and mitigating climate change effects. **DnD**'s concept descriptions for neurons not only improve the model performance, but also identify spurious correlations, suggesting a critical role of interpretability techniques in ensuring reliability and building trust in AI systems.

Setup. We evaluate our framework on two classification models. **Tile2Vec** (Jean et al., 2019) utilizes a modified ResNet-18 backbone trained to minimize triplet loss between anchor, neighbor, and distant land tiles from the NAIP dataset (Claire Boryan & Craig, 2011). Following (Jean et al., 2019), we train a random forest classifier on ResNet-18 embeddings to evaluate accuracy on 27 subclasses of Cropland Data Layer (CDL) labels. We also evaluate a ResNet-50 model trained on labeled EuroSAT images (Helber et al., 2019) with 10 land cover classes. We experiment on two probing datasets. **NAIP** is an aerial imagery dataset updated annually by the USDA. The dataset is composed of crop cover data with annotated ground truth masks from CDL labels. However, due to the fine-grained nature of the 27 subclasses, we also categorize each subclass into six broad superclasses to improve our understanding of the results: **1.** Planted/Cultivated, **2.** Herbaceous/Shrubland, **3.** Urban/Suburban, **4.** Barren, **5.** Forest, **6.** Water/wetlands. We use NAIP for all experiments on the Tile2Vec ResNet-18 model. On EuroSAT ResNet-50, we use the union of EuroSAT ∪ ImageNet ∪ Broden datasets as the probing dataset.

5.1 Locating Conceptual Groupings

Figure 4: **Layer 2 Concept Profile.** We cluster neurons with similar concepts and categorize them into 6 NAIP superclasses. Interpretable concepts have more neurons associated with them. Some superclasses do not appear due to dataset bias or intrinsic similarities between classes.



We identify neuron clusters detecting similar concepts across layer 2 of Tile2Vec ResNet-18. For a pair of candidate concepts sets for neurons n and m, we define their textual similarity $\mathcal{S}_{n,m}$ as the spatial mean of $L(T_n) \cdot L(T_m)^\top$ where $L(\cdot)$ is the CLIP-ViT-B/16 text encoder. A set of neurons is considered similar if $\mathcal{S}_{n,m} \geq \phi = 0.8$, where ϕ controls the minimum similarity threshold between concepts. For practicality, we confine each neuron to exactly one group of highly similar neurons. Low-level concepts are then classified by GPT one-shot classification into the 6 NAIP superclasses.

Fig. 4 presents concepts from layer 2, color-coded by superclass. We find that clusters (each bar in Fig. 4) containing more neurons are frequently associated with more interpretable features, while clusters with relatively less neurons are related to vague or irrelevant concepts. Based on these insights, we prune neurons

with uninterpretable concepts in Sec. 5.2. We also note that "Herbaceous/Shrubland" and "Water/wetlands" are not associated with concepts in layer 2. We hypothesize that this is likely due to the intrinsic similarity between the classes and bias in the NAIP dataset. Concepts related to "Herbaceous/Shrubland" are closely related to the "Planted/Cultivated" superclass while "Water/wetlands" images only comprises of ∼0.275% of the dataset.

5.2 Pruning Uninterpretable Neurons

We conduct a study to identify neurons within Tile2Vec ResNet-18 model that contribute minimally to model accuracy. Based on Sec. 5.1, we locate a subset of ungrouped neurons which correlate poorly to other neurons in the network. For our experiment, we define poorly correlating subsets as concepts that activate on only a single neuron.

Table 7: Pruning uninterpretable neurons in Tile2Vec ResNet18.

Layer	% of Neurons Pruned	Avg. Acc. (%)
No pruning	0.00	71.63
All pruned	100.00	35.96
Layer 1	23.44	71.07
Layer 2	50.00	71.54
Layer 3	26.95	71.75
Layer 4	58.98	72.04
Layer 5	56.45	71.76

Table 7 shows model accuracy after identifying and pruning poorly correlating neuron subsets across each layer of the model. Due to high variance in the NAIP dataset, we evaluate the baseline accuracy for a fully pruned network. In this case, we find the prediction is always the "tomatoes" subclass which achieves an accuracy of 35.96% since this subclass comprises of ∼35.2% of the data. Our results show that a significant proportion of neurons in the model do not contribute to the overall classification. Particularly in layers 4 and 5, we are able to prune over 50% of neurons in each layer while achieving a better result than the baseline (no neurons pruned). Across the entire network, the relative small difference in accuracy after pruning suggests human interpretable neurons account for more critical roles within image classification networks compared to uninterpretable neurons.

5.3 Characterizing Spurious Correlations

We characterize spurious correlations in the intermediate layers of EuroSAT-trained ResNet-50 by determining common neuron labels through Term Frequency Analysis and studying their relationship to the task. We prune these neurons to further understand the class-wise correlation of spurious concepts (Table 8). Though "fishing" is the most prevalent concept in layer 4 (41.65% of all neurons), pruning them has no impact on model accuracy. In other words, these fishing neurons are irrelevant to the classification task. "Pink" and "purple" account for 29.74% of layer 4 neurons and have a much greater impact. These concepts, which are seemingly unrelated to the task, are spuriously correlated to the Forest, Herbaceous Vegetation, Industrial, Pasture, Residential, and Sea-Lake classes. However, the Annual Crop, Highway, and River classes have weaker correlations with these concepts.

Table 8: Pruning concepts from layer 4 of EuroSAT-trained ResNet-50.

Concepts pruned	% of neurons pruned	Class-wise Accuracy (%)										Avg. Acc. (%)
		Ann.-Crop	Forest	Herb-Veget.	Highway	Industrial	Pasture	Perm.-Crop	Residential	River	Sea-Lake	
No pruning	0	95.00	98.00	93.67	96.00	96.80	91.50	90.80	98.67	92.80	97.33	95.26
Fishing	41.65	95.00	98.00	93.67	96.00	96.80	91.50	90.80	98.67	92.80	97.33	95.26
Pink/purple	29.74	88.33	0.00	0.00	62.00	0.00	0.00	18.80	0.00	76.00	0.00	24.33

6 Conclusions

In this paper, we presented Describe-and-Dissect (**DnD**), a novel method for automatically labeling the functionality of deep vision neurons without the need for labeled training data or a provided concept set. We accomplish this through three important steps including probing set augmentation, candidate concept generation through off-the-shelf general purpose models, and best concept selection with carefully designed scoring functions. Through extensive qualitative, quantitative, and use-case analysis, we show that **DnD** outperforms prior work by providing higher-quality neuron descriptions, greater generality and flexibility, and significant potential for social impact.

References

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

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.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Rick Mueller Claire Boryan, Zhengwei Yang and Mike Craig. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

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

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *ICLR*, 2022.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2Vec: Unsupervised representation learning for spatially distributed data. In *AAAI*, 2019.

Neha Kalibhat, Shweta Bhardwaj, Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-Dissect: Automatic description of neuron representations in deep vision networks. In *ICLR*, 2023.

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.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

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. In *ICML*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. URL `https://arxiv.org/abs/2102.12092`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

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.

Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. In *ICLR*, 2015.

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