CAUSAL IMAGENET: HOW TO DISCOVER SPURIOUS FEATURES IN DEEP LEARNING?

Anonymous authors
Paper under double-blind review

(a) **class:** band aid, **spurious feature**: fingers, **-41.54%**

(b) **class:** plate, **spurious feature**: food, **-32.31%**

(c) **class:** space bar, **spurious feature**: keys, **-46.15%**

(d) **class:** butterfly, **spurious feature**: flowers, **-21.54%**

(e) **class:** potter's wheel, **spurious feature**: vase, **-21.54%**

Figure 1: Examples of spurious features discovered using our framework for a standard Resnet-50 model. In the top row, red color highlights regions with spurious features. The bottom row shows images generated by visually amplifying these features. Adding small amount of gaussian noise (with $\sigma = 0.25$) to spurious regions (red regions) significantly reduces the model accuracy for that class (shown in dark red in captions). More visualizations are shown in Appendix G.

ABSTRACT

A key reason for the lack of reliability of deep neural networks in the real world is their heavy reliance on *spurious* input features that are causally unrelated to the true label. Focusing on image classifications, we define causal attributes as the set of visual features that are always a part of the object while spurious attributes are the ones that are likely to *co-occur* with the object but not a part of it (e.g., attribute "fingers" for class "band aid"). Traditional methods for discovering spurious features either require extensive human annotations (thus, not scalable), or are useful on specific models. In this work, we introduce a *scalable* framework to discover a subset of spurious and causal visual attributes used in inferences of a general model and localize them on a large number of images with minimal human supervision. Our methodology is based on this key idea: to identify spurious or causal *visual attributes* used in model predictions, we identify spurious or causal *neural features* (penultimate layer neurons of a robust model) via limited human supervision (e.g., using top 5 activating images per feature). We then show that these neural feature annotations *generalize* extremely well to many more images *without* any human supervision. We use the activation maps for these neural features as the soft masks to highlight spurious or causal visual attributes. Using this methodology, we introduce the *Causal Imagenet* dataset containing causal and spurious masks for a large set of samples from Imagenet. We assess the performance of several popular Imagenet models and show that they rely heavily on various spurious features in their predictions. We hope this study paves the way to develop improved deep models that rely mainly on causal features in their inferences.

1 INTRODUCTION

The growing use of deep learning in sensitive applications such as medicine, autonomous driving, law enforcement and finance raises concerns about the trustworthiness and reliability of these systems in the real world. A root cause for the lack of reliability of deep models is their reliance on spurious input features (i.e., features that are causally unrelated to the true label) in their inferences. For example, in Arjovsky et al. (2020), authors trained a convolutional neural network (CNN) to classify camels from cows. In the training dataset, most of the images of cows had green pastures (Beery et al., 2018) while most camels were in deserts. The CNN learned to associate green pastures with cows and failed to classify pictures of cows in sandy beaches correctly. Similarly, Bissoto et al. (2020) discovered that skin-lesion detectors use spurious visual artifacts for making predictions. The list of such examples goes on (de Haan et al., 2019; Singla et al., 2021; Wong et al., 2021).

Most of the prior work on discovering spurious features in image classification (Nushi et al., 2018; Zhang et al., 2018; Chung et al., 2019) requires humans to first come up with a possible list of spurious features. This is often followed by an expensive human-guided labeling of visual attributes which is not scalable for datasets with a large number of classes and samples such as ImageNet. In some cases, this may even require the construction of new datasets with unusual visual attributes to validate the hypothesis (such as the cows in sandy beaches example discussed previously). To address these limitations, some recent works use the neurons of robust models as visual attribute detectors thereby circumventing the need for manual annotations. However, their applicability is either limited to robust models which are less widely deployed due to their low clean accuracy (Wong et al., 2021) or to identifying background spurious features (Xiao et al., 2021), ignoring foreground ones such as food in Figure 1 or to models that achieve low accuracy on the training dataset (Singla et al., 2021). We discuss limitations of these methods in more detail in Section 2.

In this work, we introduce a *scalable* methodology to discover a subset of spurious and causal visual attributes used in model inferences and localize them on a large number of images with minimal human supervision. Our work builds upon the prior observations (Engstrom et al., 2019) that for a robust model, neurons in the penultimate layer (called *neural features*) often correspond to human-interpretable visual attributes. These visual attributes can be inferred by visualizing the *heatmaps* (e.g. via CAM, Zhou et al. (2016)) that highlight the importance of image pixels for those neural features (top row in Figure 1). Alternatively, one can amplify these visual attributes in images by maximizing the corresponding neural feature values via a procedure called the *feature attack* (Singla et al. (2021), bottom row in Figure 1).

Our framework (shown in Figure 2) is based on this key idea: to identify spurious or causal *visual attributes* used for predicting the class i, we identify spurious or causal *neural features* via limited human supervision. We say that a visual attribute is causal for class i if it is always a part of the object i and spurious otherwise. In particular, to annotate a neural feature as causal or spurious, we show only the top-5 images (with a predicted label of i) that maximally activate that neural feature to Mechanical Turk workers. We then show that these neural feature annotations *generalize* extremely well to top-k (with $k \gg 5$) images with label i that maximally activate that neural feature. For example, on Imagenet, in Section 3.4, we show a generalization accuracy of 95% for $k = 65$ (i.e., a 13 fold generalization). Thus, by using neural features and their corresponding neural activation maps (heatmaps), with *limited* human supervision, we can identify a large subset of samples that contain the relevant spurious or causal visual attributes. The neural activation maps for these images can then be used as soft segmentation masks for these visual attributes. We emphasize that the usual method of obtaining such a set would require the manual detection and segmentation of a visual attribute across all images in the training set and is therefore not scalable.

We apply our proposed methodology to the Imagenet dataset: we conducted a Mechanical Turk study using 232 classes of Imagenet and 5 neural features per class. For each neural feature, we obtained its annotation as either causal or spurious for the label. Out of the $232 \times 5 = 1,160$ neural features, 160 were deemed to be spurious by the workers. For 93 classes, at least one spurious neural feature was discovered. Next, for each (class=i, feature=j) pair, we obtained 65 images with class i showing highest activations of the neural feature j and computed their neural activation maps. The union of these images is what we call the *Causal Imagenet* dataset. The dataset contains $52,521$ images with around 226 images per class on average. Each instance in the dataset is of the form $(\mathbf{x}, y, \mathcal{M}^c, \mathcal{M}^s)$ where y is the ground truth class and $\mathcal{M}^s, \mathcal{M}^c$ represent the set of spurious and

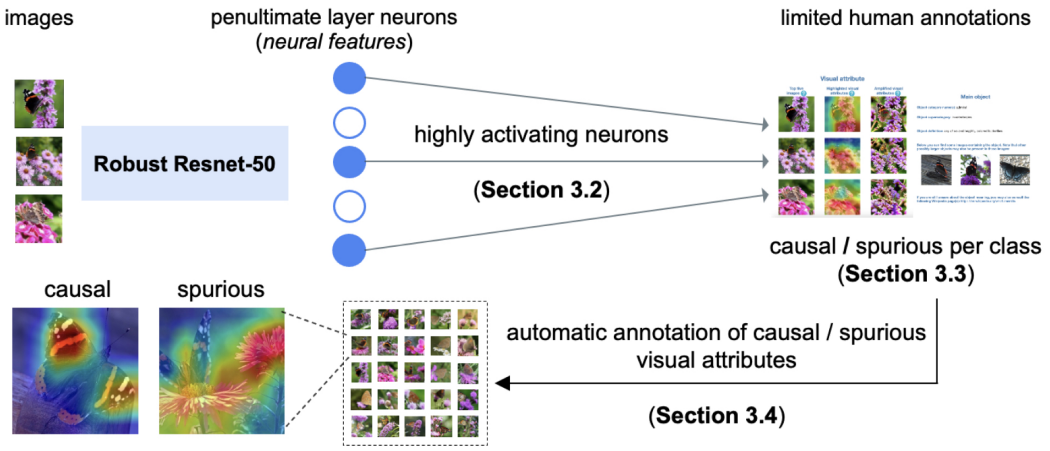


Figure 2: Our framework for constructing the Causal Imagenet dataset. We annotate *neural features* as causal or spurious using minimal human supervision. Then, we show that these annotations *generalize* extremely well to many more samples making our methodology scalable. We highlight causal and spurious regions on the input samples using activation maps of these neural features.

causal masks, respectively (obtained from neural activation maps of neural features). This dataset can be used to test the sensitivity of *any pretrained model* to different image features by corrupting images using their relevant masks and observing the drop in the accuracy.

Ideally, we expect our trained models to show a low drop in accuracy for corruptions in spurious regions and a high drop for that of the causal regions. However, using a standard Resnet-50 model, we discover multiple classes for which the model shows a trend contradictory to our expectations: a significantly higher *spurious drop* compared to the *causal drop*. As an example, for the class drake (Figure 5), we discover that corrupting any of the spurious features results in a significantly higher drop ($> 40\%$) compared to corrupting any of the causal features ($< 5\%$). We show many more examples in Appendix J. For various standard models (Resnet-50, Wide-Resnet-50-2, Efficientnet-b4, Efficientnet-b7), we evaluate their accuracy drops due to corruptions in spurious or causal regions at varying degrees of corruption noise and find no significant differences between their spurious or causal drops suggesting that none of the models differentiates between causal and spurious visual attributes in their predictions (Figure 6).

To the best of our knowledge, this is the first work that introduces a scalable methodology to discover causal and spurious features with limited human supervision by leveraging neurons in the penultimate layer of a robust model as interpretable visual feature detectors. Also, this is the first work that releases a version of Imagenet with annotated spurious and causal image features. We believe that this work opens future research directions to leverage richly annotated datasets to develop deep models that mainly rely on causal and meaningful features in their inferences.

2 Related Work

Robustness and Interpretability: In recent years, there have been many efforts towards post-hoc interpretability techniques for trained neural networks. Most of these efforts have focused on *local explanations* where decisions about single images are inspected (Zeiler & Fergus, 2014; Yosinski et al., 2016; Dosovitskiy & Brox, 2016; Mahendran & Vedaldi, 2016; Nguyen et al., 2016; Zhou et al., 2018; Olah et al., 2018; Adebayo et al., 2018; Chang et al., 2019; Carter et al., 2019; Yeh et al., 2019; Sturmfels et al., 2020; O'Shaughnessy et al., 2019; Verma et al., 2020). These include saliency maps (Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017; Singla et al., 2019), class activation maps (Zhou et al., 2016; Selvaraju et al., 2019; Bau et al., 2020; Ismail et al., 2019; 2020), surrogate models to interpret local decision boundaries such as LIME (Ribeiro et al., 2016), methods to maximize neural activation values by optimizing input images (Nguyen et al., 2015; Mahendran & Vedaldi, 2015) and finding influential inputs (Koh & Liang, 2017). However, recent work suggests that class activation or saliency maps may not be sufficient to narrow down visual

attributes and methods for maximizing neural activations often produce noisy visualizations (Olah et al., 2017; 2018) with changes imperceptible to humans. To address these limitations, recent works (Tsipras et al., 2018; Engstrom et al., 2019) show that for robust (Madry et al., 2018) or adversarially trained models (in contrast to standard models), activation maps are qualitatively more interpretable and optimizing an image directly to maximize a certain neuron produces human perceptible changes in the images which can be useful for visualizing the learned features.

Failure explanation: Most of the prior works on failure explanation either operate on tabular data where interpretable features are readily available (Zhang et al., 2018; Chung et al., 2019), language data where domain-specific textual queries can be easily generated (Wu et al., 2019; Ribeiro et al., 2020), images where visual attributes can be collected using crowdsourcing (Nushi et al., 2018; Xiao et al., 2021) or photorealistic simulation (Leclerc et al., 2021). However, collecting visual attributes via crowdsourcing can be expensive for large datasets. Moreover, these methods require humans to hypothesize about the possible failures and one could miss critical failure modes when the visual attributes used by the model are different from the ones hypothesized by humans.

To address these limitations, recent works (Wong et al., 2021; Singla et al., 2021) use the neurons of robust models as the visual attribute detectors for discovering failure modes, thus avoiding the need for crowdsourcing. However, Wong et al. (2021) cannot be used to analyze the failures of standard (non-robust) models which achieve significantly higher accuracy and are more widely deployed in practice. Moreover, to discover the spurious feature, they ask MTurk workers to identify the common visual attribute across multiple images without highlighting any region of interest. However, different images may have multiple common attributes even if they come from different classes and this approach may not be useful in such cases. Barlow (Singla et al., 2021) learns decision trees using misclassified instances to discover leaf nodes with high failure concentration. However, the instances could be correctly classified due to some spurious feature (e.g., background) and Barlow will discover no failure modes in such cases. In contrast, our approach circumvents these limitations and discovers the spurious features for standard models even when they achieve high accuracy.

3 Scalable framework for discovering Spurious Features

Consider an image classification problem $\mathcal{X} \rightarrow \mathcal{Y}$ where the goal is to predict the ground truth label $y \in \mathcal{Y}$ for inputs $\mathbf{x} \in \mathcal{X}$. For each class $i \in \mathcal{Y}$, we want to identify a set of visual attributes in the training set that can be used by neural networks to predict the label i. Some of these attributes will be causal while others spurious. We define *causal attributes* as the set of visual features that are always a part of the object while *spurious attributes* are the ones that are likely to *co-occur* with the object, but not a part of it (such as food, vase, flowers in Figure 1). Additionally, we want to localize these attributes on a large subset of training images to assess the sensitivity of models to them. However, these attributes are abstract objects and not mathematically well-defined. Moreover, localizing these visual attributes on a large number of images requires significant human annotations and can be very expensive. In this paper, we propose a methodology to identify and localize a subset of visual attributes and determine if they are spurious or causal with minimal human supervision.

3.1 Notation and Definitions

For a trained neural network, the set of activations in the penultimate layer (adjacent to logits) is what we call the *neural feature vector*. Each element of this vector is called a *neural feature*. For an input image and a neural feature, we can obtain the *Neural Activation Map* or *NAM*, similar to CAM (Zhou et al., 2016), that provides a soft segmentation mask for highly activating pixels (for the neural feature) in the image (details in Appendix A.1). The corresponding *heatmap* can then be obtained by overlaying the NAM on top of the image so that the red region highlights the highly activating pixels (Appendix A.2). The *feature attack* (Singla et al., 2021) is generated by optimizing the input image to increase the values of the desired neural feature (Appendix A.3).

3.2 Extracting, visualizing and selecting neural features

Prior works (Engstrom et al., 2019; Singla et al., 2021) provide evidence that neural features of a robust model encode human interpretable visual attributes. Building on these works, to identify spurious or causal visual attributes used for predicting class i, we identify neural features encoding

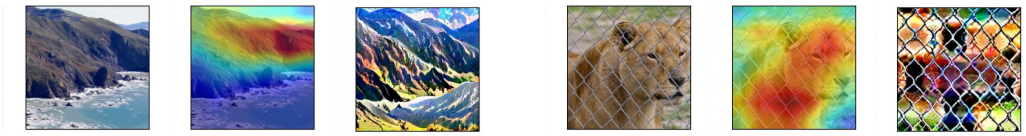


Image Heatmap Feature Attack Image Heatmap Feature Attack

Figure 3: On the left, the heatmap suffices to explain that the focus is on ground. But on the right, the heatmap covers both cougar and wires. Feature attack clarifies that the focus is on wires.

these attributes. For each image in the Imagenet training set (Deng et al., 2009), we extract the neural feature vector using a robust Resnet-50 model. The robust model was adversarially trained using the l_2 threat model of radius 3. Each neural feature is visualized using the following techniques:

- **Most activating images**: By selecting 5 images that maximally activate the neural feature, the common visual pattern in these images can be interpreted as the visual attribute encoded in it.
- **Heatmaps (NAMs)**: When there are multiple common visual attributes, heatmaps highlight the region of interest in images and can be useful to disambiguate; e.g., in Figure 3 (left panels), heatmap disambiguates that the focus of the neural feature is on the ground (not water or sky).
- **Feature attack**: In some cases, heatmaps may highlight multiple visual attributes. In such cases, the feature attack can be useful to disambiguate; e.g., in Figure 3 (right panels), the feature attack indicates that the focus of the neural feature is on the wires (not cougar).

However, the neural feature vector can have a large size (2048 for the Resnet-50 used in this work) and visualizing all of these features per class to determine whether they are causal or spurious can be difficult. Thus, we select a small subset of these features that are highly predictive of class i and annotate them as causal or spurious for i. Prior work (Singla et al., 2021) selects this subset by considering the top-k neural features with the highest mutual information (MI) between the neural feature and the *model failure* (a binary variable that is 1 if the model misclassifies $\mathbf{x}$ and 0 otherwise). However, the model could be classifying all the images correctly using some spurious features. In this case, the MI value between the (spurious) neural feature and the model failure will be small and thus this approach will fail to discover such critical failure modes.

In this work, we select a subset of neural features *without* using the model failure variable. To do so, we first select a subset of images (from the training set) on which the robust model predicts the class i. We compute the mean of neural feature vectors across all images in this subset denoted by $\overline{\mathbf{r}}(i)$. From the weight matrix $\mathbf{w}$ of the last linear layer of the robust model, we extract the ith row $\mathbf{w}_{i,:}$ that maps the neural feature vector to the logit for the class i. Next, we compute the hadamard product $\overline{\mathbf{r}}(i) \odot \mathbf{w}_{i,:}$. Intuitively, the jth element of this vector $(\overline{\mathbf{r}}(i) \odot \mathbf{w}_{i,:})_j$ captures the mean contribution of neural feature j for predicting the class i. This procedure leads to the following definition:

Definition 1 *The Neural Feature Importance of feature j for class i is defined as:*

$$\mathrm{IV}_{i,j} = (\overline{\mathbf{r}}(i) \odot \mathbf{w}_{i,:})_j \,.$$

For class i, the neural feature with k^{th} highest feature importance is said to have the feature rank k.

We thus select the neural features with the highest-5 importance values (defined above) per class.

3.3 Mechanical Turk study for discovering spurious and causal features

To determine whether a neural feature j is causal or spurious for class i, we conducted a crowd study using Mechanical Turk (MTurk). We show the MTurk workers two panels: One panel visualizes the feature j and the other describes the class i. To visualize j, we show the 5 most activating images (with predicted class i), their heatmaps highlighting the important pixels for j, and the feature attack visualizations (Engstrom et al., 2019). For the class i, we show the object names (Miller, 1995), object supercategory (from Tsipras et al. (2020)), object definition and the wikipedia links. We also show 3 images of the object from the Imagenet validation set. We then ask the workers to determine whether the visual attribute is a part of the main object, some separate objects or the background.

We also ask them to provide reasons for their answers and rate their confidence on a likert scale from 1 to 5. The design for this study is shown in Figure 9 in Appendix C. Each of these tasks (also called Human Intelligence Tasks or HITs) were evaluated by 5 workers. The HITs for which the majority of workers selected either separate objects or background as their responses were deemed to be spurious and the ones with main object as the majority answer were deemed to be causal.

Because conducting a study for all 1000 Imagenet classes can be expensive, we selected a smaller subset of 232 classes (denoted by $\mathcal{T}$) by taking a union of 50 classes with the highest and 50 with the lowest accuracy across multiple models (details in Appendix B). For each class, we selected 5 neural features with the highest feature importance values resulting in $232 \times 5 = 1160$ HITs. For 160 HITs, workers deemed the feature to be spurious for the label. Thus, for each class $i \in \mathcal{T}$, we obtain a set of causal and spurious neural features denoted by $\mathcal{C}(i)$ and $\mathcal{S}(i)$, respectively.

3.4 Generalization of visual attributes encoded in neural features

We show that the visual attribute inferred for some neural feature j (via visualizing the top-5 images with predicted class i as discussed in the previous section) *generalizes* extremely well to top-k (with $k \gg 5$) images with the ground truth label i that maximally activate feature j. That is, we can obtain a large set of images containing the visual attribute (encoded in neural feature j) by only inspecting a small set of images. To obtain the desired set, denoted by $\mathcal{D}(i, j)$, we first select training images with label i (Imagenet contains ≈ 1300 images per label). Next, we select $k = 65$ images (5% of 1300) from this set with the highest activations of the neural feature j. We use the Neural Activation Maps (NAMs) for these images as soft segmentation masks to highlight the desired visual attributes.

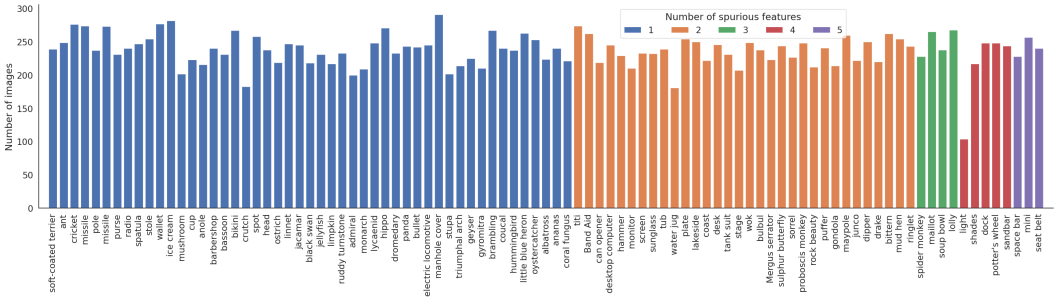
To validate that the NAMs focus on the desired visual attributes, we conducted another Mechanical Turk study. From the set $\mathcal{D}(i, j)$, we selected 5 images with the highest and 5 with the lowest activations of feature j. We show the workers two panels: The first panel shows images with the highest 5 activations (along with their heatmaps) and the second shows images with the lowest 5 activations (and heatmaps). We then ask them to determine if the focus of heatmaps in both visualizations was on the same visual attribute, different attributes or if either of the two visualizations was unclear. For each spurious feature (160 total), we obtained answers from 5 workers each. For 152 visualizations (95%), majority of the workers selected *same* as the answer, thereby validating that the NAMs focus on the same visual attribute. The design for the study is shown in Figure 10 in Appendix D.

We emphasize that although using $k = 65$ significantly increases the size of the set (i.e., a 13 fold increase over the manually annotated set), the value of k can be changed depending on the number of training images with the desired properties. One can quickly search over various possible values of k by visualizing the 5 images (and their heatmaps) with the highest and lowest activations of the relevant neural feature (in the set of k images) until they both focus on the desired visual attribute.

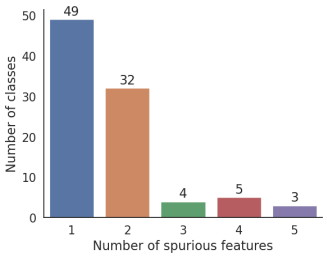
4 The Causal Imagenet dataset

Recall that for each class $i \in \mathcal{T}$, we obtain a set of causal and spurious neural features denoted by $\mathcal{C}(i)$ and $\mathcal{S}(i)$, respectively (Section 3.3). Then, using the generalization property (Section 3.4), we obtain the set $\mathcal{D}(i, j)$ for each $i \in \mathcal{T}$ and $j \in \mathcal{C}(i) \cup \mathcal{S}(i)$. Each $\mathcal{D}(i, j)$ contains 65 images along with the NAM (per image) acting as the soft segmentation mask for the visual attribute encoded in feature j. We note these soft segmentation masks overlap with the desired visual attribute and may not cover it completely. However, since we corrupt these regions using Gaussian noise which *preserves the content* of the original images (discussed in Section 5), they can be useful for evaluating the model sensitivity to various spurious/causal visual attributes. If $j \in \mathcal{C}(i)$, these are called *causal masks*, otherwise *spurious masks*. The union of all these datasets $\mathcal{D}(i, j)$ is what we call the *Causal Imagenet* dataset. Each instance in the dataset is of the form $(\mathbf{x}, y, \mathcal{M}^c, \mathcal{M}^s)$ where y is the ground truth label while $\mathcal{M}^s$ and $\mathcal{M}^c$ represent the set of spurious and causal masks (obtained from NAMs of neural features) for the image $\mathbf{x}$, respectively. Anonymized data from our Mechanical Turk studies and the Causal Imagenet dataset will be made available upon acceptance.

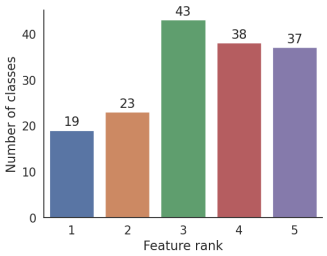
The dataset contains $52,521$ images with around 226 images per class on average. From the MTurk study (Section 3.3), 160 features were discovered to be spurious and $1,000$ to be causal for the relevant classes. We visualize 80 spurious features in Appendix Section G. Examples of background and foreground spurious features are in Appendix G.1 and G.2, respectively. For 93 classes, we



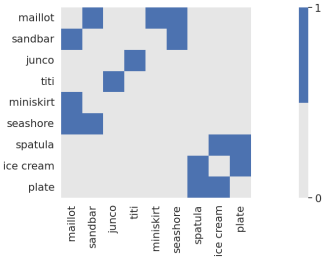
(a) Number of images for different classes with at least 1 spurious feature



(b) Number of classes vs. number of spurious features



(c) Number of classes with spurious feature at various ranks



(d) Number of common spurious neural features between classes

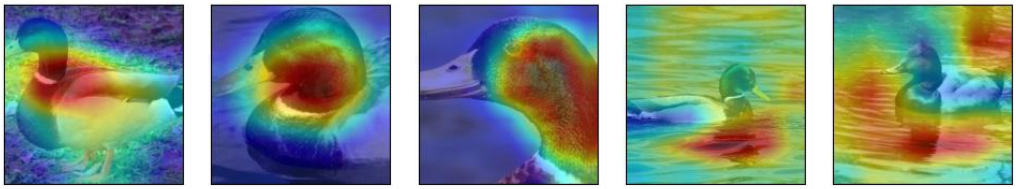
Figure 4: In the top row, we show number of images for different classes with at least 1 spurious feature in the Causal Imagenet dataset. Bottom row shows various plots using the dataset.

discover at least one spurious feature (the number of images per class shown in Figure 4a). We show a histogram of the number of classes vs. the number of spurious features in Figure 4b. For 3 classes (namely space bar, miniskirt and seatbelt) all the 5 features were found to be spurious. This finding suggests that in safety critical applications such as biomedical imaging and autonomous driving, it can be unsafe to just use the test accuracy to assess the model performance because the model may be highly accurate on the benchmark for *all the wrong reasons* (such as seatbelt, Appendix H).

We plot the histogram of the number of classes with spurious feature at rank k (from 1 to 5) in Figure 4c. We discover that across all feature ranks, a significant number of spurious features are discovered (minimum 19) and a larger number of spurious features are discovered at higher feature ranks (4 and 5). This suggests that inspecting a larger number of features per class (i.e., > 5) may be necessary for a more thorough analysis of spurious features. In Figure 4d, we show the number of shared spurious features between different classes (diagonal is zeroed out for brevity). By visualizing common neural features, we diagnose that legs are a confusing visual attribute between maillot and miniskirt (Appendix I.1); tree branches for titi, junco (Appendix I.1); food for spatula, plate, icecream (Appendix I.3); water for maillot, sandbar, seashore (Appendix I.4).

5 Evaluating deep models using the Causal Imagenet dataset

In this section, we use the Causal Imagenet dataset to evaluate the performance of several pretrained models on Imagenet. In particular, each set $\mathcal{D}(i, j)$ can be used to test the sensitivity of *any trained model* to the visual attribute encoded with feature j for predicting the class i. We can either blackout (fill with black color) or corrupt (using some noise model) the region containing the attribute (using the soft segmentation masks) for all images in this set and evaluate the accuracy of the model on these corrupted images. However, because the existing Imagenet pretrained models have not explicitly been trained on images with black regions, these images can be characterized as out of distribution of the training set. Thus, it may be difficult to ascertain whether changes in accuracy are due to the distribution shift or the removal of spurious features. Instead, we choose to corrupt the region of interest using Gaussian noise to preserve the content of the original images (which can be useful when the soft segmentation masks are inaccurate). Such Gaussian noise is also known to occur in the real world due to sensor or electronic circuit noise. Since the segmentation masks we derive are



(a) Causal, **-1.5%** (b) Causal, **-4.6%** (c) Causal, **-0%** (d) Spurious,**-41.5%** (e) Spurious,**-46.1%**

Figure 5: Each image denotes the heatmap for different neural features for the class drake. Using a standard Resnet-50 model, (clean accuracy of 95.4%), we observe a drop of *at least* 41.5% by adding gaussian noise with $\sigma = 0.25$ to spurious masks while a drop of *at most* 4.6% for causal masks. This shows that the model heavily relies on spurious features in its predictions for this class.

soft (i.e., their values lie between 0 and 1, not binary), we use these masks to control the degree of the Gaussian noise corruption across the original image. That is, given image $\mathbf{x}$ and mask $\mathbf{m}$, we add noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using the equation below (examples in Appendix F):

$$\mathbf{x} + \sigma\left(\mathbf{z} \odot \mathbf{m}\right), \tag{1}$$

where σ is a hyperparameter that can be used to control the degree of corruption in the image. The above equation ensures that regions where the mask $\mathbf{m}$ has values close to 1 are corrupted by noise with the desired standard deviation σ and the ones with values close to 0 suffer little changes.

5.1 Causal and spurious accuracy

We now introduce concepts of *causal* and *spurious* accuracy to evaluate deep models. Informally, the causal accuracy is the accuracy of the model only due to the causal regions of the images and the spurious accuracy is due to the spurious regions. To compute the causal accuracy for class i, we first obtain the sets $\mathcal{D}(i, j)$. For some image $\mathbf{x} \in \mathcal{D}(i, j)$, let $\mathbf{m}^{(j)}$ be the mask obtained for some feature j. Next, we take the union of all these sets for $j \in \mathcal{S}(i)$ denoted by $\mathcal{DS}(i)$. Note that for some images in the union, we may have multiple spurious masks but to evaluate the causal accuracy, we want to obtain a single mask per image that covers all the spurious attributes. Thus, for the image $\mathbf{x}$, the (p, q)-th element of the single spurious mask $\mathbf{s}$ (or, the single causal mask $\mathbf{c}$), is computed by taking the element-wise max over all the relevant masks:

$$\mathbf{s}_{p,q} = \max_{j \in \mathcal{S}(i),\ \mathbf{x} \in \mathcal{D}(i,j)} \mathbf{m}^{(j)}_{p,q}, \qquad \mathbf{c}_{p,q} = \max_{j \in \mathcal{C}(i),\ \mathbf{x} \in \mathcal{D}(i,j)} \mathbf{m}^{(j)}_{p,q} \tag{2}$$

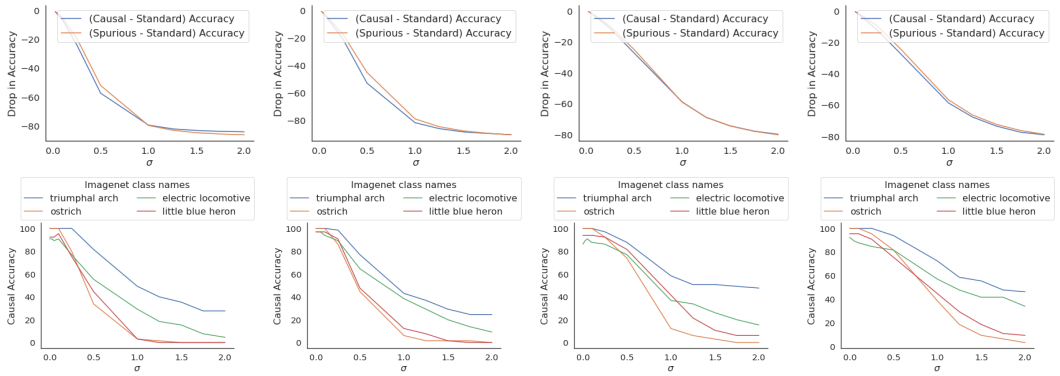
Definition 2 ***(Causal Accuracy)*** *We define the Causal Accuracy or* $\mathrm{acc}^{(\mathcal{C})}$ *as follows:*

$$\mathrm{acc}^{(\mathcal{C})} = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \mathrm{acc}^{(\mathcal{C})}(i), \quad \mathrm{acc}^{(\mathcal{C})}(i) = \frac{1}{|\mathcal{DS}(i)|} \sum_{\mathbf{x} \in \mathcal{DS}(i)} \mathbb{1}\left(h\left(\mathbf{x} + \sigma\left(\mathbf{z} \odot \mathbf{s}\right)\right) = y\right)$$

We acknowledge that our definition of causal accuracy is *incomplete* because the set of spurious visual attributes discovered using our framework may not cover *all* spurious attributes in the dataset. Moreover, we use causality in a very loose sense to analyze the sensitivity of trained models to various spurious visual attributes. Finally, note that the defined causal accuracy is a function of the noise parameter (σ) used to corrupt spurious regions. The Spurious Accuracy or $\mathrm{acc}^{(\mathcal{S})}(i)$ can be computed similarly by replacing $\mathcal{DS}(i)$ with $\mathcal{DC}(i) = \cup_{j \in \mathcal{C}(i)} \mathcal{D}(i, j)$, and $\mathbf{s}$ with $\mathbf{c}$ in equation 2. Observe that $\mathcal{DS}(i) \neq \mathcal{DC}(i)$ in general because $\mathcal{D}(i, j)$ may contain different sets of images for different j's. Thus, the standard accuracy of the model (i.e. without adding any noise) on the two sets pf $\mathcal{DS}(i)$ and $\mathcal{DC}(i)$ can be different. We want our trained models to show a low degradation in performance when noise is added to spurious regions (i.e. high causal accuracy) and a high degradation when corrupting causal regions (i.e. low spurious accuracy).

5.2 Results from evaluating deep models on the Causal Imagenet dataset

In this section, we test the sensitivity of several models to various spurious features and assess their causal and spurious accuracy using the Causal Imagenet dataset.



(a) Resnet-50 (b) Wide-Resnet-50-2 (c) Efficientnet-b4 (d) Efficientnet-b7

Figure 6: In the top row, we observe a similar drop in accuracy irrespective of whether noise is added to spurious or causal regions suggesting that trained models do not differentiate between the spurious and causal regions for predicting the desired class. In the bottom row, we show causal accuracy vs. noise parameter σ for several classes. In particular, for both classes "triumphal arch" and "ostrich", the model has a standard accuracy of 100% (at $\sigma = 0$). However, the causal accuracy for triumphal arch is high even at $\sigma = 2.0$ ($\approx 40\%$) while that for ostrich is almost 0%.

Testing model sensitivity to spurious features. We test the sensitivity of a standard Resnet-50 model to different spurious and causal features (one feature at a time) by corrupting images from the datasets $\mathcal{D}(i, j)$ and evaluating the drop in model accuracy. We use $\sigma = 0.25$ (equation 1) because we observe that it preserves the content of the images (Examples in Appendix F) so we would expect the model prediction to remain unchanged. In Figure 1, we show multiple examples of spurious features discovered using this procedure. In Figure 5, we show a class (namely drake) on which the model has a high clean accuracy (95.4%). However, it exhibits a large drop in performance when any of the spurious regions (e.g. water) are corrupted while a small drop is observed when any of the causal regions (e.g. bird's head) are corrupted. This highlights that the model relies on mostly spurious features in its predictions for this class. Out of the 93 classes with at least one spurious feature, we discovered 34 classes such that the accuracy drop due to some spurious feature was at least 20% higher than due to some causal feature. In Appendix J, we show visualizations of causal and spurious features for 17 of these classes. On 11 of these 17 classes, the model has $> 95\%$ accuracy. Since the l_2 norm of noise perturbations can be different for different features (e.g. due to different l_2 norms of their activation maps), we also evaluate accuracy drops due to corrupting spurious/causal features when the mean l_2 perturbations are similar and find similar trends (details in Appendix J).

Comparing causal and spurious accuracy of trained models. Next, we compute the causal and spurious accuracy of four standard pretrained models namely, Resnet-50 (He et al., 2015), Wide Resnet-50-2 (Zagoruyko & Komodakis, 2016), Efficientnet-b4 and Efficientnet-b7 (Tan & Le, 2019). In Figure 6 (top row), we plot the drop in accuracy when noise is added to spurious regions (i.e., causal accuracy - standard accuracy) and also to causal regions (i.e., spurious accuracy - standard accuracy). We observe a similar drop in performance for all values of σ and for all trained models. This suggests that trained models do not differentiate between spurious and causal regions of the images for predicting an object class. In some sense, this is expected because in the standard Empirical Risk Minimization (ERM) paradigm, models have never explicitly been trained to make such differentiation, suggesting that providing additional supervision during training (for example, segmenting causal/spurious regions as provided in Causal Imagenet or diversifying the training dataset) may be essential to train models that achieve high causal accuracy. In Figure 6 (bottom row), we plot the causal accuracy for 4 different classes as the noise level σ increases. In particular, consider the two classes "triumphal arch" and "ostrich" that both have standard accuracy of 100% (at $\sigma = 0$). However, the causal accuracy for triumphal arch is high even at $\sigma = 2.0$ ($\approx 40\%$) while that for ostrich is almost 0%. This provides further evidence that the standard accuracy alone is not a reliable measure for model performance in the real world because as we show that the casual accuracy for two classes with the same standard accuracy can be very different. We believe having richly annotated training datasets such as Causal Imagenet can lead to training reliable deep models that mainly rely on causal and informative features in their predictions.

REPRODUCIBILITY

We provide the design details of our Mechanical Turk studies in Appendix Sections C and D. The code for user studies and experiments is provided in the supplementary materials. We provide some samples from the Causal Imagenet dataset in supplementary materials. The full dataset has a large size and will be made publicly available upon acceptance.

ETHICS STATEMENT

This paper introduces a scalable framework for discovering spurious features in predictions made by deep neural networks. Our approach can be useful for various downstream tasks such as model debugging and developing improved models that rely on causal features in their predictions. We introduce a new dataset called Causal Imagenet that includes annotations of various causal and spurious visual attributes for a subset of images in the Imagenet training set. We do *not* collect any new samples for constructing our dataset. As this work paves the way to develop reliable models, we do not see any foreseeable negative consequences associated with our work.

REFERENCES

Julius Adebayo, J. Gilmer, M. Muelly, Ian J. Goodfellow, M. Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

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. ISSN 0027-8424. doi: 10.1073/pnas.1907375117. URL https://www.pnas.org/content/117/48/30071.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. *CoRR*, abs/1807.04975, 2018. URL http://arxiv.org/abs/1807.04975.

Alceu Bissoto, E. Valle, and S. Avila. Debiasing skin lesion datasets and models? not so fast. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3192–3201, 2020.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1MXz20cYQ.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1550–1553. IEEE, 2019.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *CoRR*, abs/1905.11979, 2019. URL http://arxiv.org/abs/1905.11979.

J. Deng, Wei Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

Aya Abdelsalam Ismail, Mohamed K. Gunady, Luiz Pessoa, Héctor Corrada Bravo, and Soheil Feizi. Input-cell attention reduces vanishing saliency of recurrent neural networks. In *NeurIPS*, 2019.

Aya Abdelsalam Ismail, Mohamed K. Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *NeurIPS*, 2020.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1885–1894. JMLR.org, 2017.

Guillaume Leclerc, Hadi Salman, Andrew Ilyas, Sai Vemprala, Logan Engstrom, Vibhav Vineet, Kai Xiao, Pengchuan Zhang, Shibani Santurkar, Greg Yang, Ashish Kapoor, and Aleksander Madry. 3db: A framework for debugging computer vision models. In *Arxiv preprint arXiv:2106.03805*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Aravindh Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120:233–255, 2016.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196, 2015. doi: 10.1109/CVPR.2015.7299155.

George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In *ICML Workshop on Visualization for Deep Learning*, 2016.

Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. In Yiling Chen and Gabriella Kazai (eds.), *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP*, pp. 126–135. AAAI Press, 2018. URL https://aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17930.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell. Generative causal explanations of black-box classifiers. In *NeurIPS*, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*, 2020.

R. R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, D. Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Sahil Singla, Eric Wallace, Shi Feng, and Soheil Feizi. Understanding impacts of high-order loss approximations and features in deep learning interpretation. In *ICML*, 2019.

Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*, 2017.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/tan19a.html`.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *ArXiv preprint arXiv:2005.11295*, 2020.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review, 2020.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11205–11216. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/wong21b.html`.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 747–763, 2019.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=gl3D-xY7wLq`.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep D. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, 2019.

Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop*, 2016.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

Matthew D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9): 2131–2145, 2018.