A Comprehensive Study of Image Classification Model Sensitivity to Foregrounds, Backgrounds, and Visual Attributes

Mazda Moayeri[1] mmoayeri@umd.edu
Phillip Pope[1] pepope@umd.edu
Yogesh Balaji[2] ybalaji@nvidia.com
Soheil Feizi[1] sfeizi@cs.umd.edu

[1] University of Maryland [2] NVIDIA

Abstract

While datasets with single-label supervision have propelled rapid advances in image classification, additional annotations are necessary in order to quantitatively assess how models make predictions. To this end, for a subset of ImageNet samples, we collect segmentation masks for the entire object and 18 *informative attributes. We call this dataset RIVAL10 (***RI***ch* ***V****isual* ***A****ttributes with* ***L****ocalization), consisting of roughly* $26k$ *instances over* 10 *classes. Using RIVAL10, we evaluate the sensitivity of a broad set of models to noise corruptions in foregrounds, backgrounds and attributes. In our analysis, we consider diverse state-of-the-art architectures (ResNets, Transformers) and training procedures (CLIP, SimCLR, DeiT, Adversarial Training). We find that, somewhat surprisingly, in ResNets, adversarial training makes models more sensitive to the background compared to foreground than standard training. Similarly, contrastively-trained models also have lower relative foreground sensitivity in both transformers and ResNets. Lastly, we observe intriguing adaptive abilities of transformers to increase relative foreground sensitivity as corruption level increases. Using saliency methods, we automatically discover spurious features that drive the background sensitivity of models and assess alignment of saliency maps with foregrounds. Finally, we quantitatively study the attribution problem for neural features by comparing feature saliency with ground-truth localization of semantic attributes.*

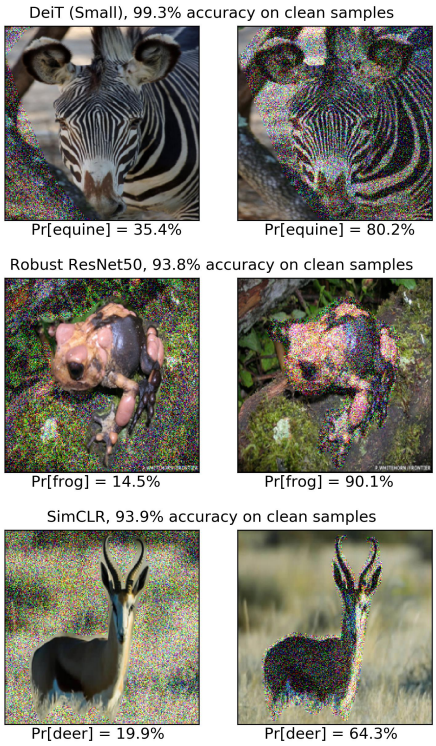

Figure 1. Examples where background noise degrades performance of highly accurate models more than foreground noise. Gaussian ℓ_∞ noise with standard deviation $\sigma = 0.24$ shown. Probabilities are averaged over ten trials. While these examples are cherry picked, we observe that they are surprisingly prevalent, and model design can affect the degree to which such cases arise.

1. Introduction

Large scale benchmark datasets like ImageNet [9] that were constructed with single class label annotation have propelled rapid advances in the image classification task [18, 21, 50, 58]. Over the last decade, several network architectures and training procedures were proposed to yield very high classification accuracies [10, 18, 45, 50]. However, methods to interpret these model predictions and to diagnose undesirable behaviors are fairly limited. One of the most popular class of approaches are saliency methods [43, 48, 49, 59] that use model gradients to produce a

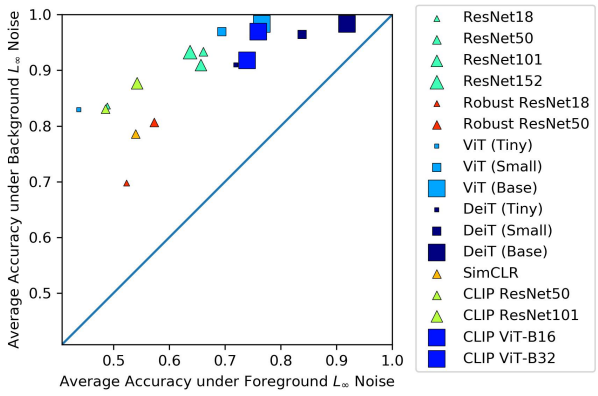

Figure 2. Accuracy under noise in foregrounds and backgrounds, averaged over multiple noise levels. Marker size is proportional to parameter count. Models with higher relative foreground sensitivity lie further from the diagonal.

saliency map corresponding to the most influential input regions that yielded the resulting prediction. However, these methods are qualitative, require human supervision, and can be noisy, thus making their judgements potentially unreliable when made in isolation of other supporting analysis.

In this paper, we argue that to obtain a proper understanding of how specific input regions impact the prediction, we need additional ground truth annotations beyond a single class label. To this end, we introduce a novel dataset, RIVAL10, whose samples include **RI**ch **V**isual **A**ttributions with **L**ocalization. RIVAL10 consists of images from 20 categories of ImageNet-1k [9], with a total of $26k$ high resolution images organized into 10 classes, matching those of CIFAR10 [26]. The main contribution of our dataset is instance wise labels for 18 informative visual attributes, as well as segmentation masks for each attribute and the entire object. We present our dataset as a general resource for understanding models trained on ImageNet. We then provide a study of the sensitivity of a diverse set of models to foregrounds, backgrounds, and attributes.

Our study of background and foreground model sensitivity is motivated by some counter-intuitive model behaviors on images whose background and foreground regions were corrupted with Gaussian noise: Figure 1 shows instances where highly accurate models have performance degraded much more due to the background noise than the foreground noise. While this is not the norm (i.e. models are more sensitive to foregrounds on average), the existence of these examples warrants greater investigation, as they expose a stark difference in how deep models and humans perform object recognition. Quantifying the degree to which different architectures and training procedures admit these examples can shed new insight on how models incorporate foreground and background information.

To this end, we conduct a *noise analysis* that leverages object segmentation masks to quantitatively assess model sensitivity to foregrounds relative to backgrounds. We proxy sensitivity to a region by observing model performance under corruption of that region. We propose a normalized metric, *relative foreground sensitivity* (RFS), to compare models with various general noise robustness. A high RFS value indicates that the model uses foreground features in its inferences more than background ones since corrupting them result in higher performance degradation.

In Figure 2, we see different architectures and training procedures lead to variations in both general noise robustness (projection onto the main diagonal) and relative foreground sensitivity (normalized distance orthogonal to the diagonal). Notably, we find that adversarially training ResNets significantly reduces RFS, surprisingly suggesting that robust ResNet models make greater use of background information. We also observe contrastive training to reduce RFS, and transformers to uniquely be able to adjust RFS across noise levels, reducing their sensitivity to backgrounds as corruption level increases. Lastly, we find object classes strongly affect RFS across models.

We couple our noise analysis with saliency methods to add a second perspective of model sensitivity to different input regions. Using RIVAL10 segmentations, we can quantitatively assess the alignment of saliency maps to foregrounds. We also show how we can discover spurious background features by sorting images based on the saliency alignment scores. We observe that performance trends that our noise analysis reveals are not captured using qualitative saliency methods alone, suggesting our noise analysis can provide new insights on model sensitivity to foregrounds and backgrounds.

Lastly, we utilize RIVAL10 attribute segmentations to systematically investigate the generalizability of neural feature attribution: for a neural feature (i.e., a neuron in the penultimate layer of the network) that achieves the highest intersection-over-union (IOU) score with a specific attribute mask on top-k images within a class, how the IOU scores of that neural feature behave on other samples in that class. For some class-attribute pairs (e.g. dog, floppy-ears), we indeed observe generalizability of neural feature attributions, in the sense that test set IOUs are also high.

In summary, we present a novel dataset with rich annotations of object and attribute segmentation masks that can be used for a myriad of applications including model interpretability. We then present a study involving three quantitative methods to analyze the sensitivity of models to different regions in inputs. We hope the RIVAL10 dataset will help study failure modes of current deep classifiers and pave the way for building more reliable models in the future.

2. Review of Literature

2.1. Related Datasets

Prior to the rise of deep learning, a number of works studied attribute classification, leading to the construction of datasets such as Animals with Attributes [27] and aPASCAL VOC 2008 [14] (adding annotations to [13]). [54] published CUB 200, a fine-grained classification dataset of bird species with object segmentations and part *localizations* in the form of single coordinates, as opposed to segmentation masks like in RIVAL10. Finally, [41] collected object attributes on a small-scale subset of ImageNet. More recently, [36] publish a large-scale object attribute dataset on a subset of ImageNet. The Celeb-A dataset [29] contains attribution with applications to generative modeling, but limited utility for general representation learning since it only contains face images. The broader dataset Visual Attributes in the Wild (VAW) [38] provides large-scale in the wild attribute annotations for 250k object instances.

Many datasets aim to stress test models to reveal limitations. [19] introduces ImageNet variants under diverse corruption types, including Gaussian noise. [20] adds two more ImageNet variants that include challenging natural samples and out of distribution samples, on which top models see massive accuracy drops. Models evaluated on [2] similarly see large drops, though this dataset differs in that it is strictly a test set. Other works introduce synthetic datasets to assess spatial biases [57] or background reliance of classifiers, such as [56] and [42], which perform some variation of swapping or altering foregrounds and backgrounds. Though similar, these works differ in objective and technical contribution to ours. [42] focuses on developing a novel distributionally robust optimization procedure. [56] emphasizes designing a multitude of test datasets through creative editing of foreground and background regions to serve as a general benchmark to evaluate models. In contrast, our work presents a novel method to analyze foreground sensitivity, and demonstrates its utility by applying it to a breadth of cutting-edge architectures and training paradigms, leading to *model-specific* observations. Further, our RIVAL10 dataset is significantly larger and richer in annotation.

Recently, [46] uses saliency maps and feature visualization in a semi-automated process to identify deep neural nodes corresponding to core or spurious features for an object of a given class, resulting in a large-scale dataset with segmentations corresponding to salient features. However, annotation of the segmented regions are limited to just labeling them as 'core' or 'spurious'.

2.2. Interpretability Methods

A number of methods have been proposed to interpret model predictions, such as saliency or class activation maps [43], influence functions [25], and surrogate white box models [40, 55]. However, saliency maps have been found to be noisy and influence functions are fragile [3, 16]. Some methods seek to interpret the function of a neural node via synthesizing inputs that maximize its activation [33, 35, 47], though these methods are limited when non-adversarially robust models are used [34], and offer qualitative insights.

A motivation behind the development of interpretability methods is to work towards addressing the 'shortcut learning' issue, where models rely on easy-to-learn features that lead to high performance on training sets, but poor generalization in other settings. [15] discusses this at length, recommending the development and usage of challenging datasets whose inputs are out-of-distribution with respect to standard benchmarks. RIVAL10's rich annotations open the door to the construction of many challenge datasets, in which shortcuts are broken via swapping backgrounds, foregrounds, and *attributes* (examples in Appendix).

Other constructive works aimed to reduce the reliance of deep models on spurious features appeal to counterfactual data generation [1, 6, 17], often appealing to disentangled representations or explicit annotations to break correlations of texture, shapes, colors, and backgrounds. Further, [23] found that removing spurious features can in fact hurt accuracy and disproportionately affect groups. Thus, the notion that spurious features are always harmful is incomplete, and a closer look is required to ground discussions regarding the shortcut learning issue. Lastly, [52] provides theoretical context for stress testing models to discern causal factors.

3. RIVAL10

3.1. Overview

RIVAL10 differs from previous attributed datasets in that it provides *attribute-specific* localizations. That is, for every positive instance of an attribute, a binary segmentation mask identifies the image region in which the attribute occurs.

Perhaps, the most similar dataset in this regard is the recent Fashionpedia [22], a dataset providing attributes and localizations of 27 apparel categories. However, the dataset is proposed for the fashion domain which limits its utility for general purpose object recognition task. To the best of our knowledge, RIVAL10 is the first *general domain* dataset to provide both rich semantic attributes and localization, the combination of which we envision to aid in analyzing the robustness and interpretability of deep networks. While other datasets used for semantic segmentation and object detection go beyond single label annotations [8, 12, 28], they are not designed with classifiers specifically in mind, like RIVAL10.

Classes were chosen to be aligned with CIFAR-10 to enable analyzing the existing architectures and training techniques developed for the object recognition task. In particular, the classes we provide are: *bird, car, cat, deer, dog,*

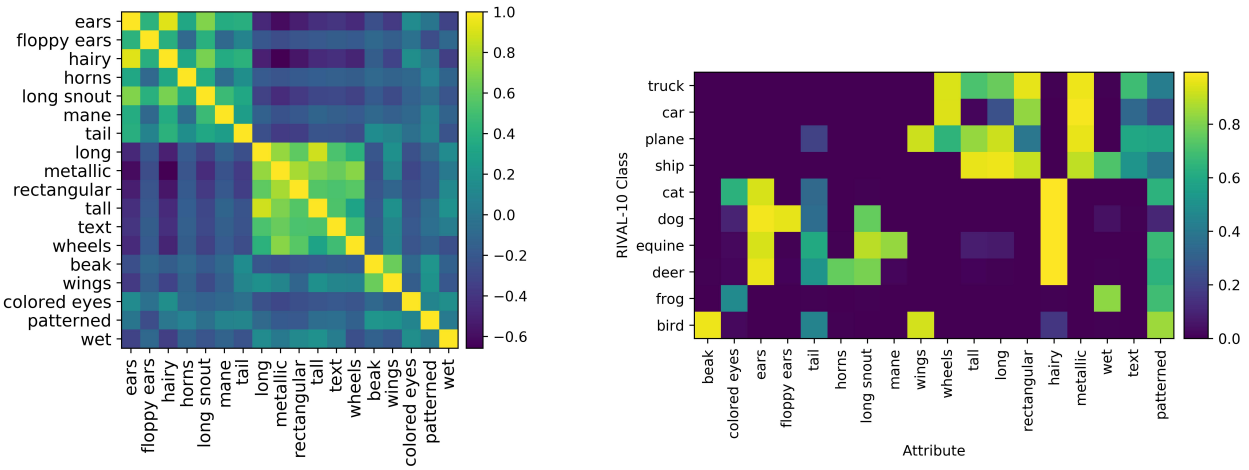

Figure 3. (**Left**): Correlations between attributes in the training split. (**Right**): Class-wise means of attribute vectors in the training split.

equine, frog, plane, ship, truck. We collected the following attributes for these object categories: *beak, colored-eyes, ears, floppy-ears, hairy, horns, long, long-snout, mane, metallic, patterned, rectangular, tail, tall, text, wet, wheels, wings*. Some attributes were inspired from [41].

We chose attributes to be intuitively informative, capturing semantic concepts that humans may allude to in classifying RIVAL10 objects. While the attributes contain some redundant information, they are nonetheless discriminative in the sense that a linear classifier on attributes achieves 93.3% test accuracy. We visualize attribute correlations and class-wise frequencies in Figure 3.

3.2. Data Collection

All images were sourced from ImageNet [9]. The images used in each RIVAL10 class were derived from pairs of related ImageNet classes. In other words, 20 classes from Imagenet were used to build the 10 RIVAL10 classes (details in appendix). To collect our attributes and localizations, we hired workers from Amazon Mechanical Turk (AMT). Data collected through AMT without careful control may be of low quality. To encourage quality annotations, we utilize strategies recommended by the HCI community [31]: providing detailed instructions, *screening* workers for aptitude, and monitoring worker performance with attention checks.

Binary attributions were collected first. Workers were required to pass a qualification test of 20 images with known ground truth attributes: only workers who achieved a minimal overall precision and recall of 0.75 were hired for full data collection. Because the task of segmentation is more involved than indicating whether or not an attribute is present, we required a second qualification test, assessing annotation quality by computing intersection-over-union (IOU) of the submitted attribute masks with ground truth masks. Workers were required to complete five segmentations with an average IOU of at least 0.7.

To ensure that quality is maintained in both the attribution and segmentation phases, roughly 5% of images provided to workers to annotate already had ground truth labels. These so-called attention checks allowed for the monitoring of annotation quality during the collection process. In the first stage of collecting binary attribute labels, the average precision and recall scores were 0.81 and 0.84 respectively. For each positive instance of an attribute marked in the first phase of data collection, an attribute segmentation was collected in the second phase. Completeing attribute segmentations in a second pass allowed for the review of the binary attributions and the removal of any false positives. Average IOU of attention checks completed during the second phase of data collection was 0.745. Further details on our data collection pipeline, including images of instructions shown to workers, payments, and quality-assurance metrics, can be found in the appendix.

4. Models

In our analyses, we focus on ResNets and Vision Transformers [10, 18]. We inspect ResNets trained (i) in a standard supervised fashion, (ii) adversarially via ℓ_2 projected gradient descent [30], and (iii) contrastively (i.e. no direct label supervision), with SimCLR and CLIP [7, 39]. We also consider CLIP Vision Transformers, as well as standard Vision Transformers (ViT) and Data efficient Image Transformers (DeiT) [51]. DeiTs differ from ViTs primarily in their training set, solely using ImageNet-1k while ViTs used ImageNet-21k. To make up for not having the inductive biases of ResNets, ViTs increased the amount of training data, while DeiTs instead rely upon extensive augmentation. All other models, with the exception of those trained with CLIP, use ImageNet-1k as the pretraining set. CLIP, on the other hand, uses a much larger dataset of images and associated text. A full discussion on models is offered in the appendix.

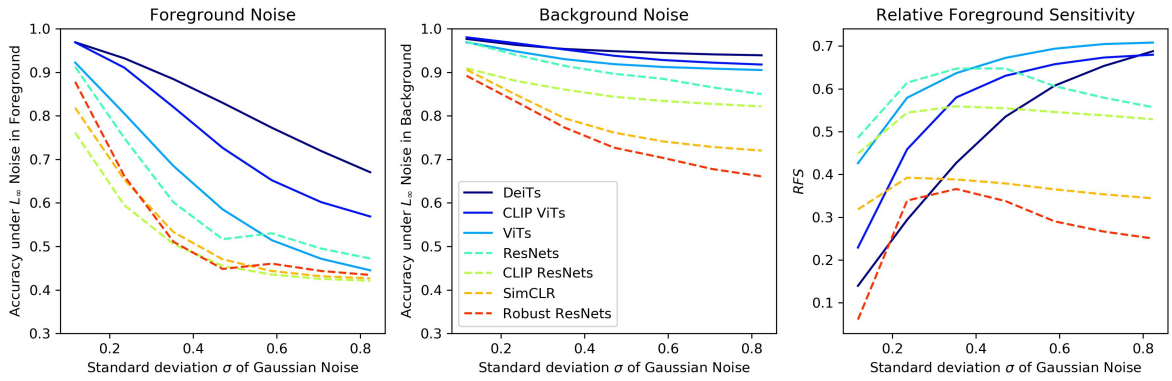

Figure 4. Accuracy under noise in foreground (**left**) and background (**middle**) at various noise levels. Models are grouped by architecture and training procedure, with a curve corresponding to the average over all models in a group. (**Right**): RFS by group.

To perform classification on RIVAL10 dataset, we append a linear head to the penultimate layer of each base model. Only the linear head is fine-tuned on RIVAL10's train split (i.e. other weights are frozen), so to preserve the feature space learned in the original pretraining. All models achieve upwards of 90% accuracy on the RIVAL10 test set, essentially controlling for classification ability. We note that while there is leakage between ImageNet-1k and the RIVAL10 test set, the purpose of this study is not to improve model's predictive accuracy directly, but instead to better understand the information used in making predictions.

Recently, a number of works compare the robustness of ViTs to ResNets. While there are mixed findings on adversarial robustness [4, 44], there is agreement that ViTs have stronger out-of-distribution generalization, likely due to self attention [5, 37]. In contrast, our work focuses on relative robustness to noise in foreground and background regions.

5. Foreground and Background Sensitivity

5.1. Noise Analysis

We add noise to the foreground and background separately to see how corrupting each region degrades model performance. Consider a sample $\mathbf{x}$ with a binary object mask $\mathbf{m}$ where $\mathbf{m}_{i,j} = 1$ if the pixel $\mathbf{x}_{i,j}$ is a part of the object. We first construct a noise tensor $\mathbf{n}$ that has pixel values drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$, where σ is a parameter controlling the noise level. Then, we obtain noisy-background $\tilde{\mathbf{x}}_{bg}$ and noisy-foreground $\tilde{\mathbf{x}}_{fg}$ samples as:

$$\tilde{\mathbf{x}}_{fg} = \text{clip}(\mathbf{x} + \mathbf{n} \odot \mathbf{m}),\ \tilde{\mathbf{x}}_{bg} = \text{clip}(\mathbf{x} + \mathbf{n} \odot (\mathbf{1} - \mathbf{m}))$$

where $\odot$ is the hadamard product, and 'clip' refers to clipping all pixel values to the $[0, 1]$ range. We add Gaussian noise so to preserve the image content. Note that additive pixel-wise noise leads to the same magnitude of perturbation in the foreground and background under the ℓ_∞ norm. We also repeat our analysis with ℓ_2 normalized noise (presented in the appendix) to avoid a bias against larger regions and obtain similar results.

We seek to quantify the sensitivity of a model to foregrounds relative to its sensitivity to backgrounds. To this end, we introduce *relative foreground sensitivity* (RFS). Let a_{fg} and a_{bg} denote accuracy under noise in the foreground and background, respectively, and $\bar{a} := (a_{fg} + a_{bg})/2$ denote their mean (referred to as the general noise robustness). We then define RFS for a model $\mathbf{F}$ as

$$RFS(\mathbf{F}) = \frac{a_{bg} - a_{fg}}{2\min(\bar{a}, 1 - \bar{a})}.$$

Essentially, RFS normalizes the gap in model performance under foreground and background noise by the total possible gap, given the general noise robustness of the model. In Figure 2, RFS takes on the geometric meaning of the ratio between the distance of (a_{fg}, a_{bg}) to $(\bar{a}, \bar{a})$, to the largest possible distance from the diagonal in the unit square for a point with general noise robustness $\bar{a}$. The scale factor in the denominator gives RFS a range of $[-1, 1]$.

We also consider an instance-wise version, $iRFS$, defined for a model $\mathbf{F}$ *and* a sample $\mathbf{x}$. Here, we use the probability that model $\mathbf{F}$ predicts sample $\mathbf{x}$ to belong to its true class as the measure of model performance instead of accuracy. Let p_{fg} and p_{bg} denote this probability for $\tilde{\mathbf{x}}_{fg}$ and $\tilde{\mathbf{x}}_{bg}$, respectively. Thus, with $\bar{p} := (p_{fg} + p_{bg})/2$,

$$iRFS(\mathbf{F}, \mathbf{x}) = \frac{p_{bg} - p_{fg}}{2\min(\bar{p}, 1 - \bar{p})}.$$

In our experiments, we consider seven equally spaced noise levels from $\sigma = 30/255$ to $210/255$. For each sample in the test set, we take ten trials of adding noise to the foreground and background separately *per noise level*. RIVAL10's test set consists of roughly $5k$ images, so for each model type, we assess $5k \times 7 \times 10 = 350,000$ trials in total.

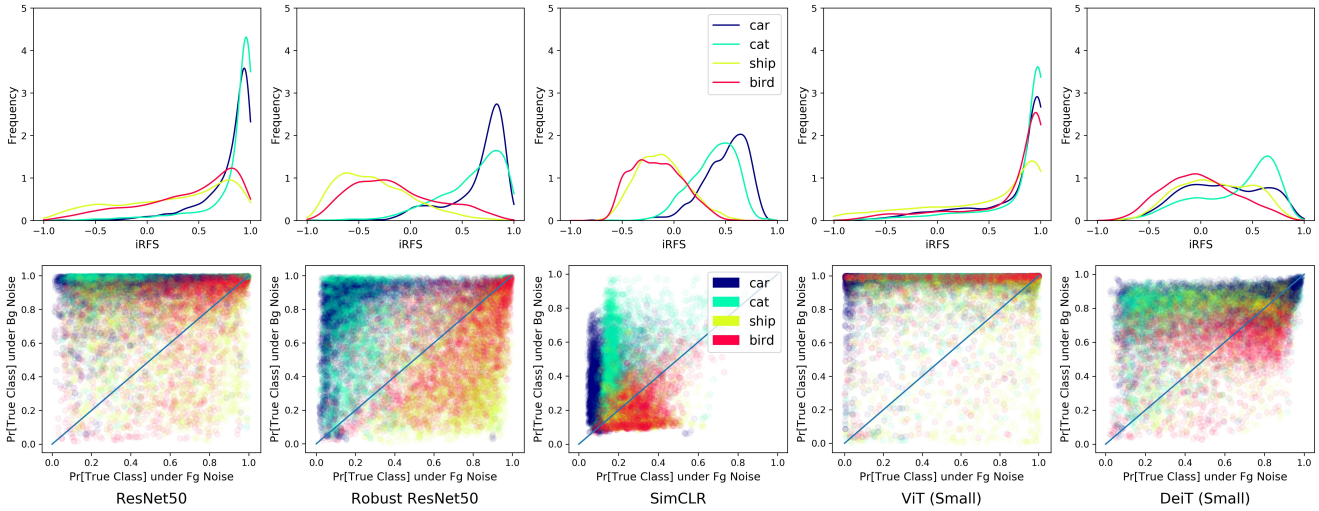

Figure 5. Relative foreground sensitivity per instance for four classes and five models of roughly equal size. (**Top**): Histogram of $iRFS$; positive denotes greater foreground sensitivity. (**Bottom**): Scatter; top left indicates high relative foreground sensitivity. Across models, ships and birds have low foreground sensitivity, often being more sensitive to noise in the background than the foreground.

5.2. Empirical Observations

Fig. 2 shows different models have vastly different performance in terms of both general noise robustness and relative foreground sensitivity. In Figure 2, transformers generally lie further up the main diagonal than ResNets, corroborating observations that transformers are more robust to common corruptions [32]. Increasing model size improves general robustness, though it does so more for transformers than ResNets. Models lie at different distances orthogonal to the diagonal as well, indicating architecture and training procedure affect relative foreground sensitivity.

In Figure 4, we categorize model types based on architecture and training procedure, averaging RFS over groups to reveal general trends. **Robust ResNets have the lowest** RFS, much lower than standard ResNets, a somewhat surprising result given that background reliance has been thought to be linked to increased adversarial vulnerability in the past [53,56]. SimCLR has the next lowest RFS overall, and generally, **contrastive training procedures (CLIP, SimCLR) seem to reduce** RFS in both ResNets and ViTs.

In comparing transformers to ResNets overall, we see at low noise levels, transformers sometimes have lower RFS than ResNets. Interestingly, **as noise level rises,** RFS **in transformers increases** as well, while RFS is mostly stable for ResNets. This suggests that transformers can adaptively alter the attention paid to different image regions based on the level of corruption. Comparing between transformers, we see DeiTs with much lower RFS than ViTs, suggesting that the heavy augmentations DeiTs leveraged to achieve increased data efficiency may have also made the models much more sensitive to backgrounds.

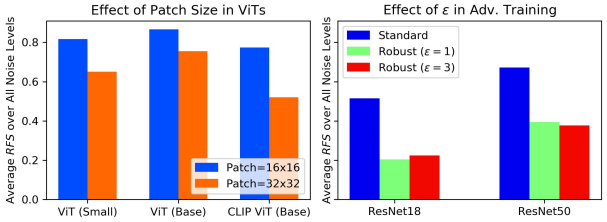

Figure 6. Controlled ablation studies. Average RFS over all noise levels presented for brevity. (**Left**): Increasing patch size in ViTs decreases relative foreground sensitivity. (**Right**) Robust models are much less relatively sensitive to foregrounds, but ϵ used in adversarial training does not affect RFS much.

In Figure 6, we more closely inspect the effect of patch sizes in ViTs and the attack budget ϵ used in adversarial training (which affects accuracy-robustness trade-off). We find that increasing the patch size in ViTs from 16×16 to 32×32 reduces RFS when averaged over all noise levels. The robustness ablation affirms that robust ResNets are much less relatively sensitive to foregrounds than standard ResNets, though the attack size seen in training does not seem to significantly affect RFS.

Moving away from comparing models, in Figure 5, we see **foreground sensitivity is largely affected by class**. In particular, across models of roughly equal size, ships and cats are often more sensitive to background noise, suggesting models learn to utilize background content more than foreground content in recognizing them. The class distinction is less pronounced in DeiTs and ViTs, with ViTs assigning high foreground sensitivity for all classes, and DeiTS having mixed sensitivity across classes, with many negative $iRFS$ scores (i.e. higher background sensitivity).

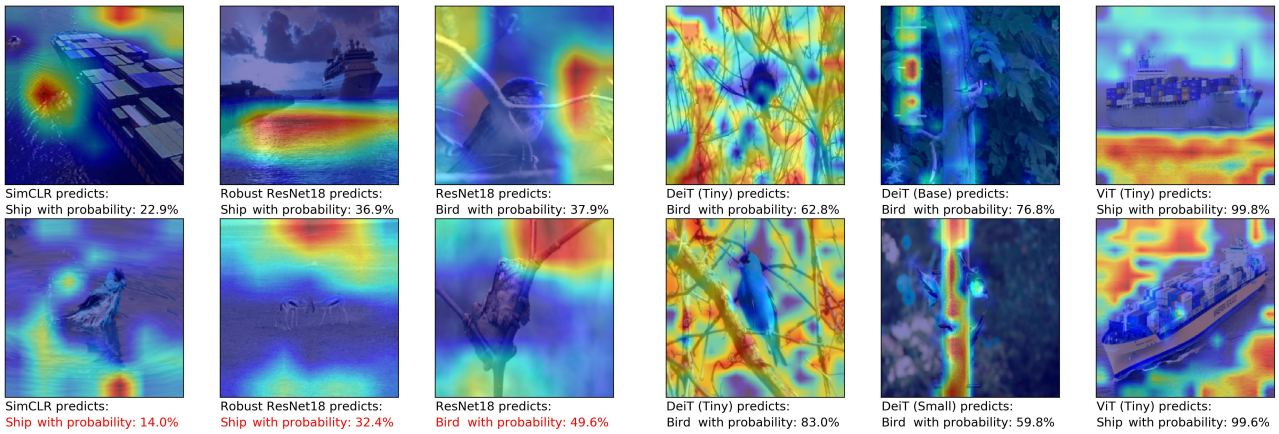

Figure 7. Instances of images with low saliency alignment, highlighting spurious features of water for ships, and branches and bird feeders for birds. (**Left**): Spurious features leading to misclassification (in red). (**Right**): Other instances of spurious features.

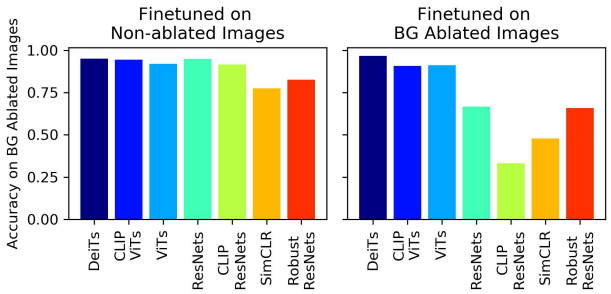

Figure 8. Model accuracy on images with backgrounds ablated via graying. The right plot shows accuracies for models finetuned on the images with backgrounds ablated. Only transformers can fit a linear layer on the features of background ablated images without compromising performance.

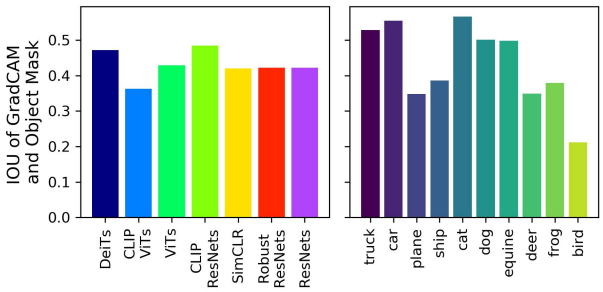

Figure 9. Alignment of binarized saliency maps with object segmentation masks, measured by intersection over union (IOU). Averaged over models (**left**) and object classes (**right**).

5.3. Removing Backgrounds Entirely

We also inspect the accuracy of models on images with backgrounds grayed out, similar to [56], though now considering ViTs, CLIP, and SimCLR, which had not been developed at the time of their study. Also, the rich annotations of RIVAL10 allow for going beyond foreground or background ablation (see the appendix for a discussion of attribute removal). Ablation via graying can be thought of as another kind of noise, where all pixels are smoothed to 0.5. In Figure 8, the left plot reveals that Robust ResNets and SimCLR see the largest drops in accuracy when evaluated on images with grayed backgrounds. Transformers do well on ablated images, consistent with the observation that transformers had high RFS at the largest noise levels. Furthermore, when we attempt to fit a linear layer to classify background-ablated images, only the features from transformer models are sufficiently informative to have high linear classification accuracy. Thus, while transformers make use of backgrounds, they still retain significant foreground information in their feature space. This result suggests transformers are much more robust to localized distribution shift. That is, distribution shift in one region (the background) may affect model perception of other unperturbed regions much less in transformers than ResNets.

5.4. Saliency Alignment

To complement the noise analysis, we use GradCAM [43] to assess the amount of saliency that models place on foreground pixels. RIVAL10's object segmentations allow us to automatically quantify saliency alignment with foregrounds, removing the need for human inspections. Sorting samples based on saliency alignment reveals failure modes, where models deem background regions as highly salient. We present several metrics to assess saliency alignment in the appendix. We find that extracting samples with the lowest difference in average pixel saliency in foreground and background yield the most interesting failure modes. We present examples selected this way in Figure 7, highlighting spurious background features that contribute to the low RFS of ships and birds observed across models in Figure 5. Specifically, models look for water and coasts when classi-

fying ships, and twigs and branches when classifying birds.

In Figure 9, we appeal to the standard metric intersection-over-union (IOU). Saliency maps are binarized using a threshold of 0.5 before being compared with object segmentation masks. On average, saliency alignment is similar across models, despite their being large differences in RFS identified in the noise analysis, suggesting saliency maps may give an incomplete picture of model sensitivity. In comparing saliency alignment across classes, we see much larger differences, emphasizing the result that **class matters** when it comes to background reliance.

6. Neural Node Attribution Analysis

The attribution of features in a neural network is a fundamental problem in modern machine learning work. Saliency, when computed with respect to a given feature, is a prominent approach for doing so [11, 24, 43, 49]. Although many works make claims of attribution based on saliency, to the best of our knowledge, quantitative validation is rarely given [60]. Here we propose to quantitatively evaluate node attribution via saliency through comparison with the ground-truth attribute localization in RIVAL10.

We propose the following procedure. Given a pretrained robust ResNet50 feature extractor and a class label, we identify the top 10 training images by activation with that label for each component in the feature layer (the penultimate layer). We then compute saliency using GradCAM at each neural feature on these top-10 images, and compare them against ground truth attribute localization. Saliencies are binarized at max-normalized threshold of $\tau = 0.5$. The intersection-over-union (IOU) with the ground truth attribute localization is then computed for each sample, and finally averaged. This obtains a score, which we interpret as measuring the quality of neural feature attribution based on saliency alignment to the attribute segmentations of the top-10 images. We then select the neural feature with highest alignment per attribute, identifying these features as the best candidates for node attribution. Note that searching by top IOU is only possible with ground truth attributes and localizations, as is the case with RIVAL10.

Next, we check if these neural features generalize to held-out data not used in the analysis, namely the test set of RIVAL10. Here we analyze one class-attribute pair and show additional results in the appendix. We visualize the GradCAMs of top testing samples with respect to the top features identified in the training set in Figure 10. We observe visually that the saliencies align well with the given attribute on these samples. We then compute the IOU scores on *all* images in the test set with the given class and attribute labels. We plot this histogram in Figure 10. We observe that IOU values are on average high (> 0.5) indicating that the neural features generalize well to held-out data for considered cases. We note that this analysis is just one approach for quantitatively evaluating feature attribution. We stress the importance of quantitative measurements rather than relying on just visualization, and envision that our RIVAL10 dataset may help refine the discourse around feature attribution.

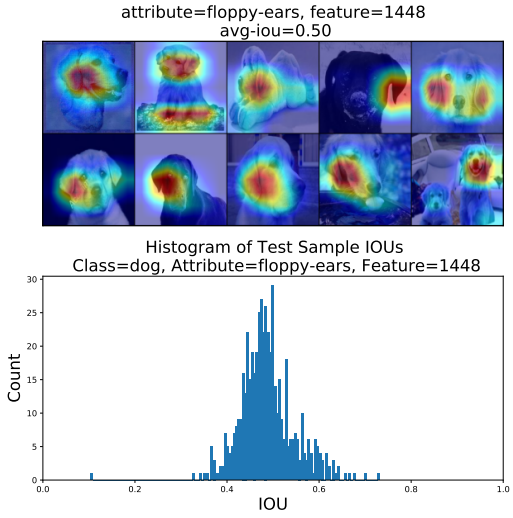

Figure 10. (**Top**): Example GradCAMs on test images with respect to the top feature identified by IOU in training set. (**Bottom**): Histograms of IOUs corresponding to this feature, attribute pair.

7. Conclusion

We present **RI**ch **V**isual **A**ttributes with **L**ocalization (RIVAL10), and quantitatively assess sensitivities of state-of-the-art models under noise corruption. Specifically, we find adversarially or contrastively training ResNets leads to reduced relative foreground sensitivity. Further, we observe that transformers adaptively raise foreground sensitivity as noise level increases, while ResNets do not. By applying automated alignment metrics to saliency maps, we reveal instances of spurious background features used by models. Lastly, we observe promising evidence that neural node attributions based on top activating images generalize to instances unseen during attribution. We hope RIVAL10's rich annotations lead future studies to gain new quantifiable insights on the behavior of deep image classifiers.

8. Acknowledgements

This project was supported in part by NSF CAREER AWARD 1942230, a grant from NIST 60NANB20D134, HR001119S0026 (GARD), ONR YIP award N00014-22-1-2271, Army Grant No. W911NF2120076 and an AWS Machine Learning Research Award.

References

[1] Andreas Geiger Axel Sauer. Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua B. Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019.

[3] Samyadeep Basu, Phillip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *CoRR*, abs/2006.14651, 2020.

[4] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *arXiv preprint arXiv:2110.02797*, 2021.

[5] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *CoRR*, abs/2103.14586, 2021.

[6] Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. *arXiv preprint arXiv:2106.01127*, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[11] Gabriel G. Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning explainable models using attribution priors. *CoRR*, abs/1906.10670, 2019.

[12] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[14] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE, 2009.

[15] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020.

[16] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, Jul. 2019.

[17] Sven Gowal, Chongli Qin, Po-Sen Huang, A. Taylan Cemgil, Krishnamurthy Dvijotham, Timothy A. Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. *CoRR*, abs/1912.03192, 2019.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[22] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge J. Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. *CoRR*, abs/2004.12276, 2020.

[23] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. *CoRR*, abs/2012.04104, 2020.

[24] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[25] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2020.

[26] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[27] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. cite arxiv:1405.0312Comment: 1) updated annotation pipeline

description and figures; 2) added new section describing datasets splits; 3) updated author list.

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[31] Tanushree Mitra, C.J. Hutto, and Eric Gilbert. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1345–1354, New York, NY, USA, 2015. Association for Computing Machinery.

[32] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021.

[33] 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. *CoRR*, abs/1602.03616, 2016.

[34] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

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

[36] Wanli Ouyang, Hongyang Li, Xingyu Zeng, and Xiaogang Wang. Learning deep representation with large-scale attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[37] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners, 2021.

[38] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[40] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[41] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In *European Conference on Computer Vision*, pages 1–14. Springer, 2010.

[42] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019.

[43] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[44] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *CoRR*, abs/2103.15670, 2021.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning?, 2021.

[47] 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.

[48] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[49] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017.

[50] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020.

[52] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *CoRR*, abs/2106.00545, 2021.

[53] Tianlu Wang, Diyi Yang, and Xuezhi Wang. Identifying and mitigating spurious correlations for improving robustness in nlp models, 2021.

[54] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[55] Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11205–11216. PMLR, 18–24 Jul 2021.

[56] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *CoRR*, abs/2006.09994, 2020.

[57] Jessica Yung, Rob Romijnders, Alexander Kolesnikov, Lucas Beyer, Josip Djolonga, Neil Houlsby, Sylvain Gelly, Mario Lucic, and Xiaohua Zhai. Si-score: An image dataset for fine-grained analysis of robustness to object location, rotation and size. *CoRR*, abs/2104.04191, 2021.

[58] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[59] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

[60] Yilun Zhou, Serena Booth, Marco Tílio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *CoRR*, abs/2104.14403, 2021.