A SAMPLE-BASED METHOD FOR SEMANTIC UNDERSTANDING OF NEURAL NETWORK DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability in deep learning is one of the largest obstacles to its more widespread adoption in critical applications. A variety of methods have been introduced to understand and explain decisions made by Deep Models. A class of these methods highlights which features are most influential to model predictions. These methods have some key weaknesses. First, most of these methods are applicable only to the atomic elements that make up raw inputs to the model (e.g. pixels or words). Second, these methods generally do not distinguish between the importance of features individually and their importance due to interactions with other features. As a result, it is difficult to explore high-level questions about how models use features during decision-making. We tackle these issues by proposing Sample-Based Semantic Analysis (SBSA). We use Sobol variance decomposition as our sample-based method which allows us to quantify the importance of semantic combinations of raw inputs and highlight the extent to which these features are important individually as opposed to due to interactions with other features. We demonstrate the ability of Sobol-SBSA to answer a richer class of questions about the behavior of Deep Learning models by exploring how CNN models from AlexNet to DenseNet use regions when classifying images. We present two key findings. 1) The architectural improvements from AlexNet to DenseNet manifested themselves in CNN models utilizing greater levels of region interactions for predictions. 2) Adversarially robust CNNs resist exploiting spurious correlations in ImageNet data by forcing these architectures to rely less on region-to-region interaction. Our proposed method is generalizable to a wide variety of network and input types and can help provide greater clarity about model decisions.

1 INTRODUCTION

Deep learning models are becoming endemic in various applications. As models are increasingly used for critical applications in medicine such as detecting lung nodules (Schultheiss et al., 2021) or autonomous driving (Li et al., 2021), it is important to either create interpretable models or to make opaque models human interpretable. This paper focuses on the latter. Existing methods developed over the last decade for doing this can be broken down into model agnostic vs model dependent. Model agnostic methods, such as Shapley values (Kononenko et al., 2013) and Integrated Gradients (Sundararajan et al., 2017) weigh the importance of input features without relying on the structure of the model. In contrast, methods such as GradCam (Selvaraju et al., 2017) and GradCam++ (Chattopadhay et al., 2018) are heavily dependent on model architecture.

While both of these methods yield valuable information about models, they share two common gaps. First, neither of them reports metrics that distinguish between the features in input space that are individually important and features that are important because of their interaction with other features. Second, the above methods are generally applied to inputs at their most granular level (pixels, words, etc..). The combination of these two gaps limits the conclusions that Machine Learning practitioners can make about individual model predictions, as well the behavior of models as a whole.

We address these limitations in two key ways. First, we introduce a two-part framework called Sample-Based Semantic Analysis (SBSA). The first part of the framework is a function that generates semantic representations of inputs and is able to map those semantic combinations to real numbers. The second part of the framework is a black-box sample-based sensitivity method. In

this case, the Sobol method which reports the importance of individual features and their interactions. Second, we demonstrate the ability of Sobol-SBSA to answer a richer set of questions than standard interpretability methods by applying it to CNN models in the context of ImageNet. We ask two particular questions. 1) How did CNN model's use of regions evolve as architectures were developed from AlexNet to DenseNet? 2) How do adversarially training robust versions of these models change the manner in which these architectures utilize regions such that they are less likely to exploit spurious correlations?

The key results and contributions of this paper are as follows:

1. We present a general-purpose framework for using sample-based sensitivity methods to analyze the importance of semantic representations of inputs and test it using a variety of black-box methods.
2. We demonstrate that the Sobol method outperforms other popular black box methods, Integrated Gradients, Shapley (Kernel Shap), and LIME, for selecting both the most and least important regions to CNN predictions.
3. We show, through direct measurement, that the main impact of the evolution of CNN architectures was primarily to increase the extent to which they used region interactions. Similarly, we show that adversarially robust versions of CNNs are less susceptible to spurious correlations because they force architectures to rely less on region interactions.

2 Related Work

In this section, we discuss prior work in three areas relevant to SBSA. Interpretability methods that use interactions, the extent to which different iterations of CNN models utilize interactions, and prior work done on analyzing how different iterations of CNN models use regions. All four of these areas represent gaps in interpretability literature that have recently seen a strong surge in interest.

In recent years, interest has grown in interpretability methods that can quantify the importance of feature interactions.Janizek et al. (2020) introduced an extension to Integrated Gradients, called Integrated Hessians, that quantified the importance of pairwise interactions of features, while Zhang et al. (2020) built on this work by presenting a similar quantification through Shapley Interactions. While both methods are powerful, they are expensive. These methods require the separate calculation of individual feature importance and interactions, as well as pairwise comparisons in order to quantify interactions.

The work most closely related to ours was done by Fel et al. (2021), who addressed the issue of additional computational expense by using the Sobol method to calculate the importance of regions in ImageNet models. They demonstrated that Sobol outperformed white box methods like Grad-CAM and another black box method, RISE (Petsiuk et al., 2018). However, they did not perform comparisons against other popular black box attribution methods like Integrated Gradients (Sundararajan et al., 2017), Shapley (Kononenko et al., 2013), and Local Interpretable Model-Agnostic Explanations, LIME (Ribeiro et al., 2016). Additionally, they did not demonstrate how the Sobol measurement of both interactive and primary feature importance could be used to answer larger overarching questions about models. Finally, they did not explore how the assumption that input features to a model are uncorrelated limits the circumstances under which Sobol can be used. We address all of these issues in our work.

To the author's knowledge, limited work has been done on exploring how region interactions evolved in CNN models. Brendel & Bethge (2019) inferred region interaction in VGG, ResNet, and DenseNet based on the Pearson correlation between a perturbed measure of a CNN's output that assumes region independence, and one that does not. They found that ResNet and DenseNet made stronger use of region interactions than VGG. While the result was powerful, it relied on a heuristic, rather than a direct measurement. Additionally, they did not explore how these interactions changed for robust versions of these architectures.

Robust CNNs were developed in order to combat the shortcomings of typical CNNs. It is well known that CNN models applied to ImageNet at times exploit correlations in the data to achieve higher accuracy in classification metrics. The result of this is that these CNN models don't always transfer ideally to other tasks. Salman et al. (2020) demonstrated that when these models are trained to be

adversarially robust they transfer to other tasks more effectively. Singla et al. (2022) showed that one of the reasons that the robust models transferred more effectively is because they were less susceptible to exploiting spurious correlations in the data when making predictions. They demonstrate this by adding noise to spurious regions of images using Salient ImageNet, a dataset that contains masks that indicate which areas in the image correspond to core regions and which correspond to spurious regions (Singla & Feizi, 2021). Although they demonstrated that Robust Models were less susceptible to exploiting spurious regions, they did not show *why* this was the case. *Sobol-SBSA tells us the mechanism by which Robust models are less susceptible to spurious correlations than their normal counterparts*.

In this paper, we will present experiments that build upon prior work in order to present a framework for interpretability that can answer a rich set of questions regarding how models utilize semantic representations of inputs.

3 METHODOLOGY

In this section, we describe the two components of SBSA and specify how we use it to analyze the importance of image regions in ImageNet. In particular, we describe how we semantically map image regions to quantities that can be analyzed with a sampling-based method, and the specifics of Sobol as a sampling-based sensitivity method.

3.1 SAMPLE-BASED SEMANTIC ANALYSIS (SBSA)

Let us define the following variables. $x \in \mathbf{R}^d$ is an input to a model, $f : x \rightarrow y \in \mathbf{R}^k$ is a model that takes x as an argument and produces y, $\tilde{x}_i \in \mathbf{R}^d$ is a sample of x, and $N \in \mathbf{Z}$ is a prescribed integer that helps to determine the number of $\tilde{x}_i$ samples generated. Most sample-based sensitivity methods operate by generating a number of samples that is some function of N and d. For example, $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{Nd}$. The model is then evaluated on these samples yielding $f(\tilde{x}_1), f(\tilde{x}_2), \ldots, f(\tilde{x}_{Nd})$. Sample-based sensitivity methods, such as Sobol, use these outputs to determine the importance of components of x to the model output, y.

One thing that immediately becomes clear is that for deep learning applications with high-dimensional inputs, such as images, videos, and long documents, applying this process to raw inputs is prohibitively expensive. This issue can be greatly minimized if one turns to semantic representations of inputs instead. In this paper, a semantic representation of an input is defined as follows. **A semantic representation of an input, x, is some combination of the raw components of that input which yields a human recognizable higher order feature, such as the colors or regions of an image or grammatical parts of sentences**. We define this semantic representation as $S \in \mathbf{R}^m$, where $m < d$. Recalling that most sample-based sensitivity methods operate on real numbers, we define two mapping objects.

$$G : x \in \mathbf{R}^d \rightarrow S \in \mathbf{R}^m, \quad G^{-1} : S \in \mathbf{R}^m \rightarrow x \in \mathbf{R}^d \quad m < d \tag{1}$$

$$H : S \in \mathbf{R}^m \rightarrow r \in \mathbf{R}^l, \quad H^{-1} : r \in \mathbf{R}^l \rightarrow S \in \mathbf{R}^m \quad l < m \tag{2}$$

G maps the raw input, x, to some lower dimensional semantic representation S, and H maps the semantic representation to some lower dimension vector of real numbers, r. Both mappings are invertible. Given these mappings, SBSA generates samples of r, $\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_{ld}$. From these samples, samples of the original input are generated by first mapping these samples of real numbers to samples of the semantic representations of x, $\tilde{S}_i$, and then mapping these samples of semantic representations to samples of the raw inputs, $\tilde{x}_i$.

$$\{H^{-1}(\tilde{r}_1), H^{-1}(\tilde{r}_2), \ldots, H^{-1}(\tilde{r}_{Nd})\} = \{\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_{Nd}\} \tag{3}$$

$$\{G^{-1}(\tilde{S}_1), G^{-1}(\tilde{S}_2), \ldots, G^{-1}(\tilde{S}_{Nd})\} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{Nd}\} \tag{4}$$

The model is then evaluated on the samples of the raw input, $\tilde{x}_i$. The key here is that since the sampling was done in r, the sensitivity analysis reports the importance of different components of the semantic representations of the input, S. In the following section we describe the use of SBSA

via image regions, however, there exist methods of automatically discovering semantic regions that could be used for different applications (Koh et al., 2020).

3.2 Sample-Based Semantic Analysis (SBSA) Applied to ImageNet and Regions

In this section, we illustrate how we perform sensitivity analysis of different regions in ImageNet models based on the SBSA framework. We will first discuss our mapping functions. Then we will detail the Sobol method for sensitivity analysis and the measures of importance that it produces.

Given an image $x \in \mathbf{R}^d$ where $d = W \times L$, our input to semantic mapping function, G, extracts regions of x. The mapping, H, from the semantic representations to a reduced real number vector, r, is simply the sum of the value of pixels in the region.

$$G : x \in \mathbf{R}^d \rightarrow \{S_1, \ldots, S_l\} \in \mathbf{R}^{l \times w \times h}, \quad S_j = x_{o:p,r:s} \in \mathbf{R}^{w \times h} \tag{5}$$

$$H : \{S_1, \ldots, S_l\} \in \mathbf{R}^{l \times w \times h} \rightarrow \{r_1, \ldots, r_l\} \in \mathbf{R}^l, \quad r_j = \sum_{t=o}^{p} \sum_{u=r}^{s} x_{tu} \tag{6}$$

Finally, the inverse mapping, H^{-1} involves normalizing each region, $\{x_{o:p,r:s}\}_j$ by the sum of the values of the pixel in that region, $\sum_{t=o}^{p} \sum_{u=r}^{s} \{x_{tu}\}_j$, and multiplying it by r_j. Then assigning the relevant indices in x the new values in the region. When r_j is equal to the summation of pixels in that region from the original image, this simply recovers the original image. When sampled, this reconstructs samples, $\tilde{x}_i$. Mathematically, the inverse mapping is as follows.

$$H^{-1} : \{r_1, \ldots, r_l\} \in \mathbf{R}^l \rightarrow \{S_1, \ldots, S_l\} \in \mathbf{R}^{l \times w \times h}, \quad S_j = r_j \frac{x_{o:p,r:s}}{\sum_{t=o}^{p} \sum_{u=r}^{s} x_{tu}} \tag{7}$$

$$G^{-1} : \{S_1, \ldots, S_l\} \in \mathbf{R}^{l \times w \times h} \rightarrow x \in \mathbf{R}^{W \times L} \tag{8}$$

Given these mappings, we apply the Sobol Sensitivy method to the low dimensional real vectors, r_j. Here we will briefly give an overview of the Sobol method. For more details about the implementation of Sobol, see Saltelli et al. (2010) and Herman & Usher (2017). The Sobol method is a variance-based sensitivity method. Given a model, $f : (x_1, \ldots, x_d) \in \mathbf{R}^d \rightarrow y$, the variance based first order effect of a component of the input, x, is:

$$V_i = V_{x_i}(E_{\mathbf{x}\sim i}(y_j|x_i)) \tag{9}$$

where y_j is a component of the model output, x_i is a component of the input, x, $\mathbf{x}_{\sim i}$ are samples of the input where all components of the input except for i are varied, and V_{x_i} and $E_{\mathbf{x}\sim i}$ are the variance and expectation over x_i and $\mathbf{x}_{\sim i}$ respectively. The Sobol first-order sensitivity is then the ratio of V_i to the variance of the model output, y_j.

$$S_i = \frac{V_i}{V(y_j)} \tag{10}$$

S_i quantifies the individual importance of the ith component of the input, x, to the jth component of the output, y. This equation essentially states that Sobol first order index, S_i, is the fraction of the variance of the output, y_j that is accounted for by x_i. What makes Sobol unique is that it simultaneously calculates the importance of the interaction between a component of the input, x_i, and all other components of the input. This is called the total effect index and is calculated as:

$$S_{T_i} = 1 - \frac{V_{\mathbf{x}\sim i}(E_{x_i}(y_j|\mathbf{x}_{\sim i}))}{V(y_j)} \tag{11}$$

This equation essentially states that the total effect index for x_i, S_{Ti}, is the difference between 1 and the fraction of the variance of the output, y_j, that is accounted for by all samples of the input in

which every component of the input varies, except for i. An important quantity in Sobol is a ratio that we will refer to as the Primary Index Ratio (PIR):

$$\text{PIR} = \frac{S_i}{S_{Ti}} \tag{12}$$

This quantity reports the extent to which a feature, x_i, is important to the model output due to interaction. The smaller the PIR, the greater the importance of interactions.

In our application to ImageNet, we produce samples of r_j that allow the calculations of the individual and total effect indices. Details about how samples are generated, and the calculation of S_i and S_{Ti} are discussed in depth in Saltelli et al. (2010). We used an open-source implementation of the aforementioned Sobol method from the Sensitivity Analysis Library (SALib) (Herman & Usher, 2017). We note that the number of samples used for the Sobol method is $N(d + 2)$. For all of the experiments that follow, $N = 50$.

4 Experiments

In this section, we demonstrate that Sobol-SBSA outperforms other standard black box methods based on multiple measures of the effectiveness of interpretability methods. Having established this, we utilize Sobol-SBSA to explore how CNNs and their robust counterparts use regions. Finally, we discuss the limitation of the Sobol-SBSA when comparing results across region sizes.

4.1 Comparison of Sobol-SBSA to Other Black Box Models

Our SBSA model can be used with a variety of black box models. Thus, we compare the following two methods against Sobol-SBSA.

- **Shapley (Shap-SBSA) (Kononenko et al., 2013):** Used with SBSA in a similar manner as Sobol-SBSA.
- **LIME (LIME-SBSA) (Ribeiro et al., 2016):** Used with SBSA in a similar manner as Sobol-SBSA.

In addition, we also compare against the following models.

- **Integrated Gradients (Sundararajan et al., 2017):** Cannot be used with SBSA since it requires gradients with respect to pixels. Instead, we aggregate pixel importance in regions and rank region importance based on this aggregation.
- **Random:** A control baseline where regions are randomly selected.

We use two standard metrics for comparing interpretability methods. The first is measuring the change in the ground truth label score of a model when regions, as specified by the sensitivity method, are masked. We mask the top and bottom 20% of regions in an image. The second measurement is sensitivity-*n* correlation. Sensitivity-n correlation is a quantity proposed by Ancona et al. (2017) for comparing the effectiveness of attribution methods when determining which features are important to a model. Quantitatively, it is a measurement of the Pearson correlation between $\sum_{i=1}^{n} R_i^c$ and $S_c(x) - S_c(x_{[x_S=0]})$. $\sum_{i=1}^{n} R_i^c$ is the sum of the attributions associated with the n input features that are masked and $S_c(x) - S_c(x_{[x_S=0]})$ is the difference in the score that the model produces when n input features are masked versus when none are. For each value of i, i random features are selected to be masked 100 times and the correlation is averaged over the examined data. A value closer to one means a more effective method.

For all methods, we evaluate the importance of regions to ResNet50 on 1000 randomly selected ImageNet validation images. To explore whether or not region size would impact our results, we performed our experiments for 4×4, 8×8, and 16×16 regions. The number of samples is selected as a function of the number of regions, as detailed in section 3.2. For 4×4, 8×8, and 16×16 regions, the number of samples used for all methods are 900, 3300, and 12900 respectively.

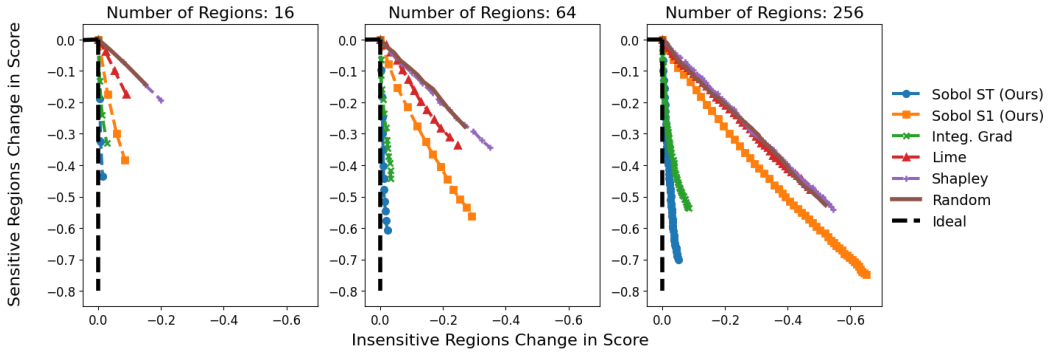


Figure 1: Plots of the change in the ground truth score of the ResNet50 when the top 20% of regions are masked vs the bottom 20% of regions. A steep slope is indicative of a more effective method since it means that the change in ground truth scores are significant when the top 20% of images are masked, but not when the bottom 20% of images are masked. From left to right this is plotted for when the regions are 4×4, 8×8, and 16×16 respectively.

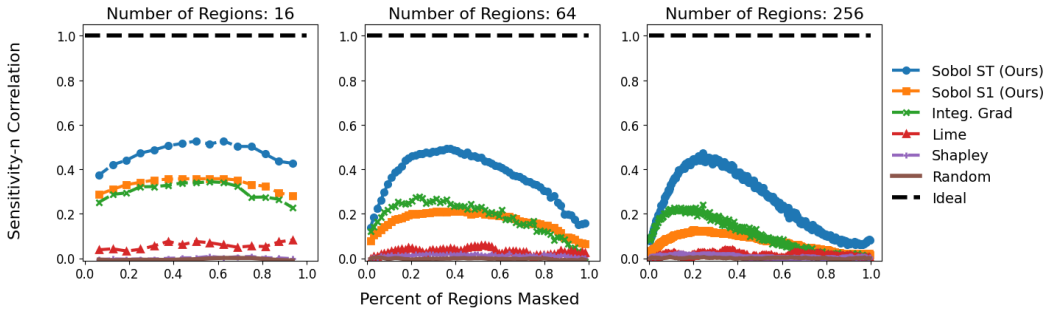
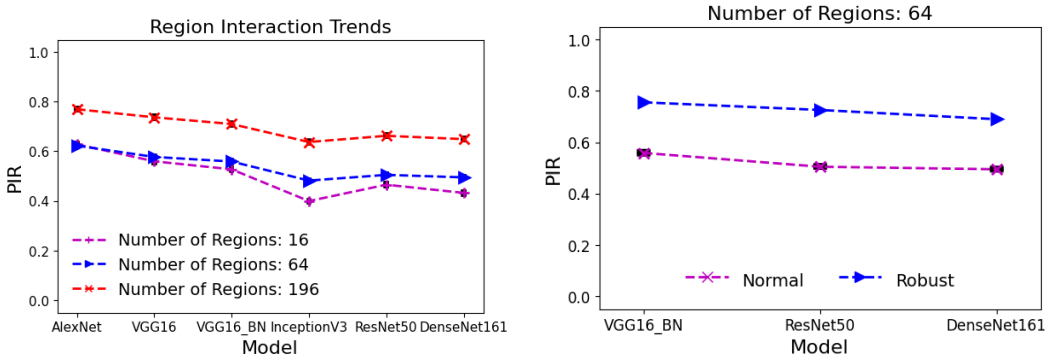


Figure 2: Plots of the sensitivity-n correlation for different sensitivity methods as a function of the percentage of regions masked. The sensitivity methods with values closer to one are more effective. From left to right this is plotted for when the regions are 4×4, 8×8, and 16×16 respectively.

Figures 1 and 2 show the results of the masking and sensitivity-n correlations respectively for all regions. Figure 1 plots the change in the model score when the most important regions are masked versus when the least. A steeper curve is more ideal since it means that the sensitivity method significantly decreases the model score when masking the most important regions, but has little impact when masking the least important regions. *The Sobol-SBSA is the most effective by both the masking and sensitivity-n measures*. The key takeaways from the plot are 1) Both Sobol measurements, S_{Ti} and S_i, are more effective at selecting the most important regions in an image, 2) S_{Ti} performs the best overall in picking the most and least important regions in an image, as well as more properly ranking the regions in between (as measured by sensitivity-n), and 3) S_{Ti} is the most robust to changes in region size, followed closely by Integrated Gradients.

4.2 What can Sobol-SBSA Tell us About CNNs Beyond Region Importance?

We have shown Sobol-SBSA's efficacy at ranking the importance of regions to CNN predictions. While Sobol-SBSA outperforms other black box methods in standard interpretability measures, the real advantage offered by Sobol-SBSA over other black box methods is the ability to answer large overarching questions about the behavior of Deep Learning models. We demonstrate this by using Sobol-SBSA to quantify how the use of regions evolves in CNN models from AlexNet to DenseNet161, and to understand the mechanism that causes adversarially robust CNN models to be less susceptible to spurious correlations in the data.



(a) The plots of the average mean of PIR for 10000 validation images for a series of CNN models when applied to images that are partitioned into different numbers of regions. This is done for the top 20% of regions in an image. The importance is determined by S_T.

(b) The plots of the average mean of PIR for 10000 validation images for normal CNN models and their robust counterparts when applied to images that are partitioned into 8×8 regions. This is also done for the top 20% of regions in an image, as determined by S_T.

Figure 3: PIR trends in CNN models and a comparison of PIR for normal and robust models.

4.2.1 How have CNN Models Evolved their Use of Regions?

To answer the question as to how the use of regions has changed as CNNs have evolved, we use the PIR, defined in equation 12. This ratio quantifies the extent to which a model relies on the interaction of a feature with other features. The smaller this ratio the larger the interactive effect. We apply Sobol-SBSA to AlexNet, VGG16, VGG16_BN, InceptionV3, ResNet50, and DenseNet161. For all models with the exception of InceptionV3, Sobol-SBSA is applied to ImageNet images when the number of regions is 4×4, 8×8, and 14×14. The number of samples for these sizes are 900, 3300, and 9900 respectively. The pre-trained InceptionV3 operates on an image that is 299×299 rather than 224×224. In order to compare similar region sizes, the number of regions for which Sobol-SBSA was applied to InceptionV3 were 5×5, 11×11, and 18×18 with sample sizes of 1350, 6150, and 16300 respectively. All of the results displayed in this section are based on 10000 random samples from the ImageNet validation set. We performed the experiments at 4×4 and 8×8 regions using all 50000 images in the validation set, and saw that the average PIR was nearly identical, figure 6 in the supplementary. We therefore ran the 14×14 region experiments using 10000 random samples in order to save computational cost.

Figure 3a shows the average PIR for AlexNet, VGG16, VGG16_BN, InceptionV3, ResNet50, and DenseNet161 for different region sizes, along with their confidence interval. The average is calculated based on the most important 20% of regions in an image, as measured by S_{Ti} since we have shown its ability to effectively select the least and most important regions for a model. The key takeaway from this figure is that *the results of the architectural changes that yielded modern CNNs was to increase the extent to which models used region interactions when making predictions*.

In particular, there was a significant increase in the importance of region interaction from AlexNet to VGG16, and then from VGG16 to InceptionV3, DenseNet161, and ResNet50. Of all the models, InceptionV3 most strongly relies on interactions. This result provides an explicit quantitative verification of the conclusions reached by Brendel & Bethge (2019)

4.2.2 Why are Adversarialy Robust CNNs Resistant to Spurious Features?

In this section, we explore the mechanism by which Adversarialy robust CNNs are less susceptible to spurious correlations in the data. In order to do this, we apply Sobol-SBSA to robust versions of VGG16_BN, ResNet50, and DenseNet161 and compare the PIR of the robust versions of these architectures to their standard versions. We use pre-trained robust weights from Salman et al. (2020) where the l_2 threat radius was 3. Figure 3b shows the PIR for the normal and robust versions of VGG16_BN, ResNet50. It is immediately clear for all region sizes that the impact of training CNN models robustly is to significantly decrease the extent to which said model utilizes interactions in regions when making decisions. Since robust models reduce the impact of spurious correlations in

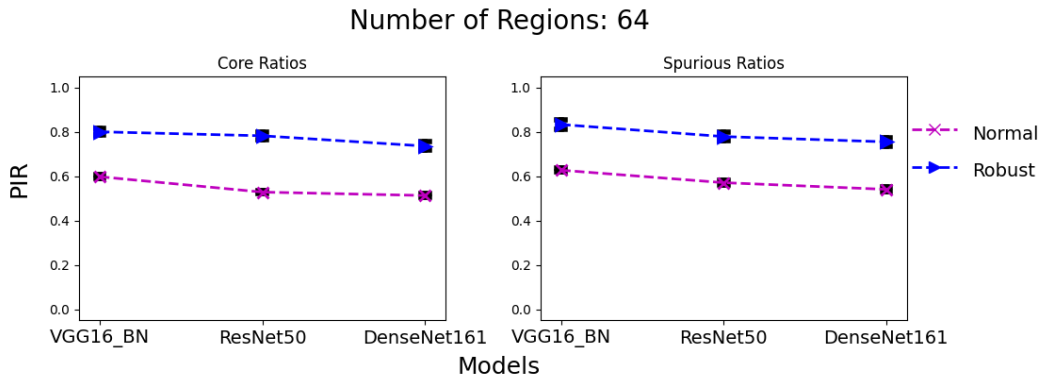


Figure 4: Plots of the average mean of PIR for ResNet50 and RobustResNet50 when decomposed into core and spurious regions.

data on model predictions, it is reasonable to conclude that modern CNN models that were trained normally used interactions in regions of an image in order to exploit correlations in the data, and that dampening the utilization of these interactions led these same models to be more robust against spurious correlations. To strengthen this conclusion we apply the Sobol-SBSA to images within the SalientImageNet (Singla & Feizi, 2021), and directly measure PTR for core and spurious regions within the SalientImageNet dataset. We sample 50000 images from SalientImageNet and subselect images that have been annotated as having both spurious and core regions. This yielded 1761 images. This experiment is run for 8×8 regions. Figure 4 shows that the robust models use interactions in both the spurious and core regions less than the vanilla versions of these models. Finally, we present an example of Sobol-SBSA being applied to Robust ResNet50 and ResNet50 for an egyptian cat and canoe image. Figure 5 shows the overlay of $\frac{1}{\text{PIR}}$ and S_i for each of these models on the image. A higher $\frac{1}{\text{PIR}}$ means larger interactive effects. It is seen that while both models view similar regions on the canoe and cat as being important individually, as measured by S_i, ResNet50 exploits background information through interactions in both cases. Similar examples can be found in figure 7 in the supplementary.

The key takeaway from this section is that *CNN models trained adversarially are resistant to exploiting spurious correlations in the data because they are forced to reduce the extent to which they rely on region interactions.*

4.2.3 The Limit of Sobol-SBSA

While the Sobol-SBSA is a powerful method, it has a similar weakness to other sample-based black box methods. It assumes that the input features that it is analyzing are uncorrelated. When this is not the case, the ability of Sobol to decompose importance into independent subsets of the feature space is weakened (Li et al., 2010). This manifests itself in the Sobol Indices, which are supposed to satisfy the property that PIR ≤ 1, having at least some input features for which PIR > 1.

In the case of regions, the assumption of uncorrelated input features breaks down as regions become smaller in size since the regions will tend more to belong to the same object on an image (the wing of a bird). We observe this in two ways. First, we examine figure 3a and note that PIR increases as the region sizes increase as region size decreases. This is intuitively an incorrect result since interactive effects should be greater for smaller regions, as demonstrated by Brendel & Bethge (2019). To confirm that this result is due to a breakdown in Sobol assumptions we calculate the average percent of regions for which PIR > 1 across all models for a given region size. For 4×4, 8×8, and 14×14 regions the percent of regions for which PIR > 1 are 22%, 33%, and 39% respectively. The key takeaway here is that *Sobol-SBSA is should be used to compare how models use identically correlated features*. A key direction for future work is to implement a version of Sobol-SBSA that accounts for correlations in the data.

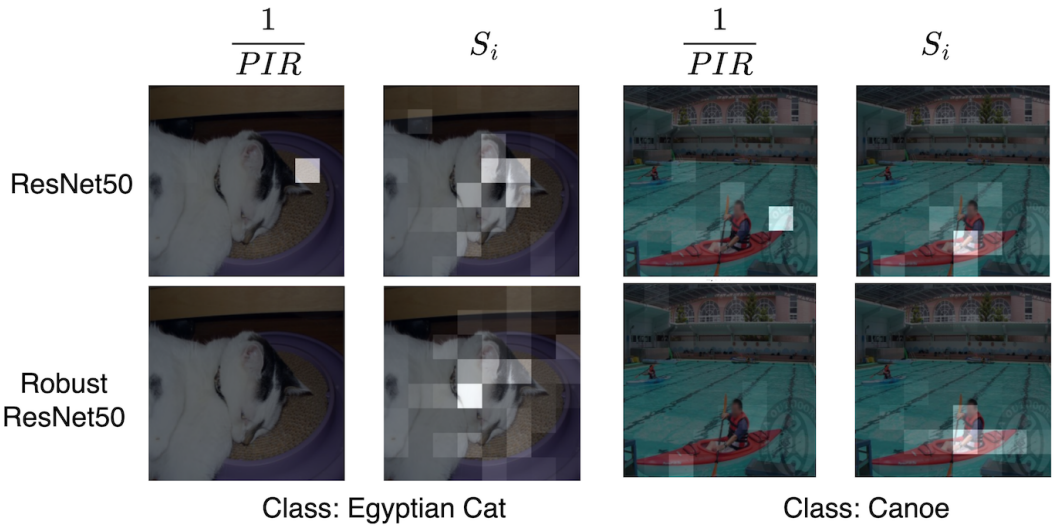


Figure 5: The result of Sobol-SBSA $\frac{1}{\text{PIR}}$ and S_i when applied to validation images for ResNet50 and RobustResNet50. From $\frac{1}{\text{PIR}}$, we see that ResNet50 exploits interactions with background spurious correlations, water background for the *Canoe* class, and pet bed background for the *Egyptian Cat* class to a much greater degree compared to RobustResNet50. However, both models highlight similar regions of primary importance, as seen from S_i.

5 Discussion and Conclusion

We proposed Sobol-SBSA, a general method for understanding how Deep Learning Models use semantic features when making decisions. We demonstrated the ability of this method to answer richer questions about models by applying it to how CNN models use regions of ImageNet models. We found (1) that the primary impact of the evolution of CNN models was to make greater use of region interactions, and (2) that adversarially robust CNN models are less susceptible to spurious correlations in the data because they force CNN architectures to rely less on region interactions.

The Sobol-SBSA method has a variety of potential applications beyond the image/region-based analysis that we presented here. Many different types of input partitioning and input modalities can be analyzed using the SBSA method. For images, the partitioning into image regions can be done via external metadata, such as object detection results. For natural language processing (NLP), the use of SBSA is even more straightforward since words/tokens provide a natural way to partition the input sequence. Hence, our model can be utilized to understand how parts of speech are utilized for different iterations of various NLP models, including translation, classification, and generative text models. Beyond unimodal content, SBSA can also be easily extended to multimodal setups. For example, Sobol-SBSA can be used to measure the strength of bias in Image Captioning systems by quantifying the extent to which parts of an image correspond to which parts of the generated caption. In addition to these applications, there are multiple avenues for future work. These include exploring the combination of Sobol-SBSA with automatic feature detectors, implementing a version of Sobol-SBSA that accounts for input correlations, and exploring whether PIR can be used as a proxy for the robustness of non-CNN Deep Learning models.

Through SBSA and Sobol-SBSA we have proposed a strong foundation for obtaining a richer understanding of how models use semantic representations of inputs, regardless of whether these representations are generated automatically or by end users. We hope that it can be used to both provide clarity about the mechanisms by which Deep Learning makes decisions, and to influence how we further develop these models.

REFERENCES

Marco Ancona, Enea Ceolini, A. Cengiz Öztireli, and Markus H. Gross. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104, 2017. URL http://arxiv.org/abs/1711.06104.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. volume 7, 2019.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *CoRR*, abs/2111.04138, 2021. URL https://arxiv.org/abs/2111.04138.

Jon Herman and Will Usher. Salib: An open-source python library for sensitivity analysis. *Journal of Open Source Software*, 2, 2017. URL https://doi.org/10.21105/joss.00097.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *CoRR*, abs/2002.04138, 2020. URL https://arxiv.org/abs/2002.04138.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

I. Kononenko, E. Štrumbelj, Z. Bosnic, Darko Pevec, M. Kukar, and M. Robnik-Sikonja. Explanation and reliability of individual predictions. *Informatica*, 37, 2013.

Genyuan Li, Herschel Rabitz, Paul E. Yelvington, Oluwayemisi O. Oluwole, Fred Bacon, Charles E. Kolb, and Jacqueline Schoendorf. Global sensitivity analysis for systems with independent and/or correlated inputs. *The Journal of Physical Chemistry A*, 114, 2010. URL https://doi.org/10.1021/jp9096919.

Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A. Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. URL https://arxiv.org/abs/1806.07421.

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

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *ArXiv preprint arXiv:2007.08489*, 2020.

Andrea Saltelli, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, and Stefano Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications*, 181, 2010. URL https://doi.org/10.1016/j.cpc.2009.09.018.

Manuel Schultheiss, Sebastian Schober A., Marie Lodde, Jannis Bodden, Juliane Aichele, Christina Müller-Leisse, Bernhard Renger, Franz Pfeiffer, and Daniela Pfeiffer. A robust convolutional neural network for lung nodule detection in the presence of foreign bodies. *Scientific Reports*, 32, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Sahil Singla and Soheil Feizi. Causal imagenet: How to discover spurious features in deep learning? *CoRR*, abs/2110.04301, 2021. URL https://arxiv.org/abs/2110.04301.

Sahil Singla, Mazda Moayeri, and Soheil Feizi. Core risk minimization using salient imagenet, 2022. URL https://arxiv.org/abs/2203.15566.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 34, 2017.

Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate interactions in dnns. *CoRR*, abs/2010.05045, 2020. URL https://arxiv.org/abs/2010.05045.

A APPENDIX

A.1 10K VS 50K EXPERIMENTS

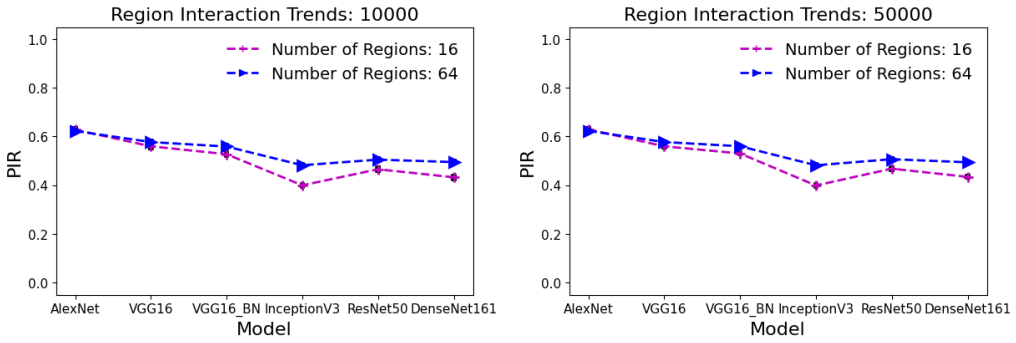


Figure 6: Average PIR for the top 20% of regions when calculated for 10000 and 50000 ImageNet validation images. This is done for 4×4 and 8×8 regions. The results are identical.

We calculated the average PIR for the top 20% of regions for 10000 and 50000 validation images in ImageNet using 4×4 and 8×8 regions respectively. We saw that the results were identical so, for the smaller regions generated when splitting an image into 14×14 regions, we calculated the average PIR for 10000 images to save computational cost.

A.2 RESNET50 VS ROBUST RESNET50 EXAMPLES

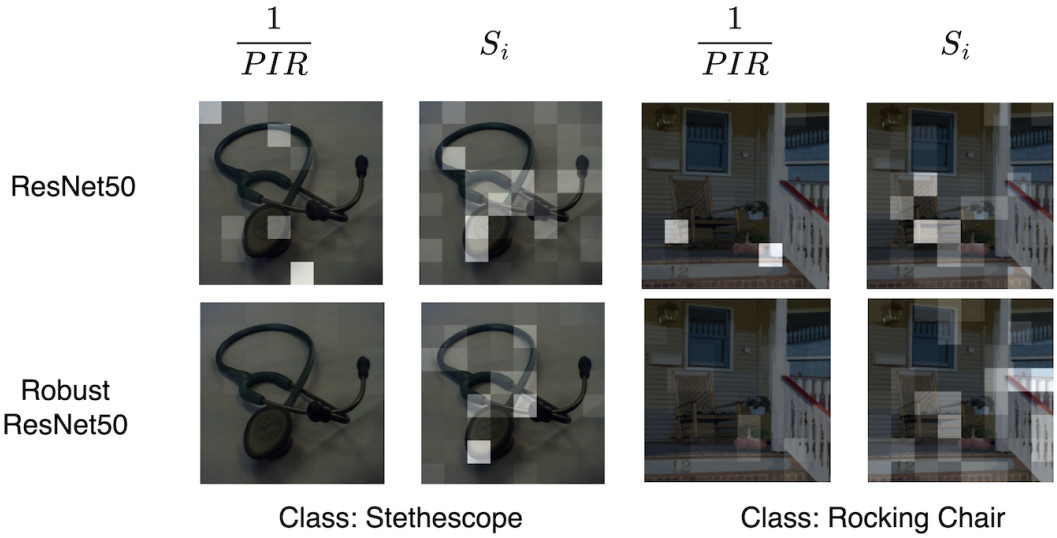


Figure 7: The result of Sobol-SBSA $\frac{1}{\text{PIR}}$ and S_i when applied to a validation images for ResNet50 and RobustResNet50. From $\frac{1}{\text{PIR}}$, we see that ResNet50 exploits interactions with background spurious correlations, white surface background for the *Stethescope* class, and porch background for the *Rocking Chair* class to a much greater degree compared to RobustResNet50. However, both models highlight similar regions of primary importance, as seen from S_i.

Figure 7 shows additional examples of images where both ResNet50 and the Robust ResNet50 looked at similar parts of the image as being of primary importance, S_i, but ResNet50 had stronger interactions with the background of the images which it exploited in making predictions. In this case, ResNet50 interacted with the white background when classifying a stethoscope and the porch when classifying a rocking chair.