Supplement to: Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations

Alexander Binder[1,2,9] Leander Weber[3] Sebastian Lapuschkin[3,0000−0002−0762−7258]
Grégoire Montavon[4,5] Klaus-Robert Müller[5,6,7,8] Wojciech Samek[3,5,6,0000−0002−6283−3265]
[1]ICT Cluster, Singapore Institute of Technology (SIT), Singapore 138683
[2]University of Oslo, 0316 Oslo, Norway
[3]Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[4]Freie Universität Berlin, 14195 Berlin, Germany
[5]BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[6]Technische Universität Berlin, 10587 Berlin, Germany
[7]Korea University, Seoul 136-713, Republic of Korea
[8]Max Planck Institut für Informatik, 66123 Saarbrücken, Germany
[9]SUTD, Singapore 487372

A. Details on Experiments

A.1. Randomization-based Sanity Checks vs. Faithfulness

The ResNet-50 and DenseNet-121 are used as provided by the Torchvision package of PyTorch [6]. For the EfficientNet-B0 we resort to a pretrained model provided by the github of Luke Melas [5]. All results are averaged over the first 1000 images from the ImageNet validation set.

For model-randomization-based sanity checks testing we reset the layers as per the initialization introduced by [3]. We report model randomization for a partial set of layers, as these results are in already known from [1], which makes an exhaustive computation for each layer unnecessary. For the ResNet-50 we randomize the fully connected layer, and in each step we randomize all layers from the last randomized layer until the next layer with name `.conv1` as per Torchvision until we have randomized 16 `.conv1`-named layers. For the DenseNet-121 we also randomize the fully connected layer, and in each step we randomize all layers from the last randomized layer until the next third layer with name `.conv1` until we have randomized 63 `.conv1`-named layers. For the EfficientNet-B0 we randomize the fully connected layer, and in each step we randomize all layers from the last randomized layer until the next layer with name `._depthwise_conv` until we have randomized 17 `._depthwise_conv`-named layers.

For the perturbation-based testing we create a blurred version of the original image, using a constant blur kernel of kernel size 15. We perform the perturbation by replacing a region of kernel size 8 or 15 in the original image by a patch from the blurred version. We do this for the 30 regions in an image which have the highest average attribution map score. Unlike the random draw for a patch used [7], using a blurred copy results in a less pronounced outlier structure due to preservation of color statistics while removing texture. We measure the decrease of the prediction function under the iterative replacement of the highest scoring patches of the image by the corresponding patches from the blurred copy.

A.2. Forward Pass-Adaptive β-rule

It is used in the experiments for model faithfulness estimation. The idea is based on the interpretation that $\frac{\beta}{1+\beta}$ in LRP-β is the fraction of redistributed negative to positive relevance. An adaptive way to determine its value can be derived by setting it equal to the corresponding fraction $\frac{-\sum_i (w_i x_i)_-}{\sum_i (w_i x_i)_+}$ of the input statistics of a neuron, and solving it for β as in:

$$\frac{\beta}{1+\beta} = \frac{-\sum_i (w_i x_i)_-}{\sum_i (w_i x_i)_+} \tag{1}$$

$$\Rightarrow \beta = \frac{-\sum_i (w_i x_i)_-}{\sum_i (w_i x_i)_+ - \sum_i (w_i x_i)_-} \tag{2}$$

We use a value of $\beta_* = \min(\beta, 3.0)$ in all experiments.

A.3. Additional Results of Model Faithfulness Experiments

Please see Figure 1 for results with a kernel size of 8. Furthermore Figure 2 shows results for 4000 additional images from the Imagenet validation set, and for 5000 images

from the MSCOCO dataset. One can see that the difference between gradient-based methods and LRP variants is also visible on MSCOCO data.

For computing faithfulness on the MSCOCO validation dataset we fine-tuned a multi-label classifier over ground truth derived from the intersection of the center crop with the bounding boxes of the MSCOCO detection task. While it could be more interesting to see results from training from scratch, we chose fine-tuning in order to arrive at a well-performing classifier within a small training time budget and to avoid the risk of having to try a large number of setups until we can obtain a properly performing network when training from scratch. It should be noted that there are very few pre-trained model initializations available which use large datasets in the order of hundred thousands without including ImageNet data, although foundation models are emerging to provide alternatives. For fine-tuning we assigned a centercrop a positive label if it contained at least 40% of the area of a bounding box of a class. We used AdamW, a learning rate of 0.0001 for all layers, a batch size of 64. Data augmentation used a 224 pixel resize, a 224 pixel center crop and RandomHorizontalFlip. We selected the model with the best validation performance within the first 20 epochs. Unlike ImageNet, MSCOCO comes with multi-label predictions. In contrast to ImageNet, where the unique ground truth label was used to obtain explanations, we chose for MSCOCO to explain the highest predicted label.

B. Proof of Theorem 1

Proof. Consider the term in Equation (1) of the main paper. Since we consider processes with $\sigma_{AB} \geq 0$, this attains the minimum at $\sigma_{AB} = 0$, resulting in

$$\min_{\sigma_{AB}\geq 0} \left| \frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1} \frac{2\sigma_{AB} + C_2}{\sigma_A^2 + \sigma_B^2 + C_2} \right| \tag{3}$$

$$= \frac{|2\mu_A\mu_B + C_1|}{\mu_A^2 + \mu_B^2 + C_1} \cdot \min_{\sigma_{AB}\geq 0} \frac{|2\sigma_{AB} + C_2|}{\sigma_A^2 + \sigma_B^2 + C_2} \tag{4}$$

$$= \frac{|2\mu_A\mu_B + C_1|}{\mu_A^2 + \mu_B^2 + C_1} \frac{C_2}{\sigma_A^2 + \sigma_B^2 + C_2} \tag{5}$$

$$\leq \frac{2|\mu_A\mu_B| + C_1}{\mu_A^2 + \mu_B^2 + C_1} \frac{C_2}{\sigma_A^2 + \sigma_B^2 + C_2} \tag{6}$$

$$\leq \frac{\mu_A^2 + \mu_B^2 + C_1}{\mu_A^2 + \mu_B^2 + C_1} \frac{C_2}{\sigma_A^2 + \sigma_B^2 + C_2} \tag{7}$$

The last inequality holds due to $\pm 2ab \leq a^2 + b^2$.

□

C. The Sensitivity of Spearman Rank Correlation Minimization Towards Noise

In Section 3 of the main paper we demonstrated the sensitivity of the SSIM metric towards random attributions. The same (in terms of ranks) holds for the other distance metric employed by [1], the Spearman Rank Correlation, given as

$$\frac{\sigma_{R(A)R(B)}}{\sigma_{R(A)}\sigma_{R(B)}}, \tag{8}$$

with $R(A)$ and $R(B)$ being the ranks derived from attribution maps A and B. The following Theorem and Proof show the sensitivity of this metric's minimization towards random noise analogously to Theorem 1 and the corresponding Proof:

Theorem 3. *Consider the set of all statistical processes with non-negative expected covariance between the corresponding ranks $\sigma_{R(A)R(B)} \geq 0$.*

Then the expected Spearman Rank Correlation is minimized by a statistical process with zero covariance between the corresponding ranks.

Proof. Consider the term in Equation (8). Since $\sigma_{R(A)R(B)} \geq 0$, and the standard deviations $\sigma_{R(A)}, \sigma_{R(B)} \geq 0$, this attains the minimum at $\sigma_{R(A)R(B)} = 0$, resulting in

$$\min_{\sigma_{R(A)R(B)}\geq 0} \left(\frac{\sigma_{R(A)R(B)}}{\sigma_{R(A)}\sigma_{R(B)}} \right) \tag{9}$$

$$= \frac{0}{\sigma_{R(A)}\sigma_{R(B)}} = 0 \tag{10}$$

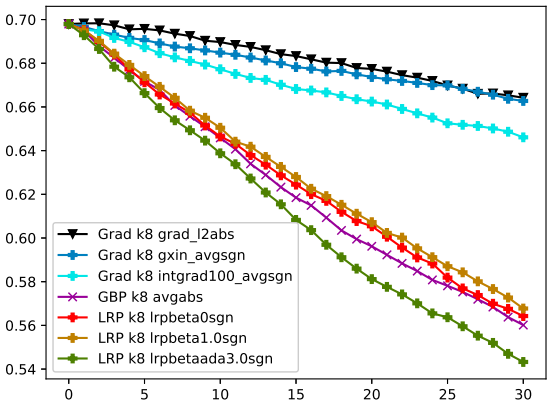
□

Of course, the ranks and their covariance depend not only on the attribution maps, but also on the employed ranking function. However, if simply the sorted indices of attribution maps (or their absolute values) are used as ranks, then $\sigma_{R(A)R(B)} \geq 0$ iff $\sigma_{AB} \geq 0$, and Theorem 3 holds for all statistical processes with non-negative expected covariance $\sigma_{AB} \geq 0$.

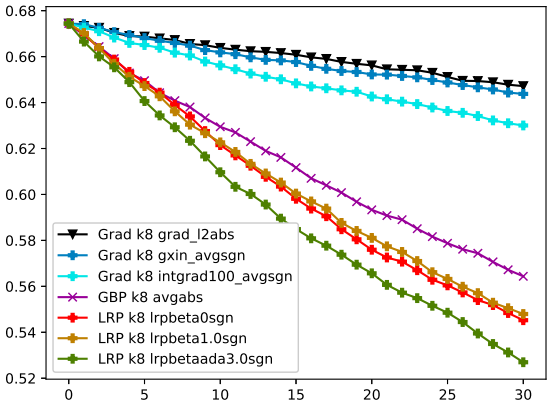
D. The Sensitivity of Normalized MSE Maximization Towards Noise

One may consider to replace the Structural Similarity Index Measure (SSIM) by a Mean-Squared Error (MSE). This comes with another topic to be considered: Different methods to compute attribution maps may exhibit different patch-wise variances, which will affect the scale of differences used in model-randomization-type sanity checks unrelated to the effects coming from the model randomization itself. This raises the question of how to normalize attribution maps in order to ensure a comparability of the distances computed using different attribution methods.

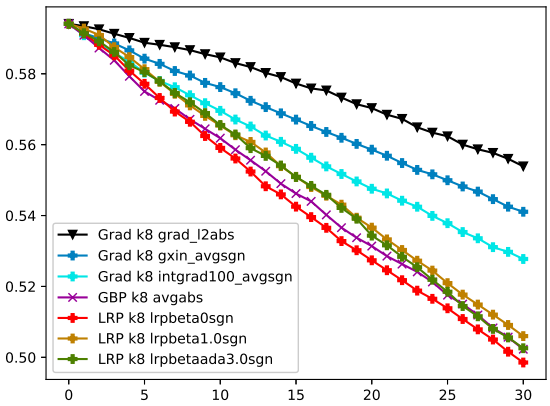
We consider for the case of MSE attribution maps which are normalized by dividing the attribution map by the square-



(a) ResNet50



(b) DenseNet121



(c) EfficientNet-B0

Figure 1. Results of model faithfulness testing by measuring the correlation to iterative occlusion with a kernel size of 8 on 1000 ImageNet validation set images. The comparison shows the gradient, gradient $\times$ input, integrated gradient, guided backpropagation and several LRP approaches. The occlusion is performed by taking patches from a blurred copy of the original image. The figure shows the softmax scores. *Lower is better.*

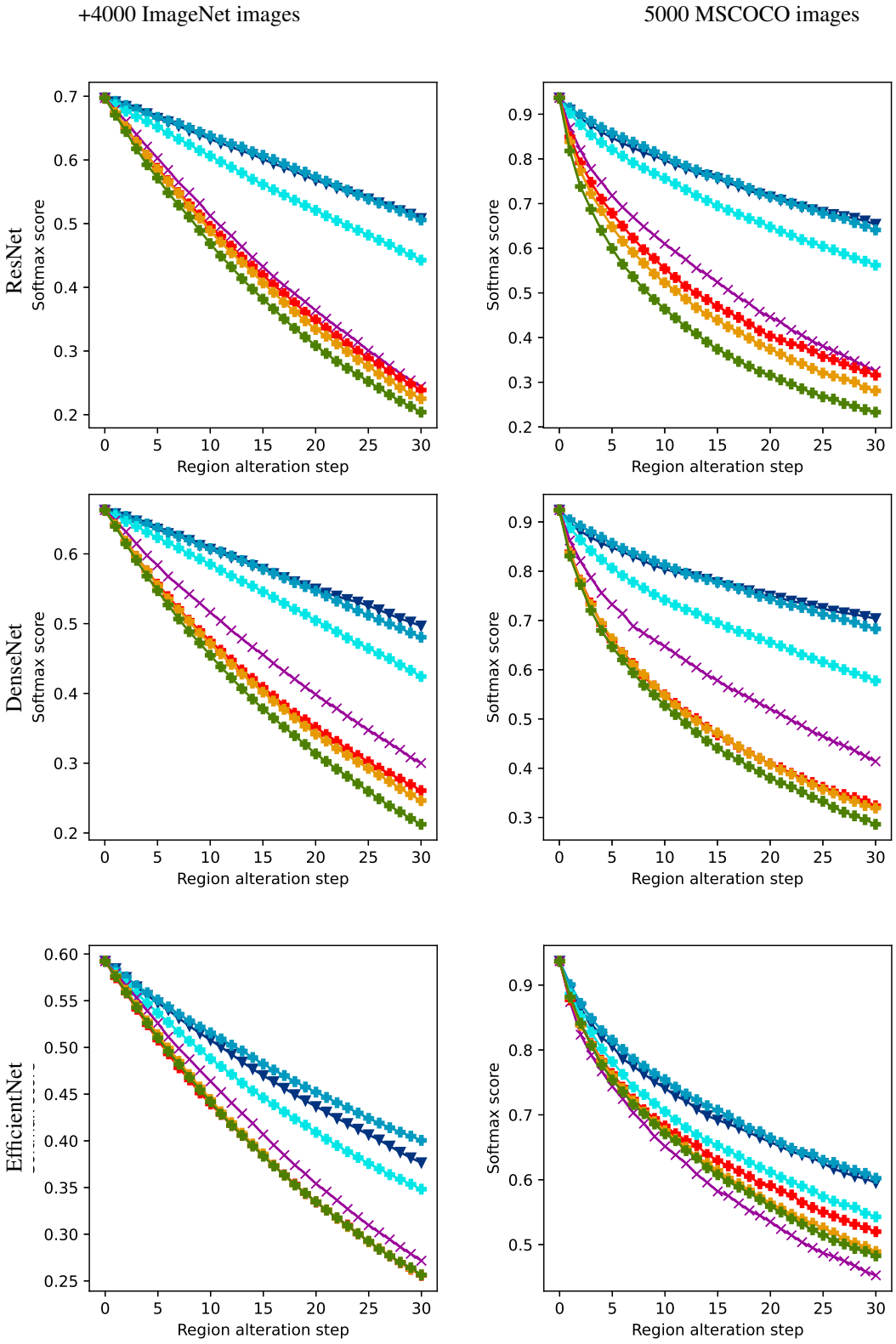


Figure 2. Additional results for Model Faithfulness, kernel size 15. *Left column:* Results on 4000 additional ImageNet images. *Right column:* Results on 5000 images from MSCOCO. The figure legend is the same as in Figure 2 in the submission and in Figure 1 of this supplement. *Lower is better.*

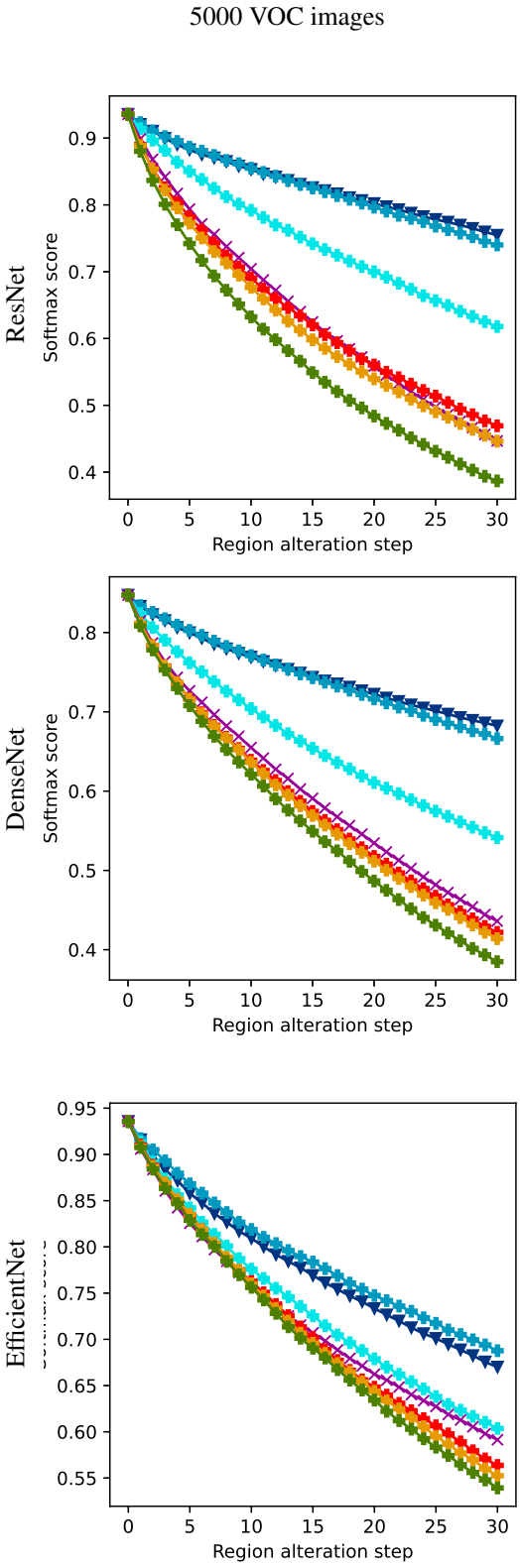


Figure 3. Additional results for Model Faithfulness, kernel size 15. Results on 5000 images from PASCAL VOC. The figure legend is the same as in Figure 2 in the submission and in Figure 1 of this supplement. *Lower is better.*

root of its average second moment estimate:

$$\frac{A}{\left(\hat{E}[A^2]\right)^{1/2}} = \frac{A_{h,w}}{\left(\frac{1}{HW}\sum_{h',w'} h^2_{h',w'}\right)^{1/2}}, \quad (11)$$

where $A_{h,w}$ is the value of the attribution map at pixel location (h, w) and H, W denote the attribution map height and width, respectively. This ensures that the average squared distance of an attribution score per pixel from the attribution value of 0 is 1. One may ask why we did not choose the more common standard deviation [1]. Standard deviation normalizes the average squared distance of a pixel-score from the mean of a patch to be one. However, the mean score over the pixels of the different attribution methods (such as gradient with ℓ_2-norm over the RGB-subpixels, also known as Sensitivity [2, 8], gradient with averaging over the RGB-subpixels, gradient $\times$ input and integrated gradients) has no particular meaning for explaining the prediction in the context of the above methods. The value of zero (0) has for all of above methods the meaning of being the estimate of non-contribution to the prediction, which justifies the choice of second moment estimates. Thus we ensure an equal average distance from the point of no-contribution by this type of normalization. This ensures better comparability of distances among attribution maps computed for different attribution map processes.

Using the MSE does not resolve the issue that a zero covariance attribution map yields the best results among all statistical processes with non-negative covariances $\sigma_{AB} \geq 0$. In the following A, B can be single subpixels. It directly translates to patches when using $E[\|A - B\|_2^2]$ instead.

The following theorem is again meant to be used with two different attribution maps A, B, e.g., coming from a model and a partially randomized variant of it, over the same patch location.

Theorem 4. *Consider the set of all statistical processes with non-negative expected covariance for each patch $\sigma_{AB} \geq 0$. Then the expected MSE can be maximized by using a statistical process with zero covariance and the maximal value is $2 - 2\frac{\mu_A\mu_B}{E[A^2]^{1/2}E[B^2]^{1/2}}$*

Proof.

$$E\left[\left(\frac{A}{E[A^2]^{1/2}} - \frac{B}{E[B^2]^{1/2}}\right)^2\right] \quad (12)$$

$$= \frac{E[A^2]}{E[A^2]} - 2\frac{E[AB]}{E[A^2]^{1/2}E[B^2]^{1/2}} + \frac{E[B^2]}{E[B^2]} \quad (13)$$

$$= 2 - 2\frac{\sigma_{AB}}{E[A^2]^{1/2}E[B^2]^{1/2}} - 2\frac{\mu_A\mu_B}{E[A^2]^{1/2}E[B^2]^{1/2}} \quad (14)$$

$$\leq 2 - 2\frac{\mu_A\mu_B}{E[A^2]^{1/2}E[B^2]^{1/2}} \quad (15)$$

[1] $\left(\hat{E}[A^2_{h,w}] - (\hat{E}[A_{h,w}])^2\right)^{1/2}$ for reference

With respect to the influence of means, for $\mu_A\mu_B \approx 0$, this would result in a MSE of 2. Note that we can observe from Figure 5 that the MSE indeed attains a value close to 2.0 for certain methods which perform well in model-randomization-type sanity checks, such as gradient and integrated gradient. In light of above theorem, this finding is conspicuous as it may indicate a correlation σ_{AB} and means μ_A, μ_B close to zero, in the sense of high gradient shattering noise.

This result shows, that the contribution from the randomization of a model and the noise from the attribution map process are still entangled when using MSE for model-randomization-type sanity checks. Consequently, using attribution map processes with a lower degree of correlation within a patch makes them appear more favorable when using model-randomization-type sanity checks to compare attribution maps. As noted before, a lower degree of correlation may originate from using a process with a high amount of zero-correlation noise, and in the worst case, from a statistically independent random process.

On a side note, using an unnormalized MSE would result in

$$E\left[(A-B)^2\right] = E[A^2] - 2E[AB] + E[B^2] \quad (16)$$

$$= \sigma_A^2 - 2\sigma_{AB} + \sigma_B^2 + E[A]^2 - 2E[A]E[B] + E[B]^2 \quad (17)$$

$$= \sigma_A^2 + \sigma_B^2 + (\mu_A - \mu_B)^2 - 2\sigma_{AB} \quad (18)$$

$$\leq \sigma_A^2 + \sigma_B^2 + (\mu_A - \mu_B)^2 \quad (19)$$

This would again be a measure that is maximized among the set of processes with non-negative covariance by using $\sigma_{AB} = 0$ and additionally be sensitive to increasing patch-wise variances σ_A^2, σ_B^2 in processes. The proposed normalization by the second moment puts a bound on the sensitivity to patch-wise variances.

In summary, this section shows that replacing minimization of a similarity by maximization of a well-known squared distance still retains a sensitivity and possible preference towards attribution map methods with low correlation when viewed as a statistical process.

D.1. A Note on Normalization

The scores obtained from different attribution methods generally do not share the same range of values. Therefore, in order to compare them, some sort of normalization is required, however, care has to be taken in doing so as to not alter or destroy any information provided by attributions. The seminal work on model-randomization-based sanity checks [1] uses normalization by division with statistics using the maximal absolute value of an attribution map.

This, however, may introduce additional variance in the measurement when computing differences of attribution maps:

Minima or maxima are the only statistics among quantile estimators which do not converge for an increasing sample size to a finite expectation. One can easily see this by considering random draws from a normal distribution. The maximum will tend to infinity as the sample size $n \to \infty$ increases.

More formally, as noted in [9], the distribution of several known quantile estimators for the p-th quantile of a distribution is approximately normal with a variance of

$$\sigma^2 = \frac{p(1-p)}{nf(F^{-1}(p))} \tag{20}$$

where $f(\cdot)$ is the density, $F(\cdot)$ the cumulative density of the distribution which is used to draw samples used to compute the p-th quantile estimator, and n is the sample size. Thus for quantile estimators $p \approx 0, p \approx 1$ with values $F^{-1}(p)$ at the tails of the distribution, where the value of the density $f(\cdot)$ is low, the variance σ^2 will become unbounded, as long as $f(F^{-1}(p))$ decays faster than $\mathcal{O}(p^{-1})$ or $\mathcal{O}((1-p)^{-1})$, which is the case for a higher degree polynomial or exponential decay.

It should be noted that normalization aiming at a proper perception by the human eye and normalization for the sake of comparability of distances are non-equivalent goals. The former needs to ensure a bounded range, and color intensities which are well perceivable.

Normalization by the maximum yields a high variance of the estimator, and, while suitable for visualization to the human eye, does not preserve a quantity useful for the comparison of distances across different models under parameter randomization. For this reason we will consider a different normalization as outlined above.

E. Proof of Theorem 2

Proof. To see this, consider two sets of non-negative input activations for a neuron, X_L and X_S. We assume that each input from X_L is by a factor of K larger than each input from X_S such that:

$$\min_{x_l \in X_L} x_l \geq K \max_{x_s \in X_S} x_s \,. \tag{21}$$

In order for a single x_s to have at least the same effect on the output as a single $x_l > 0$, it requires $w_s x_s \geq w_l x_l$ and thus for the weights $w_s \geq K w_l$. This corresponds to a ratio distribution of two zero-mean normal variables, which is known to have a Cauchy density, as for example shown in [4]

$$f\left(\frac{w_s}{w_l} = K\right) = \frac{1}{\pi\gamma}\frac{1}{K^2/\gamma^2 + 1},\ \gamma = \frac{\sigma_s}{\sigma_l} \,. \tag{22}$$

The quantity of interest in this case is the tail-CDF

$$P\left(\frac{w_s}{w_l} \geq K\right) = 1 - CDF_\gamma(K) \,. \tag{23}$$

In order for each of the neurons in the small-value set to have the same summed contribution to the output, we require

$$\sum_{x_s \in X_S} w_s x_s \geq \sum_{x_l \in X_L} w_l x_l \,. \tag{24}$$

This can be combined together as follows.

$$\sum_{x_s \in X_S} w_s x_s \text{ and } \sum_{x_l \in X_L} w_l x_l \tag{25}$$

are normally distributed random variables with respect to draws of the weights w with zero mean and variances

$$\sigma_S = \sum_{x_s \in X_S} x_s^2 \text{ and } \sigma_L = \sum_{x_l \in X_L} x_l^2 \,. \tag{26}$$

Thus, for $\sum_{x_l \in X_L} w_l x_l > 0$ the requirement in Equation (24) translates into the probability of the ratio

$$\frac{\sum_{x_s \in X_S} w_s x_s}{\sum_{x_l \in X_L} w_l x_l} \geq 1 \tag{27}$$

This is the cumulative tail probability $P(Z \geq 1) = 1 - CDF_\gamma(1)$ with a parameter γ_1 given as

$$\gamma_1 = \sqrt{\frac{\sigma_S}{\sigma_L}} = \sqrt{\frac{\sum_{x_s \in X_S} x_s^2}{\sum_{x_l \in X_L} x_l^2}} \tag{28}$$

The Cauchy distribution obtains larger cumulative tail probabilities for larger values of the parameter γ. Therefore for an upper bound on cumulative tail probabilities, we need to obtain an upper bound on γ_1.

$$\gamma_1 = \sqrt{\frac{\sum_{x_s \in X_S} x_s^2}{\sum_{x_l \in X_L} x_l^2}} \tag{29}$$

$$\leq \sqrt{\frac{\sum_{x_s \in X_S} \max_{x_s \in X_S} x_s^2}{\sum_{x_l \in X_L} \min_{x_l \in X_L} x_l^2}} \tag{30}$$

$$= \sqrt{\frac{|X_S| \max_{x_s \in X_S} x_s^2}{|X_L| \min_{x_l \in X_L} x_l^2}} \tag{31}$$

$$\overset{\text{Eq.(21)}}{\leq} \sqrt{\frac{|X_S| \frac{1}{K^2} \min_{x_l \in X_L} x_l^2}{|X_L| \min_{x_l \in X_L} x_l^2}} \tag{32}$$

$$= \sqrt{\frac{|X_S|}{|X_L|}\frac{1}{K}} \tag{33}$$

where we used Equation (21) to get a term depending on K. Plugging in this upper bound γ_1 into the CDF shows

$$CDF_{\gamma_1}(1) = 0.5 + \frac{1}{\pi}\arctan(\frac{1-0}{\gamma_1}) \tag{34}$$

$$= 0.5 + \frac{1}{\pi}\arctan\left(\frac{K}{\sqrt{\frac{|X_S|}{|X_L|}}}\right) \tag{35}$$

$$= CDF_{\gamma_2}(K),\ \gamma_2 = \sqrt{\frac{|X_S|}{|X_L|}}\,. \tag{36}$$

Therefore we obtain the cumulative tail CDF of a Cauchy distribution from the value of K onwards $P(Z \geq K)$ with a parameter $\gamma_2 = \sqrt{\frac{|X_S|}{|X_L|}}$. □

Section I in this supplement provides estimates for this probability for three trained deep neural networks which provides empirical evidence for the sparsity.

If one would consider average contributions

$$\frac{1}{|X_S|}\sum_{x_s \in X_S} w_s x_s \geq \frac{1}{|X_L|}\sum_{x_l \in X_L} w_l x_l\,, \tag{37}$$

then one would obtain the analogous result with an inverted parameter $\gamma_{2,avg} = \sqrt{\frac{|X_L|}{|X_S|}}$.

A reason to consider such averages instead of sums would be the case when one is interested to analyze when two regions of an input would achieve the same average explanation score per input element of the respective regions. This case corresponds in an attribution map to two regions with the same average color intensity per pixel.

This can be shown as follows. If we consider

$$\frac{1}{|X_S|}\sum_{x_s \in X_S} w_s x_s \text{ and } \frac{1}{|X_L|}\sum_{x_l \in X_L} w_l x_l\,, \tag{38}$$

then these are normally distributed random variables with respect to draws of the weights w with zero mean and variances

$$\sigma_S = \frac{1}{|X_S|^2}\sum_{x_s \in X_S} x_s^2 \text{ and } \sigma_L = \frac{1}{|X_L|^2}\sum_{x_l \in X_L} x_l^2\,. \tag{39}$$

The difference to the proof above is a multiplicative factor in γ_1 in Equation (29) of

$$\sqrt{\frac{\frac{1}{|X_S|^2}}{\frac{1}{|X_L|^2}}} = \frac{|X_L|}{|X_S|} \tag{40}$$

$$\Rightarrow \gamma_{2,avg} = \frac{|X_L|}{|X_S|}\gamma_2 = \frac{|X_L|}{|X_S|}\sqrt{\frac{|X_S|}{|X_L|}} = \sqrt{\frac{|X_L|}{|X_S|}} \tag{41}$$

F. The Monotonicity Property of selected Explanation Methods

We show here that several explanation methods satisfy the positive monotonicity property that if we consider two inputs x_i, x_j which have no other connections except to neuron y, then $w_i x_i \geq w_j x_j > 0$ implies $|R(x_i)| \geq |R(x_j)|$.

F.1. Positive Monotonicity for Gradient × Input

$$z = g(\sum_k w_k x_k + b) \tag{42}$$

$$R(x_i) = \frac{\partial f}{\partial z}\frac{\partial z}{\partial x_i}(x)x_i = \frac{\partial f}{\partial z}g'(\cdots)w_i x_i \tag{43}$$

$$\sum_k w_k x_k + b > 0, w_i x_i > w_j x_j > 0 \Rightarrow \tag{44}$$

$$|R(x_i)| = \left|\frac{\partial f}{\partial z}\right| |g'(\cdots)||w_i x_i| \tag{45}$$

$$> |R(x_j)| = \left|\frac{\partial f}{\partial z}\right| |g'(\cdots)||w_j x_j| \tag{46}$$

In fact, a stronger version holds here: $|w_i x_i| \geq |w_j x_j|$ implies $|R(x_i)| \geq |R(x_j)|$

F.2. Positive Monotonicity for Shapley Values

This holds when $w_i x_i > w_j x_j > 0$ and the activation function g is monotonously non-decreasing. In that case, for all subsets $S : i \notin S, j \notin S$:

$$f(S \cup \{i\}) = g(\sum_{k \in S} w_k x_k + b + w_i x_i) \tag{47}$$

$$\geq g(\sum_{k \in S} w_k x_k + b + w_j x_j) = f(S \cup \{j\}) \tag{48}$$

$$\Rightarrow \phi(i) = \sum_S c_{|S|}(f(S \cup \{i\}) - f(S)) \tag{49}$$

$$\geq \sum_S c_{|S|}(f(S \cup \{j\}) - f(S)) = \phi(j)\,, \tag{50}$$

where

$$c_{|S|} = \frac{1}{d\binom{d-1}{|S|}} \tag{51}$$

are the normalizing constants used in the exact computation of Shapley values.

F.3. Positive Monotonicity for LRP-β

$$R(i) = R(z)(1+\beta)\frac{(w_i x_i)_+}{\sum_k (w_k x_k)_+} - R(z)\beta\frac{(w_i x_i)_-}{\sum_k (w_k x_k)_-} \quad (52)$$

$$w_i x_i > 0 \Rightarrow R(i) = R(z)(1+\beta)\frac{(w_i x_i)_+}{\sum_k (w_k x_k)_+} \quad (53)$$

$$w_j x_j > 0 \Rightarrow R(j) = R(z)(1+\beta)\frac{(w_j x_j)_+}{\sum_k (w_k x_k)_+} \quad (54)$$

$$w_i x_i \geq w_j x_j > 0 \Rightarrow (w_i x_i)_+ \geq (w_j x_j)_+ \quad (55)$$

$$\Rightarrow |R(i)| = |R(z)|(1+\beta)\frac{(w_i x_i)_+}{\sum_k (w_k x_k)_+} \quad (56)$$

$$\geq |R(j)| \quad (57)$$

In fact, a stronger version holds here: $|w_i x_i| \geq |w_j x_j|$ and $sign(w_i x_i) = sign(w_j x_j)$ implies $|R(x_i)| \geq |R(x_j)|$.

G. Positive Explanation Score Dominance in ReLU Networks with Positive Logits

In this section we briefly show another property to hold, when explaining positive logits in ReLU networks with non-positive biases, irrespective of the randomization.

The property is that the positive evidence will dominate the negative evidence in every layer until the input, *under the condition that the explanation is additive for ReLU units with positive outputs.* An exception to it would occur when one has large positive biases, and one would attribute explanation scores to the bias terms itself.

Consider a positive logit $f(x)$ as a linear combination of the last layer activations $\phi^{(L)}$ with a non-positive bias $b \leq 0$:

$$0 < f(x) = \sum_i w_i \phi_i^{(L)}(x) + b \quad (58)$$

$$0 < R\left(\sum_i w_i \phi_i^{(L)}(x)\right) = \sum_i R\left(w_i \phi_i^{(L)}(x)\right) \quad (59)$$

We can see that the explanations for the last layer activations must sum to a positive value as well. Now let us consider the output of a ReLU feature

$$\phi_i^{(L)}(x) = ReLU\left(\sum_k w_k \phi_k^{(L-1)}(x) + b\right) . \quad (60)$$

If the negative contributions to it dominate, then the output value of the ReLU is zero. This has the meaning that this neuron detects no feature. In this case $R\left(w_i \phi_i^{(L)}(x)\right) = 0$, and no explanation scores will be propagated back to its inputs $\phi_k^{(L-1)}(x)$, that is $R(\phi_k^{(L-1)}(x)) = 0$ received along this path from $\phi_i^{(L)}(x)$.

If positive contributions to it dominate, then $0 <$ ReLU and we use the same idea as in the previous section:

$$0 < \text{ReLU}\left(\sum_k w_k \phi_k^{(L-1)}(x)\right) = \sum_k w_k \phi_k^{(L-1)}(x) \quad (61)$$

$$\phi_i^{(L)}(x) = ReLU\left(\sum_k w_k \phi_k^{(L-1)}(x) + b\right)$$

$$\Rightarrow R\left(w_i \phi_i^{(L)}(x)\right) = R\left(\text{ReLU}\left(\sum_k w_k \phi_k^{(L-1)}(x)\right)\right) = \sum_k R(w_k \phi_k^{(L-1)}(x)) \quad (62)$$

$$\Rightarrow 0 < \sum_i R\left(w_i \phi_i^{(L)}(x)\right) = \sum_i \sum_k R(w_k \phi_k^{(L-1)}(x)) \quad (63)$$

We use here only additivity of explanations $R(\cdot)$, and non-assignment of explanation scores to bias terms. In summary, combining equation (59) with (63) shows that the sum of relevances in layer $L-1$ is positive and equal to the initial logit relevance. Iterating this through all layers proves the claim until the input. In practice, explaining positive logits with methods which satisfy such an additivity, will result in dominantly positive explanations.

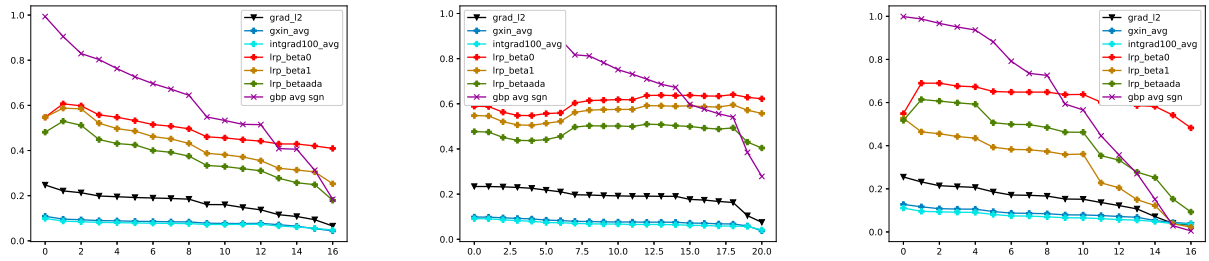
H. Top-down Model Randomization Experiments

Please see Figures 4 and 5 for the results on ImageNet, and Figure 6 for results on MSCOCO. For better comparability all attribution maps were normalized by the square root of their second moment (not their variance) as discussed in Section D. The results are in principle known from [1]. The behaviour on ImageNet and on a model finetuned from ImageNet to MSCOCO is qualitatively very similar.

I. Probabilities of overtaking large activations from forward pass activation statistics

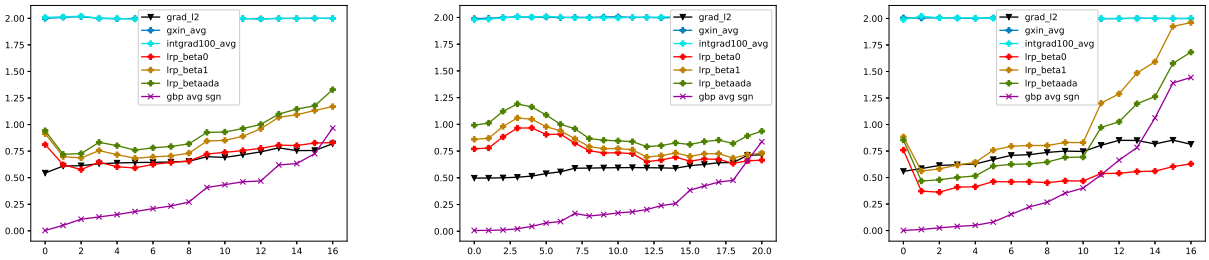
This Section computes an upper bound for the probability of overtaking according to Theorem 2 for given trained models from Resnet-50, DenseNet-121 and EfficientNet-B0 architectures. This shows that in practice these probabilities are small.

To do this, we compute for a given image the forward pass activations, and pool them in every layer across spatial



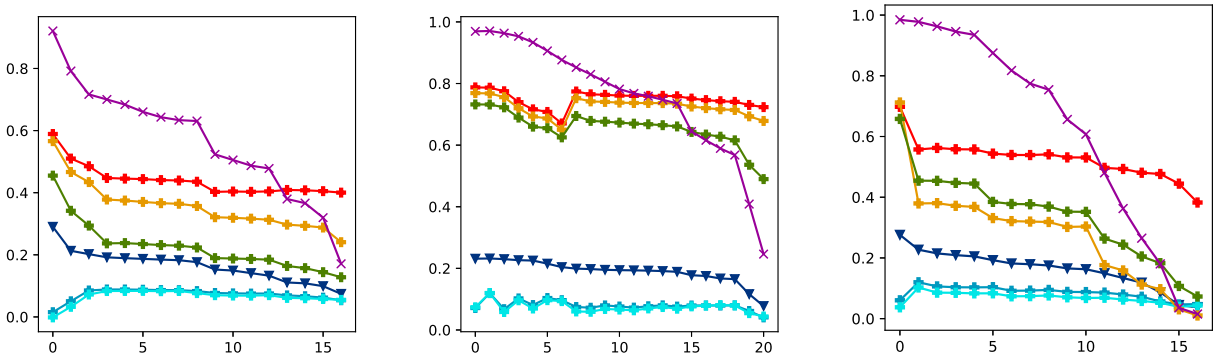
(a) ResNet50 (b) DenseNet121 (c) EfficientNet-B0

Figure 4. The figure shows the results of top-down model randomization-based sanity checks with SSIM after normalization of attribution maps by their second moment. *Lower is better.*



(a) ResNet50 (b) DenseNet121 (c) EfficientNet-B0

Figure 5. The figure shows the results of top-down model randomization-based sanity checks with MSE after normalization of attribution maps by their second moment. Of note is also the score of gradient-based results close to the value of 2 in comparison with the upper bound in Equation (15). *Higher is better.*



(a) ResNet50 (b) DenseNet121 (c) EfficientNet-B0

Figure 6. The figure shows the results of top-down model randomization-based sanity checks for the MSCOCO dataset with SSIM after normalization of attribution maps by their second moment. *Lower is better.*

and channel dimensions (because usually a convolution kernel takes all channels as input). Next we compute a set of quantile estimators for these activation values in the range from 0.95 down to 0.1 in 0.05 decrements. This yields 18 quantile estimators for every layer of the net and for one image. We compute the mean value of these estimators over 1000 images from the ImageNet validation set.

After that we can compute estimates for the value of $\gamma = \sqrt{\frac{|X_S|}{|X_L|}}$ and K from this information for every pair

(q_h, q_l) of a high quantile $q_h \in \{0.95, \ldots, 0.85\}$ and a low quantile $q_l \in \{0.5, \ldots, 0.1\}$. K is given as the ratio of quantile estimator values $K \geq \frac{V(q_h)}{V(q_l)}$, whereas γ is given as $\sqrt{\frac{|X_S|}{|X_L|}} = \sqrt{\frac{q_l}{1-q_h}}$, which corresponds to the relative fractions of the amount of bottom-k% activations to the amount of observed top-k% activations.

Finally we can plug this into the Cauchy cumulative tail density $P_\gamma(Z \geq K)$ to obtain the probabilities.

Each plot shows on the x-axis the low quantile $q_l \in \{0.5, \ldots, 0.1\}$, and on the y-axis $P_\gamma(Z \geq K)$. It shows one graph of probabilities $P_\gamma(Z \geq K)$ for each value of the high quantile $q_h \in \{0.95, 0.9, 0.85\}$. The graphs are color coded according to q_h.

The results are shown in Figures 7, 8 and 9.

We can see rather low probabilities despite the Cauchy distribution having a low order polynomial decay. Note that the EfficientNet can have negative activation statistics for some lower layers. In this case K is computed using the inverse (because in this case one wants to overtake the absolute larger negative values using the absolute smaller negative values).

Some graphs, like for Resnet-50 levels 9 and 12 remain almost flat zero because the mean activation is very close to zero for the bottom-50% values due to a strong sparsity in these layers. See Section J for the fraction of non-positive activations as an explanation, and compare the graph against Resnet-50 Level 6 and the corresponding statistics in Section J. We have verified that for higher bottom-% values one would see small positive overtaking probabilities $P_\gamma(Z \geq K)$.

J. Activation Statistics

This section shows the fraction of non-positive activations. Results are shown in Figure 10. One can see that for ResNet-50 and DenseNet-121, most layers have at least 30% zero activations. The amount of nonpositive activations is less for the EfficientNet-B0, which makes sense as this is less wide than other architectures. From layer 11 onwards it has also at least 20% zeros. Note that activations can get truly negative for the Efficientnet as a result of using the Swish activation function. Therefore seeing 100% in layer 0,1,3,5 is not a mistake.

Acknowledgements

AB was supported by the SFI Visual Intelligence, project no. 309439 of the Research Council of Norway. KRM was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). This work was supported in part by the German Ministry for Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A and 01IS18037A-I. LW, SL and WS were also supported by the European Union's Horizon 2020 research and innovation programme as grant [iToBoS (965221)], and the state of Berlin within the innovation support program ProFIT as grant [BerDiBa (10174498)]. WS was also supported by the German Research Foundation (ref. DFG KI-FOR 5363).

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31*, pages 9525–9536, 2018. 1, 2, 6, 9

[2] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010. 6

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[4] D. V. Hinkley. On the ratio of two correlated normal random variables. *Biometrika*, 56(3):635–639, 12 1969. 7

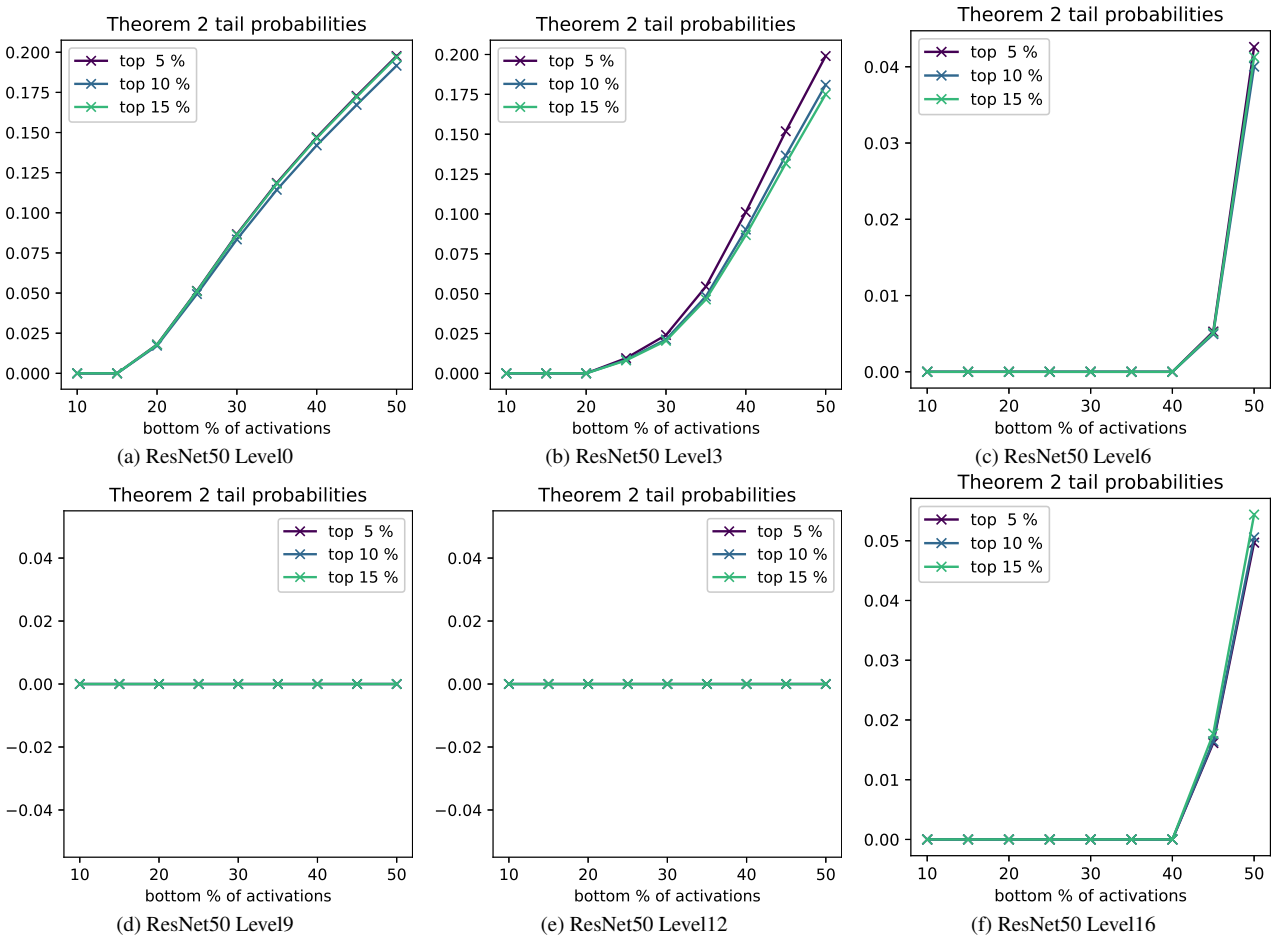
[5] Luke Melas. 2019. 1

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 1

[7] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. 1

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 6

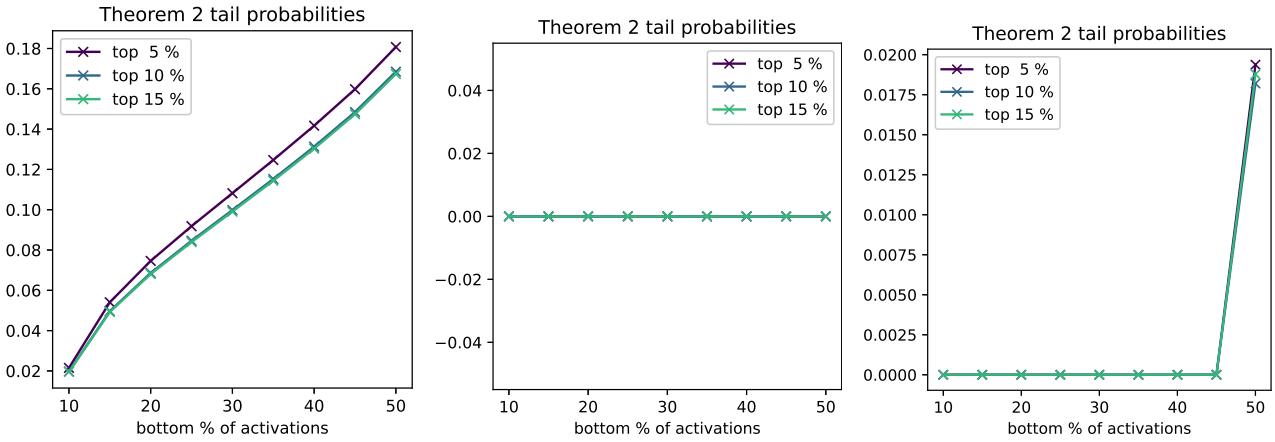
[9] A. Stuart and K. Ord. *Kendall's advanced theory of statistics*, volume 2, Classical Inference and Relationship. Wiley, fifth edition, March 1991. 7



(a) ResNet50 Level0 (b) ResNet50 Level3 (c) ResNet50 Level6

(d) ResNet50 Level9 (e) ResNet50 Level12 (f) ResNet50 Level16

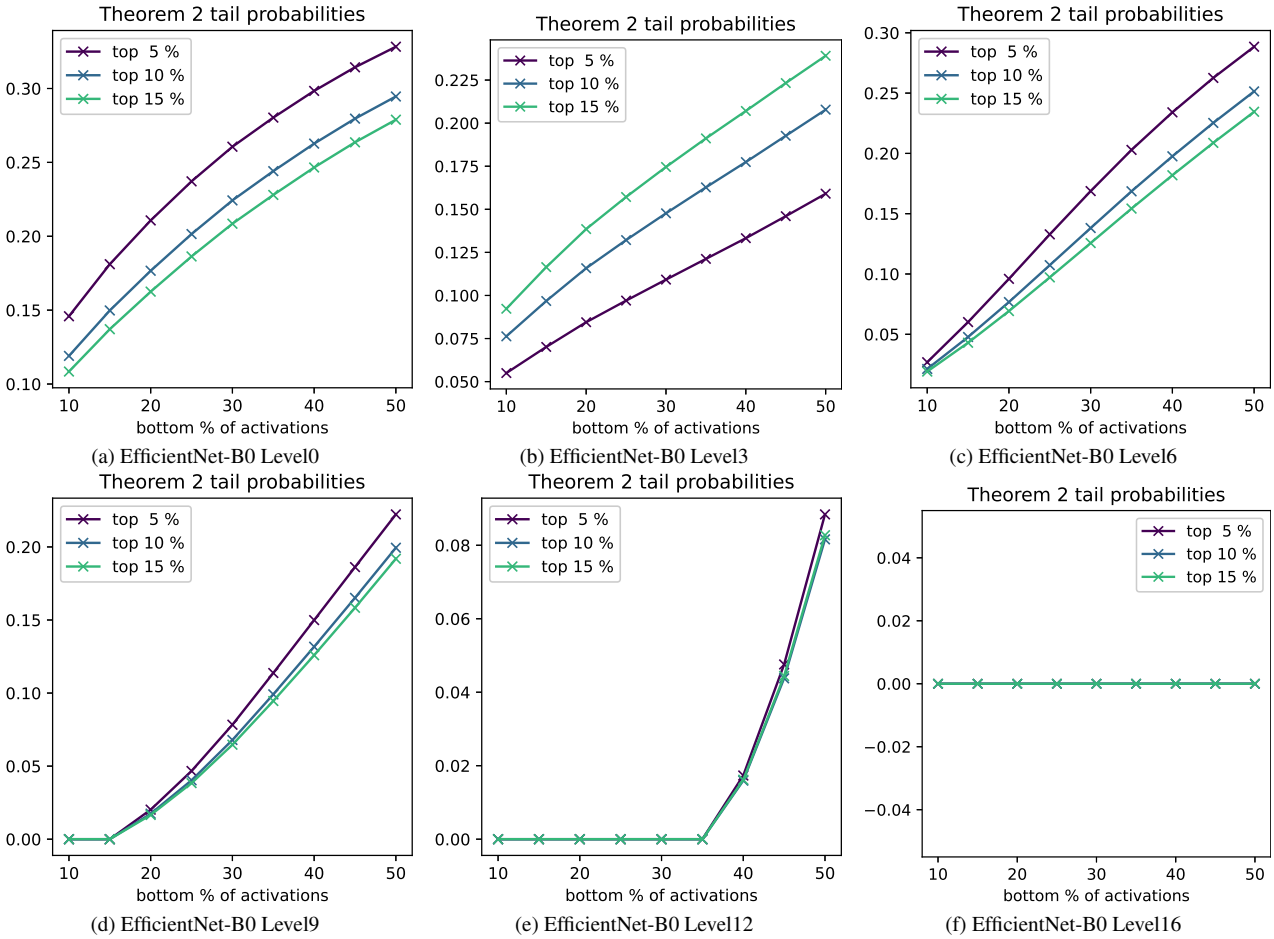
Figure 7. Lower probabilities support Theorem 2 better.



(a) DenseNet121 Level0 (b) DenseNet121 Level10 (c) DenseNet121 Level20

(d) DenseNet121 Level30 (e) DenseNet121 Level73 (f) DenseNet121 Level74

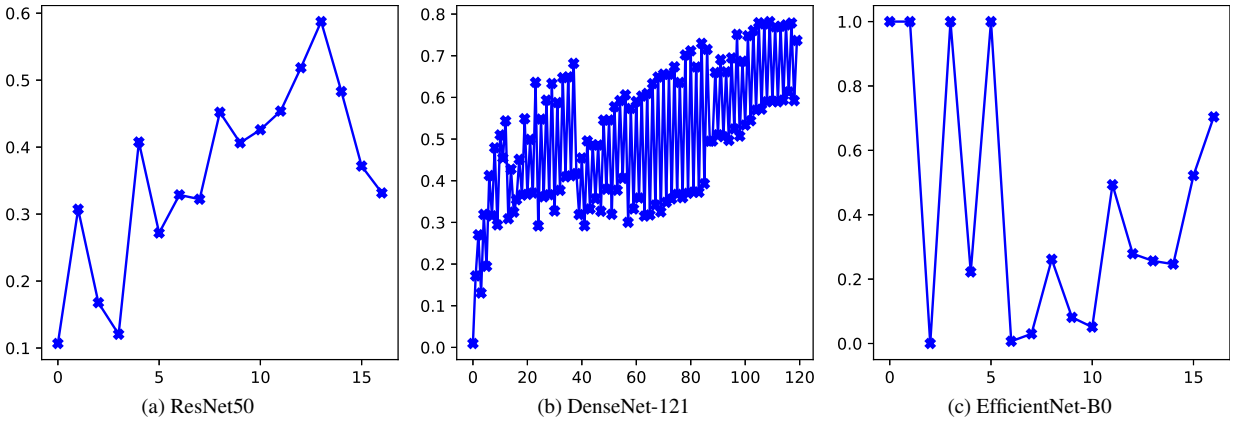
Figure 8. Lower probabilities support Theorem 2 better.



(a) EfficientNet-B0 Level0 (b) EfficientNet-B0 Level3 (c) EfficientNet-B0 Level6

(d) EfficientNet-B0 Level9 (e) EfficientNet-B0 Level12 (f) EfficientNet-B0 Level16

Figure 9. Lower probabilities support Theorem 2 better.



(a) ResNet50 (b) DenseNet-121 (c) EfficientNet-B0

Figure 10. Non-positive activations per layer. Higher values indicate a higher fraction of non-positive activations.