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, SIT Singapore, [2]UiO Oslo, [3]Fraunhofer HHI, [4]FU Berlin, [5]BIFOLD Berlin
[6]TU Berlin, [7]Korea University Seoul, [8]MPI Saarbrücken, [9]SUTD Singapore

Abstract

While the evaluation of explanations is an important step towards trustworthy models, it needs to be done carefully, and the employed metrics need to be well-understood. Specifically model randomization testing can be overinterpreted if regarded as a primary criterion for selecting or discarding explanation methods. To address shortcomings of this test, we start by observing an experimental gap in the ranking of explanation methods between randomization-based sanity checks [1] and model output faithfulness measures (e.g. [20]). We identify limitations of model-randomization-based sanity checks for the purpose of evaluating explanations. Firstly, we show that uninformative attribution maps created with zero pixel-wise covariance easily achieve high scores in this type of checks. Secondly, we show that top-down model randomization preserves scales of forward pass activations with high probability. That is, channels with large activations have a high probility to contribute strongly to the output, even after randomization of the network on top of them. Hence, explanations after randomization can only be expected to differ to a certain extent. This explains the observed experimental gap. In summary, these results demonstrate the inadequacy of model-randomization-based sanity checks as a criterion to rank attribution methods.

1. Introduction

Parallel to the progressively astounding performances of machine learning techniques, especially deep learning methods, in solving even the most complex tasks, the transparency, trustworthiness, and lack of interpretability of these techniques has increasingly been called into question [11, 14, 15]. As potential solutions to these issues, a vast number of XAI methods have been developed in recent years [21], that aim to explain a model's behavior, for instance, by (locally) attributing importance scores to features of singular input samples, indicating how (much) these features influence a specific model decision [6, 22, 25, 27]. However, the scores obtained for different attribution map methods tend to differ significantly, and the question arises how well each explains model decisions. This is generally not answered easily, as there are a number of desirable properties proposed to be fulfilled by these attributions, such as localization on relevant objects [4, 5, 30] or faithfulness to the model output [2, 8, 20], among others, with several quantitative tests having been proposed for each.

In parallel to these empirical evaluations, several works have proposed that explanations should fulfill a certain number of 'axioms' or 'unit tests' [1, 12, 16, 27], which need to hold universally for a method to be considered good or valid. We place our focus on the model-randomization-based sanity checks [1], which state that the explanation should be sensitive to a random permutation of parameters at one or more layers in the network. Specifically, the authors proposed to apply measures such as Structural Similarity Index Measure (SSIM) [28] between attribution maps obtained from the original model and a derived model for which the top-layers are randomized. The idea is to require that methods used to compute attribution maps should exhibit a large change when the neural network model — i.e., its defining/learned parameter set — is randomized from the top. The authors of [1, 23] suggested to discard attribution map methods which perform poorly under this test — i.e., have a high SSIM measure between attributions obtained with the original and the randomized model — under the assumption that those XAI methods are not affected by the model's learned parameters.

However, we observe a significant experimental gap between top-down randomization checks when used as an evaluation measure, and occlusion-based evaluations of model faithfulness such as region perturbation [20]. Concretely, Guided Backpropagation (GB) [25] and Layer-wise Relevance Propagation (LRP) [6] exhibit low randomization scores under the first type of measure and yet clearly outperform several gradient-based methods in occlusion-based evaluations. We are interested to resolve this discrepancy.

We identify two shortcomings of top-down randomiza-

tion checks when used as a measure of explanation quality. Firstly, we show that uninformative attribution maps created with zero pixel-wise covariance — e.g., attribution maps generated from random noise — easily achieve high scores in top-down randomization checks. Effectively, this makes top-down randomization checks favor attribution maps which are affected by gradient shattering noise [7].

Secondly, we argue that the randomization-based sanity checks may always reward explanations that change under randomization, even when such randomizations do not affect the output of the model (and its invariances) significantly. Such invariance to randomization may result, e.g., from the presence of skip connections in the model, but also due to the fact that randomization may be insufficient to strongly alter the spatial distribution of activations in adjacent layers, something that we explain by the multiplicity and redundancy of positive activation paths between adjacent layers in ReLU networks. In setups which optimize parameters of attribution methods while measuring top-down randomization this might lead to the selection of explainers with higher noise.

Along with our contributed theoretical insights and supporting experiments, the present note warns against an unreflected use of model-randomization-based sanity checks as a sole criterion for selecting or dismissing a particular attribution technique, and proposes several directions to enable a more precise and informative use of randomization-based sanity checks for assessing how XAI performs on practical ML models.

1.1. Related work

Evaluating Attributions. Comparing different attribution methods qualitatively is not sufficiently objective, and for that reason, a vast number of quantitative tests have been proposed in the past in order to measure explanation quality, focusing on different desirable properties of attributions. Complexity tests [8, 9, 18] advocate for sparse and easily understandable explanations, while robustness tests [3, 8, 17] measure how much attributions change between similar samples or with slight perturbations to the input. Under the assumption of an available ground truth explanation (e.g., a segmentation mask localizing the object(s) of interest), localization tests [4, 5, 30] ask for attributed values to be concentrated on this ground truth area. Faithfulness tests [3, 8, 20] compare the effect of perturbing certain input features on the model's prediction to the values attributed to those features, so that optimally perturbing the features with the largest attribution values also affects the model prediction the most. Model randomization tests [1], which are the main focus of this work, progressively randomize the model, stating that attributions should change significantly with ongoing randomization.

Caveats of Model Randomization Tests. The authors of [1] find that a large number of attribution methods seems to be invariant to model parameters, as their explanations do not change significantly under cascading model randomization. However, various aspects of these sanity checks have recently been called into question: For instance, these tests were performed on unsigned attributions. Specifically for Integrated Gradients (IG) [27], [26] show that if the signed attributions are tested instead, this method suddenly passes cascading model randomization instead of failing. This indicates that some of the results obtained in [1] for attribution methods where the sign carries meaning may be skewed due to the employed preprocessing. Furthermore, [29] argue for the distribution-dependence of model-randomization based sanity checks. The authors demonstrate that some methods seem to fail the sanity checks in [1] due to the choice of task, rather than invariance to model parameters. A similar observation is made by [13], who find that the same attribution methods can perform very differently under model randomization sanity checks when the model and task are varied. Note that the underlying assumption of [1] — that "good" attribution methods should be sensitive to model parameters — is not called into question here. Rather, we posit that methods can fail the model randomization sanity checks for other reasons than invariance to model parameters.

2. Observation: The Gap between Randomization-based Sanity Checks and Measures of Model Faithfulness

Based on the assumption that a good explanation should attribute the highest values to the features that most affect a model's predictions, occlusion-type measures of model faithfulness [2, 8, 20] aim to quantify explanation quality by measuring the correlation between attribution map scores and changes of the model prediction under occlusion.

As such, these tests progressively randomize the data, and can thus be understood as complementary to model-randomization-based sanity checks, which progressively randomize the model. Consequently, model-randomization-based sanity checks depart towards the implausible due to partially randomized prediction models, while occlusion-based testing departs towards the implausible due to partially modified and outlier-like images. As both types of test apply the same intuition of increasing randomization to different variables (model and data) that (should) influence attribution maps, it is meaningful to compare their results, and determine whether both tests agree in terms of explanation quality.

In the following, we therefore empirically compare the scores measured by randomization-based sanity checks to the respective scores measured by faithfulness testing, for several methods. We use a variant of occlusion in the spirit

of [2] which replaces a region with a blurred copy to stay closer to the data manifold. Details on our experimental setup can be found in the Supplement (Section A.1).

As already known from [1], Guided Backpropagation [25] performs poorly under model randomization-based sanity checks when compared to three gradient-based attribution methods, namely the Gradient itself, Gradient $\times$ Input (GI) and IG. However, when measuring model faithfulness by a modified iterative occlusion test similar to [20] on the attribution maps, we find that the same GB, and also several LRP variants outperform the Gradient, Gradient $\times$ Input and Integrated Gradient substantially, as can be seen in Figure 1 and in the Supplement in Section A.3.

Due to the conceptual parallels between both tests discussed above, we find this extreme divergence surprising, and are interested in resolving this gap. Therefore, we will investigate the underlying reasons for this theoretically and experimentally in the following sections.

3. The Sensitivity of SSIM Minimization Towards Noise

The model-randomization-based sanity checks proposed by [1] use SSIM as a measure of distance between attribution maps. As we will demonstrate in this section, SSIM (and, by extension, several other distance measures, see Supplement Sections C and D) may be flawed in this application, with randomly generated attributions scoring optimally. We consider a setup where we use two different models, yielding two different attribution maps A and B. The following considerations apply to patches of the two attribution maps or whole attribution maps.

We can identify a fundamental issue: The SSIM between any two attribution maps can be minimized by a statistically uncorrelated random attribution process. This is due to the reason that the SSIM contains a product where one term relies on a covariance between two patches, see e.g. Equation 6 in [19], which is reproduced here:

$$\frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1}\frac{2\sigma_{AB} + C_2}{\sigma_A^2 + \sigma_B^2 + C_2}\,. \tag{1}$$

In the above term, μ_A, μ_B and σ_A^2, σ_B^2 are the per-patch means and variances for one patch location computed for two different input attribution maps A and B, σ_{AB} their covariance. C_1 and C_2 are constants depending on the possible input range of A and B, e.g. $[0, 1]$ or $[0, 255]$.

In the following, we will consider attribution maps within the framework of random variables. The next theorem is applied to patches of two attribution maps A, B coming from different prediction mappings, such as those obtained by a model and a partially randomized model. The patches are extracted at the same position of an image.

Theorem 1. *Consider the set of all random variables with expected means μ_A, μ_B for each image patch being fixed and with non-negative expected covariance for each patch $\sigma_{AB} \geq 0$.*

Then the expected SSIM absolute value is minimized by a random variable with zero covariance. In particular, an upper bound on the minimum is given by

$$\frac{C_2}{\sigma_A^2 + \sigma_B^2 + C_2}\,. \tag{2}$$

The proof is in the Supplement in Section B. This theorem has two consequences.

Firstly, even if we question the requirements of the theorem and thus allow negative patch correlations $\sigma_{AB} < 0$, the observation remains valid that we can obtain *very small* expected absolute values of the SSIM measure by using any randomized attribution map which is statistically independent over pixels of input images and therefore not informative.

Secondly, the proof of Theorem 1 is not affected by division of the term σ_{AB} by constants. Consequently, when using normalization on the attribution maps, the result from Theorem 1 still holds that attributions with zero patch-wise correlation attain very low scores among all normalized attribution maps.

Interestingly, this explains why certain gradient-based methods with rather noisy attribution maps pass this type of model randomization-based sanity checks with the best scores in the sense of lowest SSIM values. Gradients are known for ReLU-networks to have statistics which resemble noise processes, as has been shown in [7]. This carries over to Gradient $\times$ Input and to a lesser degree to smoothed versions like Integrated Gradient [27] and SmoothGrad [24].

Theorem 1 has one important consequence: One cannot disentangle the effects of model randomization from the amount of noise in an attribution process in model randomization sanity checks. Therefore it is problematic to use this type of model-randomization-based sanity check to compare or rank different attribution maps against each other.

4. Randomization Leaves the Model and Explanations Partly Unchanged

The section above has highlighted that explanations may score highly in the sanity check due to including further random factors in the explanation, which contradicts the principle that an explanation should faithfully depict the function to explain and not a random component. This concerns the measurement process after randomization.

In this section, we review the top-down randomization process itself. We will explain why it actually makes sense to underperform in model-based randomization checks, contrary to a first glance intuition.

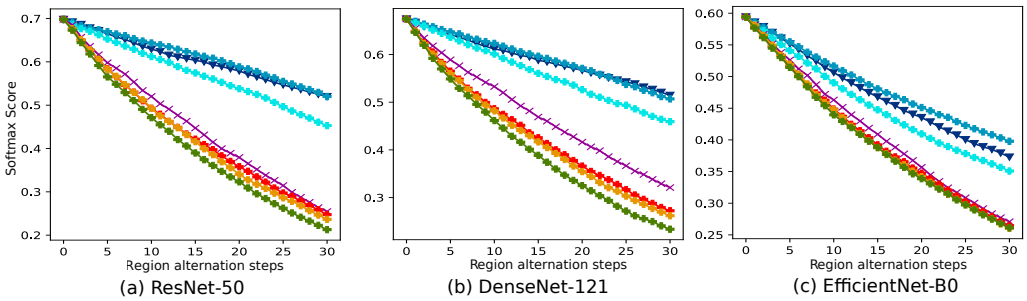

Figure 1. Results of model faithfulness via occlusion testing, by measuring the correlation to iterative occlusion with a kernel size of 15. Softmax scores are shown for Gradient, Gradient $\times$ Input, Integrated Gradients, Guided Backpropagation and several variants of LRP. The occlusion is performed by taking patches from a blurred copy of the original image. Legend is given in Figure 2. *Lower is better.*

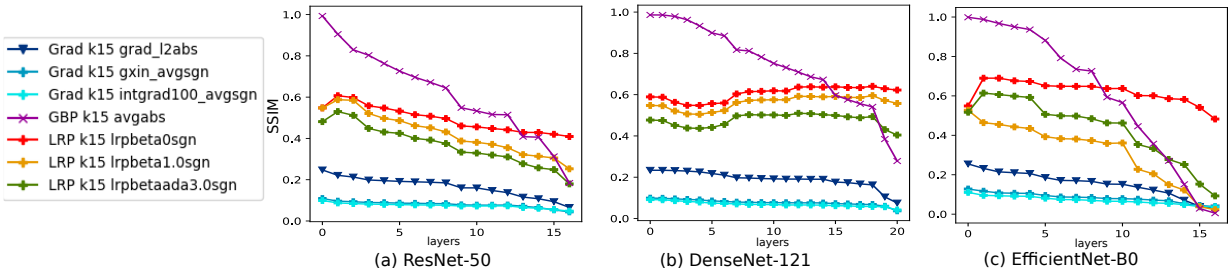

Figure 2. Results of top-down model randomization-based sanity checks with SSIM after normalization of attribution maps by their second moment. *Lower is better.*

Specifically, we will show that certain activation statistics of the network are only mildly affected by top-down randomization and thus cause low-noise explanations before and after randomization to retain some similarity.

4.1. Preservation of Irrelevance in Explanations

We first start with an empirical observation that features found to be irrelevant for a given task tend to remain irrelevant after randomization. Our experiment is based on torchvision's VGG-16 pretrained model, where we keep the mapping from the input to layer 12 unchanged and randomize the remaining layers. We apply LRP with the $z^{\mathcal{B}}$-rule in the first layer, redistribution in proportion to squared activations in the pooling layers, LRP-γ in the convolution layers with layer-wise exponential decay from $\gamma = 1.0$ to $\gamma = 0.01$, and LRP-0 in the dense layers. We inspect in Fig. 4 (right) explanations produced before and after randomization.

We observe that many spatial structures are retained before and after randomization, specifically, relevant or negatively contributing pixels are found before and after randomization on the facial and hat features, on the outline of the fish, on the finger contours, on the flagstick, on the ball, on the hole, etc. Conversely, some features remain irrelevant before and after randomization, e.g. the lake surface, the skin and the grass. Such similarities lead to similarity scores before and after randomization that remain significantly above zero, especially if considering heatmaps absolute scores.

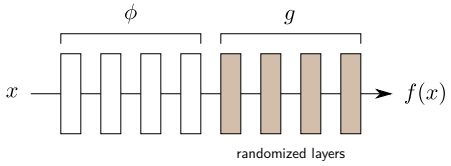

Figure 3. Diagram of a neural network where the top few layers have been randomized (shown in brown).

We now provide a formal argument showing that for an explanation to be faithful, some irrelevant features must *necessarily* remain irrelevant after randomization, thereby raising the similarity score. Let us denote by θ_R the parameters that are randomized and write the model as a composition of the non-randomized and the randomized part:

$$f(x, \theta_R) = g(\phi(x), \theta_R)\ . \tag{3}$$

The function is depicted in Figure 3. The first part ϕ contains the non-randomized layers (and can be understood as a feature extractor). The second part g contains the randomized layers (and can be interpreted as the classifier). We make the following two observations:

1. If the function ϕ is does not respond to some input feature x_i, then $g \circ \phi$ also should not respond to x_i

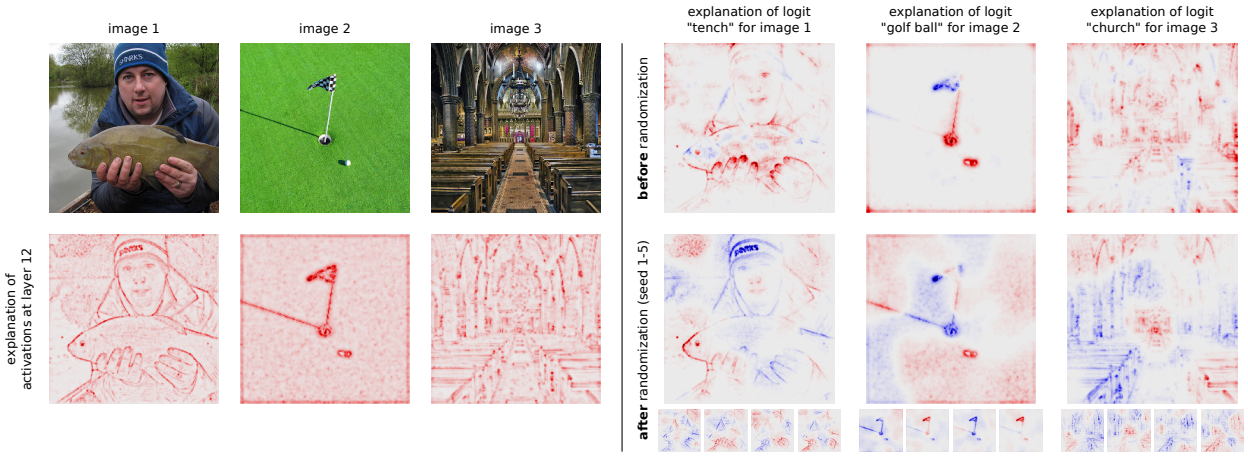

Figure 4. Experiment where one randomizes parameters of torchvision's VGG-16 pretrained model between layer 12 and the model output, and compute LRP explanations (results shown for 5 different seeds). Explanations of the neural network output before and after randomization for the true class are shown on the right. Explanation of activations at layer 12 are shown on the bottom left.

(no matter whether the function g is the classifier or its randomized variant),

2. If $g \circ \phi$ does not respond to x_i then an attribution technique should reflect this lack of response by assigning a score 0 to that feature. We note that this property of attributing low relevance input features to which the model does not respond is present in common explanation methods, for example, methods such as IG, where the gradient occurs as a multiplicative factor, LRP-type explanations, where relevance propagates mostly along connections with non-zero weights, or explanations derived from axioms such as the Shapley value whose 'null-player' axiom also relates explanation properties to model unresponsiveness.

These two observations can be summarized in the following logical clause:

$$\begin{aligned} &\phi(x) \text{ unresponsive to } x_i \\ &\quad \Rightarrow \forall g : g \circ \phi(x) \text{ unresponsive to } x_i \\ &\quad\quad \Rightarrow \forall g : \mathcal{E}_i\{g \circ \phi(x)\} \text{ small}, \end{aligned} \tag{4}$$

where $\mathcal{E}_i\{\cdot\}$ denotes the relevance of feature x_i for explaining the prediction given as argument. In other words, one should expect that any function g (randomized or not) built on ϕ shares a similar pattern of low relevances, and such a pattern originates from the lack of response of ϕ to certain input features. Therefore, we conclude that a top-down randomization process as performed in [1] can only alter explanations to a limited extent, and only a less faithful (e.g. noisy) explanation would enable further improvement w.r.t. the top-down randomization metric.

To verify that the explanation structure is indeed to some extent controlled by ϕ, we compute explanations directly at the output of the function ϕ (sum of activations) and show the results in Figure 4 (bottom left). We observe a correlation between feature relevance w.r.t. those activations and feature relevance w.r.t. the model output. For example, the lake, the grass, or more generally uniform surfaces are already less relevant at the output of ϕ, and continue to be so when considering the output of g. This is consistent with our theoretical argument that feature irrelevance of some features to classifier output g is inherited to a significant degree from the feature map ϕ.

4.2. Preservation of a Baseline Explanation

We show that for certain neural network architectures, specifically architectures that contain skip connections, a faithful explanation must further retain an additive baseline component before and after randomizations. We first demonstrate the presence of such additive component on the popular ResNet [10] model and then propose an explanation for its necessity. The ResNet is structured as a sequence of multiple modules where each module is structured as a sequence of parameterized layers, equipped with skip connections. The skip connections enable to better propagate the forward and backward signal as they simply replicate the activation and gradients from layer to layer.

Fig. 5 shows for a ResNet-34 model and the same images as before how randomizing weights at some layer affects logit scores before and after randomization. Each point in the scatter plot is one of the 1000 class logits. We observe significant correlation between the logit before and after randomization. This suggests that the model remains unchanged to a large extent and a faithful explanation should reflect such lack of change by producing a similar explanation.

Corresponding explanations are shown in Figure 7 for

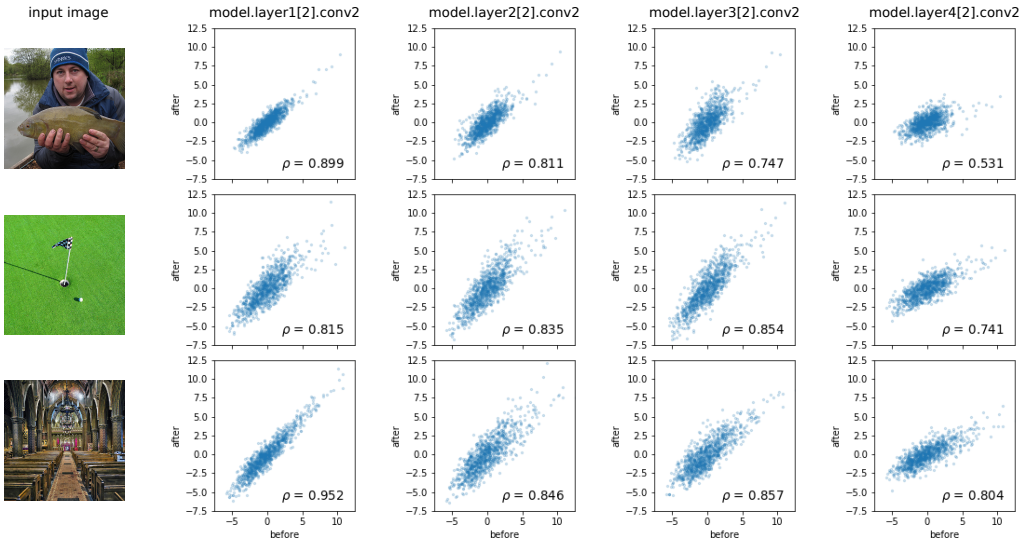

Figure 5. Effect of randomization on output logits on a ResNet-34 model for three images from ImageNet. Each point in the scatter plot is a logit for a particular class before and after randomization. Columns correspond to different layers being randomized.

the logit associated to the true class: When randomizing "`layer4.2.conv2`" of ResNet-34, the explanation remains largely the same (cf. column 2 and 5). The LRP explanation technique enables to assess contribution of different components of the neural network, and in our case, we can identify the role of the skip connection and the weighted path (cf. columns 3, 4, 6, 7). Interestingly, the explanation component that passes through the skip connection remains practically unchanged after randomization, thereby faithfully reflecting the lack of change at the output of the network (cf. Figure 5). The (weaker) contribution of the weighted path is strongly affected by randomization but its addition does not affect the overall explanation significantly.

We propose a formal argument that predicts the presence (and necessity) of an additive component for a broader range of faithful explanation methods, beyond LRP. Consider the simple architecture drawn in Fig. 6 (top) that mimics parts of a ResNet: a feature extractor, a skip connection layer, and a few top-layers. Locally approximating the top layers as a linear model (and verifying that the approximation holds under a sufficient set of perturbations of the input x), then one can decompose this approximated model in two terms, one that depends on the randomized parameter θ_R, and another term that is constant w.r.t. θ_R.

$$\hat{f}(x;\theta_R) = A(x;\theta_R) + B(x) \tag{5}$$

(cf. Fig. 6, bottom). In this model, randomization only affects the first component and thus preserves some of the original logit information. This is what we observe empirically in Fig. 5 through high correlation scores of logit before and after randomization. If we further assume usage of an

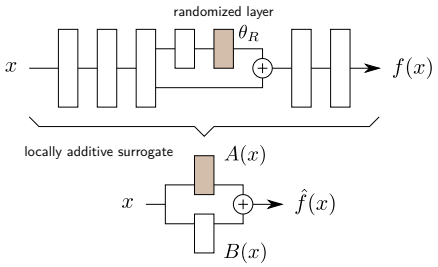

Figure 6. Top: ResNet-like structure where only one branch contains (randomizable) parameters at a particular layer. Bottom: Additive surrogate of the original model.

explanation technique satisfying the linearity property, then the explanation of the surrogate $\hat{f}$ decomposes as:

$$\mathcal{E}\{\hat{f}(x;\theta_R)\} = \mathcal{E}\{A(x;\theta_R)\} + \mathcal{E}\{B(x)\} \tag{6}$$

i.e. an explanation component that is affected by randomization and another explanation component that remains constant. This constancy under randomization prevents that optimal scores in terms of a single-layer randomization-based sanity check metric are achieved. Hence, our analysis predicts that attempts to score higher in the sanity check metric would require degrading the faithfulness of the explanation (e.g. by introduction of noise in the explanation, or by spuriously removing the additive component).

4.3. Probabilistic Preservation of Highly Activated Features in the Unrandomized Feature Layers

In this section we show that with high probability over draws of random parameters θ_R in Equation (3), regions

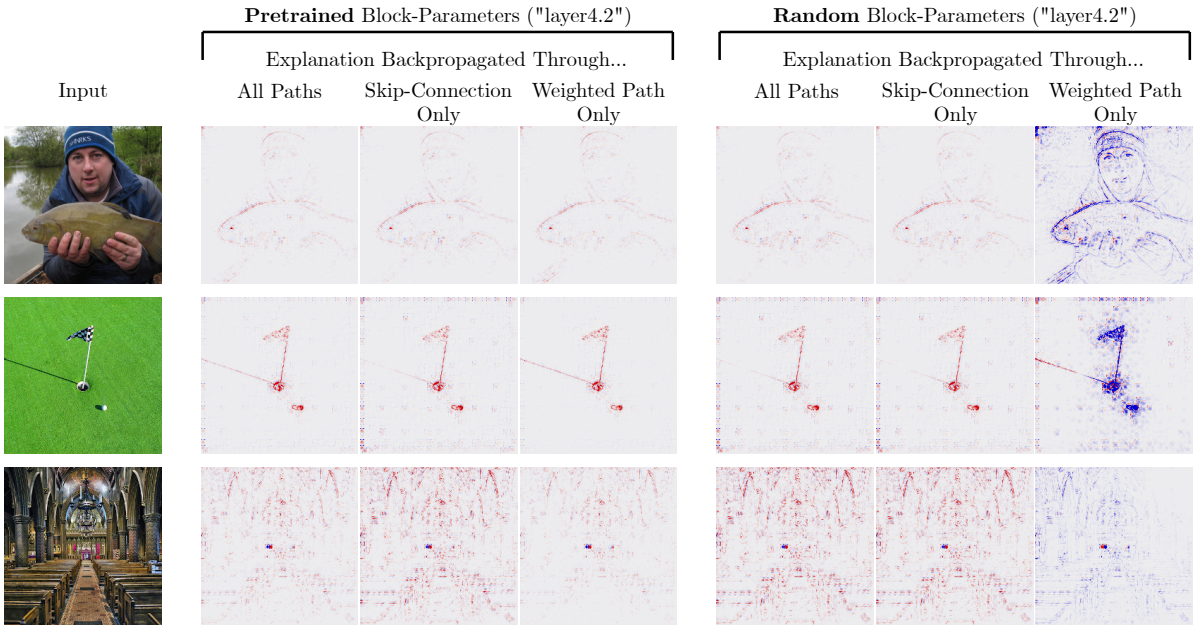

Figure 7. Effect of parameter randomization of "layer4.2.conv2" on ResNet-34 Explanations. To generate explanations, LRP-β was applied in convolution layers, and LRP-ε in dense layers. Despite random parameter re-initialization, only the part of the explanation that is propagated through the weighted path changes significantly. Due to the skip connections being unaffected, the total explanations barely change despite randomization.

of $\phi(x)$ with high activations will contribute highly to the output f, even when its value changes due to randomization in g. Unlike in the previous section, this holds for *any* ReLU network.

This observation can be explained when considering how activations are propagated to the next layer in a randomized network. One can show that small activations have a rather low probability to obtain the same average contribution to the output of a neuron as large activations, when being weighted in a linear combination with zero-mean normal weights. This statement is formulated for a single neuron in Theorem 2.

Theorem 2 (Low probability for small activations to achieve the same average contribution to the output as large activations)**.** *Suppose we have two sets of non-negative activations, X_L and X_S such that the activations of one set are by a factor of at least K larger than of the other set:*

$$\min_{x_l \in X_L} x_l \geq K \max_{x_s \in X_S} x_s \tag{7}$$

Then the probability under draws of zero-mean normal weights $w \sim N(0, \sigma^2)$ that the summed contribution of neurons in X_S surpasses the summed contribution of neurons in X_L, that is

$$0 < \sum_{x_l \in X_L} w_l x_l \leq \sum_{x_s \in X_S} w_s x_s \ , \tag{8}$$

is the tail-CDF $P(Z \geq K)$ of a Cauchy-distribution with parameter $\gamma = \sqrt{\frac{|X_S|}{|X_L|}}$ and input value of at least K.

The proof of Theorem 2 is in Supplement Section E. For probability estimates based on activation statistics of trained networks see Section I of the supplement.

To note, Theorem 2 is independent of any explanation method used. It is a statement about the preservation of relative scales of forward pass activations. It says that even though the function output value itself changes substantially under randomization, channels with large activation values still contribute highly to the output.

This effect has an impact on explanation value scales: In ReLU networks, with neurons being modeled as $y = \max(0, \sum_i w_i x_i + b)$, the differences in contributions of two inputs $w_i x_i$ to a neuron output y in many cases translate to differences in explanation scores $R(x_i)$ which the inputs x_i will receive.

Many explanation methods $R(\cdot)$ satisfy for non-decreasing activations the monotonicity property that if we consider two inputs x_i, x_j which have no other connections except to neuron y, and the network assigns positive relevance $R(y) > 0$ to y, then

$$w_i x_i \geq w_j x_j > 0 \text{ implies } |R(x_i)| \geq |R(x_j)| \ . \tag{9}$$

This holds for Gradient $\times$ Input, Shapley values, and β-LRP. See Supplement Section F for a proof.

Using the monotonicity property to go from activations to explanations, we can conclude that the probability is low to achieve equally large absolute explanation values

$\sum_{x_i \in X_S} |R(w_i x_i)|$ for inputs from the small valued set X_S in Theorem 2, when compared to $\sum_{x_i \in X_L} |R(w_i x_i)|$ from the set of large values X_L.[1]

Therefore, with high probability, explanations are also dominated by regions which have channels in the last unrandomized feature map with large activation values.

Theorem 2 and the subsequent backward pass argument hold for a single neuron. We can see that what the theoretical result predicts for a single neuron is consistent with what we observe empirically for the whole network exemplarily in Figure 4 and generally in explanations computed with GB and LRP. The above provides a theoretical justification for the exemplary observations in Figure 4, where one can see salient structures from the input image in the explanation heatmaps after top-down network randomization.

In brief, explanation heatmaps will be dominated with high probability by regions with high activations irrespective of randomization on top, thus showing limited variability under randomization. This has implications regarding the usage of top-down randomization tests to compare attribution methods: a higher variability does not imply a better explanation, when it is beyond what can be expected from the dominating high activations in the forward pass.

A further property which is preserved is shown in the Supplement Section G. The randomization-based sanity check fails to account for these necessary invariances of the model and explanation under randomization. This misalignment is particularly strong if testing the effect of randomization on the *absolute* attribution scores instead of the signed attribution scores. The necessity to use signed scores rather than absolute ones, as well as the limited change to the explanation one can expect under randomization of parameters was also emphasized in [26].

Given the discrepancy between model faithfulness measures and top-down model-based randomization checks, we remark that model faithfulness testing changes an input sample towards a partially implausible sample. Therefore it is not a perfect criterion. Another drawback is the non-uniqueness of local modifications. Different choices of local modifications will yield different measurements. However, model faithfulness testing assesses a property of explanations for a given trained model. Model randomization changes a trained model into a predominantly implausible model given the training data. Therefore it is not clear what practical aspect of a *given realistically trained* model top-down model randomization intends to measure. It seems to be unrelated to any use-case in the deployment of a fixed well-trained model.

[1] If we intend to achieve equal averaged (instead of summed) absolute explanation values $\frac{1}{|X_{S/L}|} \sum_{x_i \in X_{S/L}} |R(w_i x_i)|$, corresponding to two regions X_S and X_L with equal explanation scores, then a version of Theorem 2 holds in which $\gamma_* = \sqrt{\frac{|X_L|}{|X_S|}}$ is inverted. See Supplement Section E for a proof.

If aiming at showing sensitivity to randomization, bottom-up randomization could exhibit different (and ecologically valid) properties, because it removes strongly activated features from a model, which were the starting point in Theorem 2.

5. Conclusion and Outlook

In this paper we caution against the use of any singular method as a sole criterion to evaluate explanations, and suggest to rely on a combination of different methods for a more robust evaluation. Our study is motivated by a substantial empirical discrepancy between the scores produced by the randomization approach, and occlusion-based methods for evaluating faithfulness, in particular region perturbation [20].

Note that our theoretical and empirical results do not contradict the overall claim of [1] that a perturbation of the parameters of the model should induce a perturbation of its prediction behavior, which in turn should also perturb the explanation. The issue is rather that the similarity score should only be used as a binary test to support the presence or absence of an effect of randomization on the explanation, but not to discriminate between methods.

We have presented two main factors that explain the discrepancy between randomization-based similarity scores and the outcome of input perturbation tests: Firstly, the similarity scores used to measure the effect of randomization can be decreased artificially (and significantly) by introducing noise in the explanation. Such noise can be inherited from the gradient, which is typically highly varying and largely decoupled from the actual prediction for deep architectures.

Secondly, model randomization only alters the prediction behavior to a certain extent, often due to fixed elements in the model such as skip connections or invariances inherited from the lower layers. Hence, a maximally dissimilar explanation after randomization may not account for the partly unchanged prediction behavior of the randomized model.

These factors suggest directions for achieving a better correlation between similarity scores after randomization and evaluations of explanation faithfulness. These include (1) to only measure change w.r.t. input features to which the model is not invariant to (such features can be identified by attributing intermediate-layer activations to the input layer and retaining only input features with non-zero attribution scores), and (2) to identify the non-baseline component of the function, and only assess whether the explanation of that non-baseline component has been randomized (e.g. to exclude from the explanation what passes through the skip connections). Bottom-up randomization might be partially addressing the issue of large activations of (1).

Nevertheless, these possible refinements must address the specificity of individual architectures, thereby losing the universality of the original randomization test.

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, 3, 5, 8

[2] Chirag Agarwal and Anh Nguyen. Explaining image classifiers by removing input features using generative models. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part VI*, volume 12627 of *Lecture Notes in Computer Science*, pages 101–118. Springer, 2020. 1, 2, 3

[3] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7786–7795, 2018. 2

[4] Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Gimenez-Abalos. Focus! rating XAI methods and finding biases. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padua, Italy, July 18-23, 2022*, pages 1–8. IEEE, 2022. 1, 2

[5] Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Inf. Fusion*, 81:14–40, 2022. 1, 2

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015. 1

[7] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pages 342–350. PMLR, 2017. 2, 3

[8] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3016–3022. ijcai.org, 2020. 1, 2

[9] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1383–1391. PMLR, 2020. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[11] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.*, 37:100270, 2020. 1

[12] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham, 2019. 1

[13] Narine Kokhlikyan, Vivek Miglani, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. Investigating sanity checks for saliency maps with image and text classification. *arXiv preprint arXiv:2106.07475*, 2021. 2

[14] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, Mar 2019. 1

[15] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretations, interpretability, trustworthiness, and beyond. *arXiv preprint arXiv:2103.10689*, 2021. 1

[16] Grégoire Montavon. *Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison*, pages 253–265. Springer International Publishing, Cham, 2019. 1

[17] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018. 2

[18] An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020. 2

[19] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 3

[20] 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 Net-*

works and Learning Systems, 28(11):2660–2673, 2017. 1, 2, 3, 8

[21] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE*, 109(3):247–278, 2021. 1

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[23] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified BP attributions fail. In *International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pages 9046–9057. PMLR, 2020. 1

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[25] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015. 1, 3

[26] Mukund Sundararajan and Ankur Taly. A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1806.04205*, 2018. 2, 8

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017. 1, 2, 3

[28] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[29] Gal Yona and Daniel Greenfeld. Revisiting sanity checks for saliency maps. *arXiv preprint arXiv:2110.14297*, 2021. 2

[30] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.*, 126(10):1084–1102, 2018. 1, 2