Evaluating Input Perturbation Methods for Interpreting CNNs and Saliency Map Comparison

Lukas Brunke, Prateek Agrawal, and Nikhil George
Volkswagen Group of America, Belmont CA 94002, USA
prateek.agrawal@vw.com

Abstract. Input perturbation methods occlude parts of an input to a function and measure the change in the function’s output. Recently, input perturbation methods have been applied to generate and evaluate saliency maps from convolutional neural networks. In practice, neutral baseline images are used for the occlusion, such that the baseline image’s impact on the classification probability is minimal. However, in this paper we show that arguably neutral baseline images still impact the generated saliency maps and their evaluation with input perturbations. We also demonstrate that many choices of hyperparameters lead to the divergence of saliency maps generated by input perturbations. We experimentally reveal inconsistencies among a selection of input perturbation methods and find that they lack robustness for generating saliency maps and for evaluating saliency maps as saliency metrics.

Keywords: Saliency methods, saliency maps, saliency metrics, perturbation methods, baseline image, RISE, MoRF, LeRF

1 Introduction

Understanding and interpreting convolutional neural networks’ (CNN) predictions through saliency methods has become an active field of research in the recent years [3,14,15,17]. Saliency methods create saliency maps, which highlight relevant parts of the input image for a classification task.

Input perturbation methods, which are also referred to as occlusion methods [2], are one of the saliency methods for understanding CNNs. Input perturbation methods follow a simple principle: covering up or masking an important part of an input results in relevant information loss and should reduce the prediction score for a specific class. If an occluded part of the image results in an increase of the prediction score, then this part of the input will be negatively correlated with the target class. Furthermore, there might be parts of the input, which when covered, do not affect the prediction score. Occluding or masking an input is defined as substituting specific elements with a baseline image. The masking of an input image with a given baseline image and a mask is illustrated in Figure 1. In practice, neutral images are chosen as the baseline image, in such a
way that the baseline image’s impact on the classification probability is minimal. Furthermore, input perturbation methods are also used to compare and evaluate saliency maps from different saliency methods for the same input image. Input perturbation methods have two major advantages over other saliency methods like gradient-based methods. First, input perturbation methods allow the analysis of black-box models, whose weights and gradients are inaccessible. This is, for example, relevant in the automotive industry for validating black box models received from suppliers. Second, input perturbation methods’ saliency maps are easily interpretable by humans unlike gradient-based methods, which often result in noisy or diffuse saliency maps [11,2]. In contrast to these advantages of input perturbation methods, a disadvantage is that they are typically more computationally intensive.

In this paper we experimentally evaluate the robustness of perturbation methods for image classification tasks against different parameters and baseline images. We pick three representative input perturbation methods to illustrate our findings: Randomized Input Sampling for Explanation (RISE) [12], most relevant first (MoRF) [13], and least relevant first (LeRF). However, our results also apply to other input perturbation methods. The goal is to determine the effects of changing the perturbation methods’ parameters on the reliability of the perturbation method. We evaluate the generation of saliency maps with RISE by varying its parameters and selecting various baseline images, which are used to occlude an input. MoRF, and LeRF are metrics used to objectively compare different saliency maps for the same input by increasingly perturbing the original input [13]. We also vary the baseline images for MoRF and LeRF. The experiments reveal shortcomings regarding the robustness of input perturbation methods.

The next section introduces the investigated perturbation methods. Section 3 gives an overview of related work. The material used for the experiments is presented in Section 4. The experiments on RISE, MoRF, and LeRF are

Fig. 1: Masking an input image $X$ with a baseline image $A$. The input is elementwise multiplied with the mask $M$, keeping all the parts of the input, where the mask has a value of 1, and setting the parts of the input to zero, where the mask has a value of 0. The baseline image is elementwise multiplied with the inverse mask $(1_{m \times n} - M)$, which is the matrix of ones minus the mask, and added to the masked input, such that the parts, which were set to 0 for the input are set to the corresponding value of the baseline image. The displayed mask shows an intermediate mask from LeRF.
described in detail in Section 5 and subsequently discussed in Section 6. We
conclude our findings in Section 7.

2 Input Perturbation Methods

Perturbation methods rely on a baseline image [15] to occlude parts of the origi-
nal input, see Figure 1. The application of a baseline image
$A \in \mathbb{R}^{m \times n \times 3}$ with
mask $M \in [0, 1]^{m \times n}$ to an input image $X \in \mathbb{R}^{m \times n \times 3}$ is defined in [7,9] as:

$$\Phi(X, A, M) = M \odot X + (1 - M) \odot A,$$

where $\odot$ denotes the Hadamard product and $1_{m \times n}$ is an $m \times n$-dimensional
matrix of ones. The mask can also be of a different dimension, depending on the
scaling and cropping operations [12].

No systematic pattern of how to choose a suitable baseline image seems
identifiable in the literature. It has been argued to be a neutral input such
that $\tilde{f}(X) \approx 0$, where $\tilde{f} : \mathbb{R}^{m \times n \times 3} \rightarrow \mathbb{R}^{K}$ is a CNN function and $K$ is the
number of predicted classes [15]. The function $f = \tilde{f} \circ g$ is the DNN function
with preprocessing, where $g : \mathbb{R}^{m \times n \times 3} \rightarrow \mathbb{R}^{m \times n \times 3}$ is the preprocessing function
applied to the input $X$ of a DNN. The black image $A = 0_{m \times n \times 3}$, with the
$m \times n \times 3$-dimensional matrix of zeros $0_{m \times n \times 3}$, is the baseline image used in
[15]. Samek et al. applied baseline images from a uniform distribution, Dirichlet
distribution, constant baseline images and blurred baseline images [13]. Petsiuk
et al. use the baseline image for which $g(A) = 0_{m \times n \times 3}$ [12]. We refer to this
baseline image as the zero baseline image after preprocessing.

2.1 Saliency Map Generation: RISE

Petsiuk et al. introduced RISE for creating saliency maps for interpreting CNNs
for classification tasks [12]. RISE generates a saliency map by applying $N$
random masks to an input and creating a weighted sum of the $N$
output probabilities for the target class $c$ with the random masks. RISE uses a set of $N$
random masks $\mathcal{M}_N(p, w, h) = \{M_1(p, w, h), ..., M_N(p, w, h)\}$, where $M_i(p, w, h) \in \{0, 1\}^{w \times h}, i \in \{1, ..., N\}$ and the probability $p \in [0, 1]$ of each element of the
mask to be equal to 1 or in this context to be "on". The parameters $w$ and $h$
indicate the size of the low resolution mask before scaling and cropping. In the
following $w = h$, since we only use square masks and input images. Each mask
is scaled to the input image’s size with a transformation $q : \{0, 1\}^{w \times h} \rightarrow [0, 1]^{m \times n}$,
which applies a bilinear transformation and a random cropping operation. For a
detailed description of the mask generation for RISE we refer the reader to the
original paper [12]. The saliency map $S$ is then given by:

$$S = \frac{1}{\mathbb{E}(\mathcal{M}_N)} \cdot \sum_{i=1}^{N} \tilde{f}^c(g(M_i) \odot g(X)),$$
where $\tilde{f}^c$ are the outputs for class $c$ for the DNN function without preprocessing. Directly applying the mask to the preprocessed input is equal to implicitly using the zero after preprocessing baseline image, therefore, using a constant baseline image for the saliency map generation:

$$\tilde{f}^c(q(M_i) \odot g(X)) = f^c(\Phi(X, g^{-1}(0_{m \times n \times 3}), q(M_i))$$

with the inverse preprocessing function $g^{-1}$ and $f^c$, which is the output for class $c$ for the DNN function with preprocessing.

### 2.2 Saliency Map Comparison: MoRF and LeRF

The MoRF and LeRF metrics quantitatively evaluate the quality of saliency maps [13]. This is relevant for comparing and assessing different saliency methods like RISE [12], Grad-CAM [14], and integrated gradients [15]. MoRF and LeRF measure the effect of occluding pixels from the input image on the target class' output probabilities. MoRF and LeRF replace elements from the input with the corresponding elements from a baseline image. Given a saliency map for an input image as a ranking of importance for individual pixels, MoRF replaces pixels in the input image in decreasing order of the importance (= most relevant first), whereas LeRF replaces pixels in the input image in increasing order of importance (= least relevant first). Replacing pixels in the input image with the corresponding pixels from the baseline image results in a change of the output probabilities. Recording the output probabilities as further elements are occluded results in a curve over the relative number of occluded elements $\alpha$. The number of additionally occluded pixels $r$ in each step is variable. Here, all presented experiments are run with $r = 1$, occluding only one additional pixel in each step.

MoRF iteratively replaces elements with decreasing importance according to a given ordering by a saliency map. A greater area over the curve (AOC) for MoRF is desirable and suggests a superior saliency map. On the other hand, LeRF iteratively replaces elements with increasing importance and its score is given by the area under the curve (AUC). In this paper we limit the experiments to the MoRF and LeRF metrics, because the MoRF and LeRF metrics can be transformed into the insertion and deletion metrics, which are used in [9,12].

### 3 Related Work

There has been a limited effort to determine the robustness of input perturbation methods against varying parameters and baseline images. Ancona et al. [2] have tested different square sizes for masks for the sliding window method [17]. Increased square sizes exhibited a reduction of details provided in the saliency maps. This inspired our experiments on varying the $w$ and $p$ parameters for RISE, which affect the shape and size of the random masks. Similarly to Ancona et al. [2], we find that less fine grained saliency maps are created by less
fine grained masks. This corresponds to small values for $w$ in the case of RISE and bigger values for $n$ in the case of Occlusion-$n$.

Samek et al. [13] have investigated various baseline images for the MoRF and LeRF metrics. They determined that the uniform distribution baseline image gives the best MoRF and LeRF scores when averaged over the whole dataset. Samek et al. applied MoRF and LeRF with $r = 9 \times 9$ non-overlapping regions and only perturbed up to 15.7% of the input image. Furthermore, they evaluated the blurred baseline image with $\sigma = 3$, which is arguably small and still contains information from the original input.

Fong and Vedaldi [9] generate saliency maps by optimizing perturbation masks, which minimize or maximize the output probabilities. They find that different baseline images yield different saliency maps, which highlights the dependency on the choice of baseline image. Dabkowski and Gal [7] note similar findings and therefore use a set of baseline images to circumvent the dependence on a single baseline image. Petsiuk et al. [12] determined through qualitative reasoning for MoRF and through quantitative analysis for LeRF, that a constant gray baseline image and a blurred baseline image work best for their evaluation using MoRF (= deletion) and LeRF (= insertion), respectively. However, our quantitative analysis suggests that this is not the case.

Recently, [16] have investigated the reliability of saliency metrics like MoRF and LeRF. Using two different baseline images, specifically the constant dataset mean baseline image and a uniform noise baseline image, they show that MoRF and LeRF are dependent on the choice of the baseline image. However, since they used a random baseline image for comparison, the baseline image changes for each run of MoRF and LeRF, which lacks consistency and could affect their findings and yield issues regarding reproducibility. In contrast, we use a set of constant baseline images (with the exception of the blurred baseline image). Additionally, they only analyzed the scores for MoRF and LeRF for 100 steps, whereas we run both metrics for all steps.

4 Material

In this work we are exclusively considering CNNs for image classification. Specifically we are using the pre-trained CNN ResNet-50 [10] from Keras [6] in TensorFlow [1]. We process the input in the same way as described by [10], such that the input shape is $224 \times 224 \times 3$. Then we apply a mean shift, which is the preprocessing function used for the pretrained network.

All experiments are run with the validation set from ImageNet [8], which we refer to as dataset $D_{\text{val}}$. Saliency maps are visualized by normalizing the resulting maps in the range $[0, 1]$ and applying OpenCV’s perceptually uniform sequential inferno colormap [4]. This colormap displays parts supporting the prediction with lighter colors and parts opposing the prediction with darker colors.
5 Experiments

This section presents experiments on the robustness of input perturbation methods against different parameter settings and different baseline images. First, we investigate how saliency maps change under varying parameters for the mask generation with RISE. Second, we analyze the impact of using different supposedly uninformative baseline images on the robustness of RISE, MoRF, and LeRF.

5.1 Experiments on RISE

Petsiuk et al. apply RISE with $N = 8000$ for ResNet-50, $h = w = 7$, $p = 0.5$ and the zero baseline image after preprocessing [12]. The choice of these parameters is not elaborated on in [12]. Therefore, this section determines the effect of varying the parameters on the convergence of a saliency map and the subjective quality of the saliency map. Throughout this section, saliency maps are generated for their maximum activated class $c_{\text{max}} = \arg \max_c f^c(x)$. In contrast to Petsiuk et al. [12], we applied $N_{\text{max}} = 32768$ masks to evaluate the convergence. Unless otherwise stated the experiments are executed with $h = w = 7$, $p = 0.5$, $N = N_{\text{max}}$, and the zero baseline image after preprocessing.

First, the number of masks needed for convergence of a single saliency map is determined. In order to check the convergence RISE is run three times, with three independent random sets of masks $\mathcal{M}_{l N_{\text{max}}}$, where $l \in \{1, 2, 3\}$ and $N_{\text{max}} = 32768$. Figure 2 displays the resulting saliency maps for an input $X$ after applying $N_{\text{max}}$ masks. Figure 2 shows that the saliency maps from multiple runs have converged, since they are nearly indistinguishable from each other.

The visual assessment of convergence for saliency maps is not efficient, the goal is to determine convergence quantitatively. We define a function $d$ that calculates the similarity between two saliency maps $S_i$ and $S_j$: $d(S_i, S_j)$. The similarity $d$ is defined as the normalized cross-entropy between two maps.

\[ d(S_i, S_j) = -\frac{1}{\log_2|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log_2 \frac{S_i(x)}{S_j(x)} \]

For the class bull_mastiff, the convergence plot is shown in Figure 2(e). The plot displays the similarity $d$ between different sets of saliency maps $\{S_1, S_2\}$, $\{S_1, S_3\}$, and $\{S_2, S_3\}$. The similarity decreases as the number of masks $N$ increases, indicating that the saliency maps converge.

Fig. 2: Convergence for RISE for the class bull_mastiff.
L²-distance between two saliency maps $S_i$ and $S_j$, where $i \neq j, i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3\}$, from different runs and records the L²-distance for every incremental saliency map. In practice, we only consider saliency maps with a maximum distance between saliency maps from independent runs smaller than $d_{\text{max}} = 2000$, which is represented by the dashed horizontal line in Figure 2e. We chose this threshold, because it indicates the maximum in the histogram in Figure 3. The histogram shows the L²-distances for RISE saliency maps from 1000 randomly sampled input images from ImageNet. Manual inspection confirmed that the threshold also yields subjectively good results. Note, that for this input image the use of the suggested number of masks $N = 8000$, which is represented by the dotted vertical line in Figure 2e, results in $d > d_{\text{max}}$ for some combinations of $i$ and $j$.

![Figure 3: Selection of $d_{\text{max}} = 2000$.](image)

The input image and saliency maps in Figure 4 show an example, where the standard parameters for $w$ and $p$ with $N_{\text{max}} = 32768$ masks do not lead to the same saliency map for the three independent runs. The graph in Figure 4e shows that the L²-distance between the saliency maps in Figure 4 never get close to the threshold $d_{\text{max}}$, and therefore do not converge. Doing the same investigation on a random subset of 1000 images from the validation set from ImageNet, yields only 389 converged saliency maps. In the following we investigate the variation of the $p$ and $w$ parameters. We observed, that the example in Figure 4a-e converged when setting $p = 0.1$, which is shown in Figure 4f-j. Note, that the threshold $d_{\text{max}}$ is only reached when $N$ approaches $N_{\text{max}}$. Applying different values for $w$, while $p$ is constant, results in distinct saliency maps, which also impacts the average L²-distance $\bar{d}$ between saliency maps from different runs at $N_{\text{max}}$. Figure 5 shows the effect of selecting different $w \in \{3, 4, 5, 7\}$. In the two examples, an increasing value for $w$ leads to more defined distinct local optima in the saliency maps, which focus on the individual objects of the target class. While $\bar{d}$ decreases with increasing $w$ for the example in Figure 5a-e, $\bar{d}$ decreases with increasing $w$ for the example in Figure 5f-j.

Subsequently, the impact of varying $p$ and $w = h$ jointly is investigated. We analyzed the variation of $p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $h = w \in \{5, 7, 9, 11\}$, where each set of parameters is run three times to determine convergence. In
Figure 4: Convergence for RISE with $p = 0.5$ at the top and $p = 0.1$ at the bottom for the class goldfish. The choice of $p = 0.5$ does not lead to convergence, while $p = 0.1$ converges. We found that only 389 of 1000 randomly sampled input images led to convergence with $p = 0.5$.

Figure 6 we only report the saliency maps, which converged. Saliency maps with $p \in \{0.5, 0.7, 0.9\}$ did not converge for the input image from Figure 4. For some combinations the maximum number of masks $N_{\text{max}} = 32768$ might not be sufficient for the saliency maps to converge. However, we did not run any experiments with $N > N_{\text{max}}$, since the application of more masks increases the number of calculations, which leads to computational infeasibility in practical settings.

The systematic choice of a suitable baseline image is still an open research question [7,9,12,15,17]. The proposed condition for the baseline image is, that it is uninformative [15] and therefore different baseline images have been presented in the literature [13]. The following gives an overview over the types of baseline images considered in our work. The constant baseline images are given by $A_{\gamma} = \gamma \cdot \mathbf{1}_{m \times n \times 3}, \forall \gamma \in \Gamma \subseteq \{0, \ldots, 255\}$. Note that $\gamma = 0, \gamma = 127, \text{and } \gamma = 255$ refer to a black, gray and white baseline image, respectively. The zero baseline image after preprocessing, which is used by [12], is $A_{\text{inv}} = g^{-1}(\mathbf{0}_{m \times n \times 3})$. Applying the
Fig. 5: Convergence for RISE with varying $w$ for the classes Yorkshire terrier and pug, respectively. The choice of $w$ can either reduce or increase $\bar{d}$.

Fig. 6: RISE with different values for $(w, p)$ for the class goldfish. The choice of different hyperparameters yield different saliency maps.
preprocessing function $g$ to $A_{\text{inv}}$, yields $g(A_{\text{inv}}) = 0_{m \times n \times 3}$. The blurred baseline image is defined as $A_{\sigma} = \pi(X, \sigma), \forall \sigma \in \Sigma$, where $\pi$ applies a Gaussian blur with standard deviation $\sigma$ and $\Sigma$ is the set of standard deviations. In the past, values for the standard deviation of $\sigma = 3$ [13] and $\sigma = 5$ [12] have been chosen, however, applying these standard deviations can result in an output probability for the target class much greater than 0, which is not an uninformative baseline image. Therefore, we follow the choice of $\sigma = 10$ as in [7,9], which yields a baseline image with output probability of approximately 0. Note, that the blurred baseline image is the only baseline image which uses local information from the current input image $X$. We combine the above baseline images in a set of baseline images:

$$A(\Gamma, \Sigma, g) = \{A_{\gamma}, A_{\text{inv}}, A_{\sigma}, \forall \gamma \in \Gamma, \forall \sigma \in \Sigma\}.$$  

(4)

Specifically the set $A_{e} = A(\{0, 127, 255\}, \{10\}, g)$ is used in the following experiments. The set could also be extended to include a baseline image sampled from a uniform or Dirichlet distribution, the constant average from the current input, and a constant input image averaged over the dataset. However, the chosen set of baseline images is already sufficient for raising robustness issues for input perturbation methods. We adapt RISE to enable the application of different baseline images:

$$S = \frac{1}{E(M_N) \cdot N} \sum_{i=1}^{N} f_c(\Phi(X, A, q(M_i))),$$  

(5)

where $A \in A_{e}$. Each saliency map is generated for examples $X \in D_{\text{val}}$ for the class $c_{\text{max}}$ with the highest output probability for that input image and $N = 16384$. This choice of $N$ is a compromise between accuracy and computational effort. If the $L^2$-distance among saliency maps for an input image is greater than $d_{\text{max}}$, the input will not be considered for this experiment. Therefore, we only take into account examples, which have converged. An overview of four examples, where different baseline images yield different results, is given in Figure 7. The top row shows the tested input images and the two bottom rows each display the application of a different baseline image when creating the RISE saliency map for the specific input image. Depending on the baseline image, RISE attributes more or less importance to certain parts of the input. Comparing Figures 7g and 7k, fewer or more people are highlighted based on the choice of baseline image. Similarly, in Figures 7h and 7l the bird’s left wing and the bird’s feet are highlighted, respectively. Depending on the selected baseline image the saliency map for RISE can completely change. RISE is therefore highly dependent on the choice of the baseline image.

5.2 Experiments on MoRF and LeRF

MoRF and LeRF provide an objective and automatable metric for assessing different saliency methods. Here, we use MoRF to compare one saliency map generated by RISE with one generated by Grad-CAM [14]. We limit the presented
Fig. 7: RISE with different baseline images.

Experiments to MoRF and refer the reader to the supplementary material [5] for experiments on LeRF, which yield the same propositions. RISE and Grad-CAM have been chosen for demonstration purposes only. The choice of different saliency methods for the comparison is equally reasonable. The generation of each saliency map uses the standard parameters presented in the original paper for Grad-CAM [14]. We run MoRF with $r = 1$, such that the input images are covered pixel by pixel. The function $m : [0, 1] \times [0, 1]^{224 \times 224} \rightarrow \{0, 1\}^{224 \times 224}$ gives the mask induced by a saliency map $S$, where the relative occluded area of the input image is given by $\alpha$.

In the first set of experiments we investigate the behavior of MoRF for different single input images. Similar as in [12], we also only use the zero after preprocessing baseline image for MoRF and the blurred baseline image for LeRF. In the examples in Table 1, we find that comparing the AOC for MoRF for the single input images results in a conflicting assessment on which saliency method is preferable. For example, for the input image on the left side of Table 1, the AOC for the saliency map from Grad-CAM is greater than the AOC for the saliency map from RISE. However, for the input image on the right the AOC for the saliency map generated by RISE is greater. Note, that the output probability of the original input for the target class is displayed at $\alpha = 0$, when the image is not covered at all. According to the graphs in Table 1 the selected baseline images at $m(1, S) = A$ also fulfill the requirement of being uninformative since the output probabilities are close to zero.
Table 1: MoRF with the zero after preprocessing baseline image for input images of the classes chimpanzee (left) and beacon (right).

| Input | RISE | Grad-CAM |
|-------|------|-----------|
| AOC   | 0.5173 | 0.8521 |
| AOC   | 0.8289 | 0.7145 |

Table 2: MoRF with different baseline images for the class flamingo.

| Baseline | A_σ | A_inv |
|----------|-----|-------|
| RISE AOC | 0.8260 | 0.7874 |
| Grad-CAM AOC | 0.7720 | 0.8077 |
The subsequent set of experiments illustrate the dependency of the input perturbation methods on the choice of the baseline image. Again we use MoRF to compare one saliency map generated by RISE with one generated by Grad-CAM. Here only the results for the $A_\sigma$ and $A_{inv}$ baseline images are reported, because these are the baseline images used for the insertion and deletion metrics by [12]. However, the results below can also be observed when applying different baseline images. The results for MoRF on different input images are shown in Table 2. Each graph compares the saliency maps from RISE and Grad-CAM. The table shows that by only changing the baseline image, the result of MoRF inverts. While one saliency method is superior using a specific baseline image, it is inferior using another baseline image. Furthermore, RISE performs better than Grad-CAM with the $A_\sigma$ baseline image for MoRF but worse than Grad-CAM using the same baseline image for LeRF. Therefore, the results from MoRF and LeRF are highly dependent on the choice of the baseline image. Note, that even though the RISE saliency map was created using the zero after preprocessing baseline image, the AOC for MoRF with the blurred baseline image $A_\sigma$ is greater than the AOC for MoRF with the zero after preprocessing baseline image. This shows that running MoRF with the same baseline image, that was used to create the RISE saliency map, does not lead to a greater AOC for RISE.

6 Discussion

RISE does not converge with the parameters presented by Petsiuk et al. [12] for every input image. Therefore, we implemented a convergence check with three independent sets of masks using the $L^2$-distance. As shown by the experiments, different sets of parameters do not necessarily converge for all input images. Furthermore, the experiments show that varying the parameters for the mask generation can change the resulting saliency map significantly. In the case of Figure 4 multiple goldfish are visible and RISE does not converge using the standard parameters. In this example, the probability of at least one goldfish being visible is very high when covering half the image ($p = 0.5$). This results in a high prediction score for each masked input image. This could be one possible explanation, why the RISE saliency map is not able to converge in this case. The above example is more likely to converge if fewer goldfish are revealed with each single mask. In practice, a smaller value for $p$ led to convergence, if multiple instances of the same object are visible in the input image. While $w$ can impact convergence, we did not find indicators for a correlation between $w$ and the object size. In general, it is not clear how to decide on the parameters for running RISE on an input image. Determining the set of parameters requires multiple executions of RISE, which increases the computational effort. It is also not obvious how to select one saliency map for assessment, in the case that multiple sets of parameters result in a converged saliency map for an input image. The presented experiments on RISE also raise the issue of reproducibility of saliency maps with RISE.
The application of MoRF and LeRF on single input images when using the same baseline images, shows that the outputs for MoRF and LeRF do not agree for the saliency maps for the different input images in a dataset, which supports the recent findings in [16]. While, [13] average the AOC scores for MoRF or AUC scores for LeRF over the whole dataset, we want to highlight that MoRF and LeRF are not normalized. We argue that this puts input images with a low output probability for the target class at a disadvantage, since they generally exhibit a lower AOC for MoRF or a lower AUC for LeRF. For example, consider MoRF and two input images, which receive an output probability for the target class of 0.4 and 0.9, respectively. The maximum possible AOC for the first image is significantly lower than the maximum possible AOC for the second one.

The results from MoRF, LeRF, and RISE suggest, that these metrics and methods are highly dependent on the choice of the baseline image. Each baseline image introduces a bias into the input image. Even an all zero input can yield activations for the ResNet-50 because of nonzero bias terms. The choice of a neutral baseline image is therefore nontrivial. Creating saliency maps from RISE with different baseline images can result in highlighting contradicting parts of the input image. Consequently, a specific baseline image can lead to an incorrect conclusion when assigning importance to certain elements in the input. Similarly, the experiments on the variation of baseline images display that by changing baseline images it is possible to invert the proposition from MoRF and LeRF on which saliency method is superior.

7 Conclusion

In this paper we investigated input perturbation methods and metrics under changing parameters and baseline images. First, the experiments on RISE revealed, that convergence is not guaranteed for the majority of analyzed input images. Convergence could still be achieved for certain inputs, by varying the parameters for $w$, $h$ and $p$. However, different sets of parameters yield significantly distinct results, which might contradict each other by highlighting different parts of the input image. Determining the best parameters by multiple independent runs of RISE with different parameters and checking convergence for each input image increases the amount of required computations. Furthermore, these additional runs and variations still do not guarantee non-contradicting saliency maps from RISE. Second, varying the baseline image has led to notably different saliency maps for RISE and changed the AOC and AUC values for MoRF and LeRF, respectively, such that a reliable evaluation of saliency methods is challenging.

Our results suggest that input perturbation methods are unreliable for understanding CNNs' predictions and the current available methods should not be used in practice. Since RISE, MoRF, and LeRF lack desirable robustness properties, future work will explore the development of robust input perturbation methods along with additional experiments for determining their robustness.
References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Man´e, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015), http://tensorflow.org/

2. Ancona, M., Ceolini, E., Öztireli, A.C., Gross, M.: A unified view of gradient-based attribution methods for Deep Neural Networks. In: NIPS Workshop on Interpreting, Explaining and Visualizing Deep Learning (2017)

3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7), e0130140 (07 2015)

4. Bradski, G.: The OpenCV Library. Dr. Dobb’s Journal of Software Tools (2000)

5. Brunke, L., Agrawal, P., George, N.: Supplementary Material: Evaluating Input Perturbation Methods for Interpreting CNNs and Saliency Map Comparison (2020)

6. Chollet, F., et al.: Keras. https://keras.io (2015)

7. Dabkowski, P., Gal, Y.: Real Time Image Saliency for Black Box Classifiers. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 6967–6976. Curran Associates, Inc. (2017)

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)

9. Fong, R.C., Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

11. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: Advances in Neural Information Processing Systems 32 (2019)

12. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)

13. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Transactions on Neural Networks and Learning Systems 28(11), 2660–2673 (Nov 2017)

14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (Oct 2017)

15. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. In: ICML (2017)

16. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., Preece, A.: Sanity Checks for Saliency Metrics. Proceedings of the AAAI Conference on Artificial Intelligence (2020)
17. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 818-833. Springer International Publishing, Cham (2014)