

# Aggregating explanation methods for stable and robust explainability

Laura Rieger<sup>1</sup> Lars Kai Hansen<sup>1</sup>

## Abstract

Despite a growing literature on explaining neural networks, no consensus has been reached on how to explain a neural network decision or how to evaluate an explanation. Our contributions in this paper are twofold. First, we investigate schemes to combine explanation methods and reduce model uncertainty to obtain a single aggregated explanation. We provide evidence that the aggregation is better at identifying important features, than on individual methods. Adversarial attacks on explanations is a recent active research topic. As our second contribution, we present evidence that aggregate explanations are much more robust to attacks than individual explanation methods.

## 1. Introduction

Despite the great success of neural networks especially in classic visual recognition problems, explaining the networks' decisions remains an open research problem (Samek et al., 2019). This is due in part to the complexity of the visual recognition problem and in part to the basic 'ill-posedness' of the explanation task. This challenge is amplified by the fact that there is no agreement on what a sufficient explanation is and how to evaluate an explanation method.

Many different explanation strategies and methods have been proposed (Simonyan et al., 2013; Zeiler & Fergus, 2014; Bach et al., 2015; Selvaraju et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017). Focusing on visual explanations for individual decisions, most methods either use a backpropagation approach or aim to construct a simpler linear model with an intuitive explanation. The plethora of explanation approaches is a signature of the high-level epistemic uncertainty of the explanation task.

This paper is motivated by a key insight in machine learn-

<sup>1</sup>DTU Compute, Technical University Denmark, 2800 Kgs. Lyngby, Denmark. Correspondence to: Laura Rieger <lauri@dtu.dk>.

ing: Ensemble models can reduce both bias and variance compared to applying a single model. A related approach was pursued for *functional* visualization in neuroimaging (Hansen et al., 2001). Here we for the first time explore the potential of aggregating explanations of individual visual decisions in reducing epistemic uncertainty for neural networks.

We test the hypothesis that ensembles of multiple explanation methods are more robust than any single method. This idea is analyzed theoretically and evaluated empirically. We discuss the properties of the aggregate explanations and provide visual evidence that they combine features, hence are more complete and less biased than individual schemes. Based on this insight, we propose two ways to aggregate explanation methods, *AGG-Mean* and *AGG-Var*. In experiments on Imagenet, MNIST, and FashionMNIST, the aggregates identify relevant parts of the image more accurately than any single method and are more robust to adversarial attacks.

## 2. Related Work

### 2.1. Explanation methods

The open problem of explainability is reflected in a lot of recent work (Kindermans et al., 2017; Selvaraju et al., 2017; Bach et al., 2015; Zhang et al., 2018a; Zhou et al., 2016; Ancona et al., 2018; Ribeiro et al., 2016; Rieger et al., 2018; Kim et al., 2018; Lundberg & Lee, 2017; Zintgraf et al., 2017; Simonyan et al., 2013; Zeiler & Fergus, 2014; Selvaraju et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017; Shrikumar et al., 2017; Montavon et al., 2017; Chang et al., 2018). We focus on generating visual explanations for single samples. To our knowledge the first work in this direction was Simonyan et al. (2013) with *Saliency Maps* (*SM*) that proposed backpropagating the output onto the input to gain an understanding of a neural network decision. The relevance for each input dimension is extracted by taking the gradient of the output w. r. t. to the input. This idea was extended by Springenberg et al. (2014) into *Guided Backpropagation* (*GM*) by applying ReLU non-linearities after each layer during the backpropagation. Compared to Saliency, this removes visual noise in the explanation. *Grad-CAM* (*GC*) from Selvaraju et al. (2017) is an explanation method, developed for use with convolutional neural net-

works. By backpropagating relevance through the dense layers and up-sampling the evidence for the convolutional part of the network, the method obtains coarse heatmaps that highlight relevant parts of the input image. *Integrated Gradients* (*IG*) Sundararajan et al. (2017) sums up the gradients from linearly interpolated pictures between a baseline, e.g. a black image, and the actual image. *SmoothGrad* (*SG*) filters out noise from a basic saliency map by creating many samples of the original input with Gaussian noise (Smilkov et al., 2017). The final saliency map is the average over all samples.

Finally, we also consider *LIME* (Ribeiro et al., 2016). In contrast to the other methods, *LIME* is not based on back-propagation. Instead, it approximates the neural network with a linear model locally around the input to be explained. The coefficients of the linear model for the respective input dimensions give the importance of each dimension. Compared to the other methods this is much more computationally expensive as it requires many passes through the neural network.

Considering the plethora of explanation methods, evaluating them objectively has been a topic of recent work. In this paper we will evaluate explanation methods using two approaches:

Ancona et al. (2018) proposed a different approach to evaluate explanation methods, called Sensitivity- $n$ , based on the notion that the decrease in output when a number of inputs are canceled out should be equal to the sum of their relevances. For a range of  $n$  (between 1 and the total number of inputs) they sample a hundred subsets of the input. For each  $n$ , the Pearson Correlation Coefficient (PCC) between the decrease in output, when the subset of features is removed, and the sum of their relevances is reported. The result is a curve of the PCC dependent on the percentage of the input being removed. For a good explanation method, the PCC will decrease slowly.

We will also use a variant of the method proposed by (Samek et al., 2016) as introduced in (Anonymous, 2020). High-dimensional images get divided up into squares. Squares with high relevance (as measured by the explanation method to be evaluated) consecutively get replaced with a non-informative baseline. The difference between the original output and the output for the iteratively degraded images indicates the quality of the explanation method. For good explanation methods that accurately identify areas of high relevance, the output will go down earlier as relevant areas are replaced earlier. We calculate the area over this curve. For details we refer to (Anonymous, 2020), which is provided in the supplements (currently under review elsewhere).

## 2.2. Adversarial attacks on explanation methods

While adversarial examples for classification are well-known, recently there has been growing interest in adversarial manipulation of explanations (Ghorbani et al., 2019; Heo et al., 2019; Dombrowski et al., 2019). Attacks on explanation can serve multiple purposes including "fairwashing" (Aïvodji et al., 2019). All of these methods exploit the fully differentiable nature of neural networks and iteratively update the input (or the model weights) to change the explanation while only minimally changing the input and output. The goal is to manipulate the explanation while keeping the input and output (visually) similar. It is assumed that the network architecture and weights are known and that either the input ((Ghorbani et al., 2019; Zhang et al., 2018b; Dombrowski et al., 2019)) or the network weights (Heo et al., 2019) can be changed by the attacker.

Focussing on changing the input, (Ghorbani et al., 2019), (Zhang et al., 2018b) and (Dombrowski et al., 2019) attack the explanation by manipulating the image, not changing the network weights. Interestingly, (Zhang et al., 2018b) discuss the transferability of attacks and conclude that attacks are not that transferable. If the attacker is allowed to modify networks weights, as in (Heo et al., 2019), the attacks generalize to all the considered explanation methods.

(Zhang et al., 2018b) also discuss various strategies where the adversarial attacks the label without modifying the explanation and vice versa. The former is interesting if the explanations is used to validate the decision.

Here we are interested in the situation where the attacker can modify the input but not the network. We investigate the transferability, c.f., (Zhang et al., 2018b), and hypothesise that the limited transferability leads to improved robustness of the ensemble explanation. While ensemble methods have been proposed earlier as a defense for attacks on the label (Tramèr et al., 2017), they have not previously been investigated as a defense mechanism against attacks on explanations.

## 3. Aggregating explanation methods to reduce variance

All currently available explanation methods have weaknesses that are inherent to the approach and include significant uncertainty in the resulting heatmap (Kindermans et al., 2017; Adebayo et al., 2018; Smilkov et al., 2017). A natural way to mitigate this issue and reduce noise is to combine multiple explanation methods. Ensemble methods have been used for a long time to reduce the variance and bias of machine learning models. We apply the same idea to explanation methods and build an ensemble of explanation methods.

We assume a neural network  $F : X \mapsto y$  with  $X \in \mathbb{R}^{m \times m}$  and a set of explanation methods  $\{e_j\}_{j=1}^J$  with  $e_j : X, y, F \mapsto E$  with  $E \in \mathbb{R}^{m \times m}$ . We write  $E_{j,n}$  for the explanation obtained for  $X_n$  with method  $e_j$  and denote the mean aggregate explanation as  $\bar{E}$  with  $\bar{E}_n = \frac{1}{J} \sum_{j=1}^J E_{j,n}$ . While we assume the input to be an image  $\in \mathbb{R}^{m \times m}$ , this method is generalizable to inputs of other dimensionalities as well.

To get a theoretical understanding of the benefit of aggregation, we hypothesize the existence of a 'true' explanation  $\hat{E}_n$ . This allows us to quantify the error of an explanation method as the mean squared difference between the 'true' explanation and an explanation procured by an explanation method, i.e. the MSE.

For clarity we subsequently omit the notation for the neural network. We write the error of explanation method  $j$  on image  $X_n$  as  $\text{err}_{j,n} = \|E_{j,n} - \hat{E}_n\|^2$  with

$$\text{MSE}(E_j) = \frac{1}{N} \sum_n \text{err}_{j,n}$$

and  $\text{MSE}(\bar{E}) = \frac{1}{N} \sum_n \|\bar{E}_n - \hat{E}_n\|^2$  is the MSE of the aggregate. The typical error of an explanation method is the mean error over all explanation methods

$$\overline{\text{MSE}} = \frac{1}{J} \sum_j \text{MSE}(E_j).$$

With these definitions we can do a standard bias-variance decomposition (Geman et al., 1992). Accordingly we can show the error of the aggregate will be less than the typical error of explanation methods,

$$\begin{aligned} \overline{\text{MSE}} &= \frac{1}{N} \sum_n \frac{1}{J} \sum_j \|\hat{E}_n - E_{j,n}\|^2 & (1) \\ &= \frac{1}{N} \sum_n \|\hat{E}_n - \bar{E}_n\|^2 & (2) \\ &\quad + \frac{1}{NJ} \sum_{n,j} \|\bar{E}_n - E_{j,n}\|^2, \end{aligned}$$

hence,

$$\begin{aligned} \overline{\text{MSE}} &= \frac{1}{J} \sum_j \frac{1}{N} \sum_n \underbrace{\|\bar{E}_n - E_{j,n}\|^2}_{\text{epistemic uncertainty}} + \text{MSE}(\bar{E}) & (3) \\ &\geq \text{MSE}(\bar{E}). \end{aligned}$$

A detailed calculation is given in the appendix. The error of the aggregate  $\text{MSE}(\bar{E})$  is less than the typical error of the participating methods. The difference - a 'variance' term - represents the epistemic uncertainty and only vanishes if all methods produce identical maps. By taking the average over all available explanation methods, we reduce the variance of the explanation compared to using a single method. To obtain this average, we normalize all input heatmaps such that the relevance over all pixels sum up to one. This reflects our initial assumption that all individual explanation methods

are equally good estimators. We refer to this approach as *AGG-Mean*.

$$E_{\text{Agg-Mean},n} = \frac{1}{J} \sum_{j=1}^J E_{j,n} \quad (4)$$

This estimator however does not take into account the estimate of the local epistemic uncertainty, i.e. the disagreement between methods. A way to incorporate this information is to form an 'effect size' map by dividing the mean aggregate locally with its standard deviation (Sigurdsson et al., 2004). Intuitively, this will assign less relevance to segments with high disagreement between methods.

For stability, we divide not directly by the local variance but add a constant  $\epsilon$  to the estimate of the local variance. This can be interpreted as a smoothing regularizer or a priori information regarding epistemic and aleatoric uncertainties. We refer to this approach as *AGG-Var*.

$$E_{\text{AGG-Var},n} = \frac{1}{J} \sum_{j=1}^J \frac{E_{j,n}}{\sigma(E_{j \in J, n}) + \epsilon} \quad (5)$$

where  $\sigma(E_{j \in J, n})$  is the point-wise standard deviation over all explanations  $j \in J$  for  $X_n$

#### 4. Aggregating explanation methods to reduce vulnerability

With the increasing interest and practical importance of explainability of neural networks the interest in methods for manipulation and control of explanations is also increasing. A typical scenario is to make imperceptible changes to the input of the neural network such that the output/label is unchanged while the explanation changes according to a given goal. Such effort could, e.g., be used to hide bias or other fairness issues a given classifier might have .

(Dombrowski et al., 2019) showed that explanations can be made more robust by replacing the ReLU nonlinearity with a Softplus function. However, this requires changing the network and using a different architecture for classification and explanation, which is highly undesirable as it defeats the purpose of the explanation. The analysis of (Zhang et al., 2018b) showed that transferability of attacks is limited, hence, our ensemble of multiple explanations may offer robustness also towards certain types of adversarials.

In the following we will assume an attacker who has full knowledge of the neural network, including the architecture and weights. In contrast to (Heo et al., 2019), however, we will assume that the attacker cannot *change* the neural network, following (Dombrowski et al., 2019; Ghorbani et al., 2019). Furthermore the attacker has full control over the input to the neural network. The goal is to adversarially manipulate the image to according to a predefined objective.

In the following,  $x$  will refer to the original image.  $\hat{x}$  is the ‘target’ image. The objective is to produce an adversarial input  $x'$  with  $x' \approx x$  but the explanation  $E_{x'} \approx E_{\hat{x}}$ . While we focus on assimilating the explanation map of another input as in (Dombrowski et al., 2019), all techniques introduced can be readily adapted to other objectives, f.e. to move the mass center of the explanation.

Exploring the robustness of aggregates of multiple explanation methods we concentrate on the following two scenarios:

**Arsenal of explanation methods** In this scenario we have a pool of potential explanation methods. The attacker does not have knowledge of what explanation method is used and optimizes for a different explanation method than is used by the defender. The success of the attack depends on how readily an attack of one explanation method translates to another method.

We hypothesize that attacks on explanation methods are fragile and do not translate well across explanation methods, as they exploit locally high variances in the gradient landscape. This hypothesis is examined in section 5.5.

**Aggregation of explanation methods** In this more challenging scenario we aggregate multiple explanation methods as described in eq. (4). The attacker knows the exact explanation methods and ratio going into the mixture and attacks this aggregation.

Many attribution-based methods are utilizing the gradient  $\frac{\delta y}{\delta x}(x)$  of the output to create an explanation. Due to the non-linearity of the neural network, the gradient can change rapidly with small distances in input space (Dombrowski et al., 2019; Ghorbani et al., 2019). Attacks on explanation methods exploit this vulnerability.

AGG-Var as defined in eq. (5) will not be considered as a defence against adversarial manipulation since training with an objective containing a variable in both the nominator and denominator is computationally undesirable. While this is an issue for the attack, it also makes evaluation considerably harder.

## 5. Experiments

We first evaluate aggregations of explanation techniques against vanilla techniques with a variation of the evaluation method proposed in (Bach et al., 2015), Sensitivity- $n$  and qualitative evaluation. Subsequently we evaluate the effectiveness of aggregating methods against adversarial attacks on explanation methods in section 5.5. In the appendix we compare aggregated methods on a dataset of human-annotated heatmaps.

### 5.1. Experimental details

We tested our method on five neural network architectures that were pre-trained on ImageNet: VGG19, Xception, Inception, ResNet50 and ResNet101 (Deng et al., 2009; Simonyan & Zisserman, 2014; He et al., 2016; Chollet, 2017; Szegedy et al., 2016).<sup>1</sup> Additionally, we ran experiments on CNN trained on the MNIST and FashionMNIST dataset (LeCun & Cortes, 2010; Xiao et al., 2017).

### 5.2. Quantitative evaluation

To quantitatively compare the quality of the explanation methods on a more challenging dataset, we use an evaluation method similar to the approach described in (Samek et al., 2016) (see section 2). Instead of dividing the image into squares, we use an established segmentation algorithm, SLIC, to divide the image into semantically meaningful areas (Achanta et al., 2012). This makes the information that is being removed each step more meaningful and reduces variance in the result (Anonymous, 2020) A manuscript with a detailed description of this approach, including an evaluation of evaluation methods is currently under review elsewhere and included in the supplements (Anonymous, 2020).

We compared the aggregation methods against Saliency (SM), Guided Backpropagation (GB), SmoothGrad (SG), Grad-CAM (GC) and Integrated Gradients (IG) to have a selection of attribution-based methods. Additionally we compared against LIME as a method that is not based on attribution but rather on local approximation (Ribeiro et al., 2016). The aggregations are based on all attribution-based methods.

Some of the methods result in positive and negative evidence. We only considered positive evidence for the ImageNet tasks to compare methods against each other. To check that this does not corrupt the methods, we compared the methods that do contain negative results against their filtered version and found negligible difference between the two versions of a method in the used metrics. For AGG-Var we introduced an additional stabilizing parameter,  $\epsilon$  to the divisor. In our experiments we set  $\epsilon$  to be ten times the mean  $\sigma$  over the entire dataset.

In table 1 we show the results of this evaluation. High scores indicate that the explanation method correctly labels parts of the image that are relevant for the classification as important. We include two non-informative baselines. *Random* randomly chooses segments to remove. *Sobel* is a sobel edge detector. Neither of them contain information about the neural network.

All explanation methods have a higher evaluation score

<sup>1</sup>Models retrieved from <https://github.com/keras-team/keras>.

Table 1. Scores across methods and architectures. *AGG-Mean* and *AGG-Var* surpass all methods in all scenarios. All SE < 0.05.

METHOD	INCEPTION	RESNET101	RESNET50	VGG19	XCEPTION
LIME	74.0	67.6	66.7	73.4	74.0
SM	76.2	75.6	76.2	81.4	72.6
GB	73.7	77.6	80.9	84.3	74.8
IG	76.0	76.0	76.2	79.8	73.3
SG	77.9	77.7	77.5	83.9	75.0
GC	78.7	77.8	78.5	86.1	73.5
AGG-MEAN	79.9	81.1	80.9	86.6	<b>76.7</b>
AGG-VAR	<b>80.4</b>	<b>81.2</b>	<b>81.0</b>	<b>86.7</b>	<b>76.7</b>

than the random baseline on all architectures tested, indicating that all methods contain information about the image classification. Except for LIME, all methods also surpass the stronger baseline, *SOBEL*. The ranking of unaggregated methods varies considerably between architectures. This variance indicates that the accuracy of an explanation method depends on the complexity and structure of the neural network architecture. For all architectures *AGG-Mean* and *AGG-Var* have a higher score than any non-aggregated method. For ResNet101 the difference between the best unaggregated method and the aggregated methods is especially large. We hypothesize, that the benefit of aggregating explanation methods increases for more complex neural network with large epistemic uncertainty on the explanation. We empirically confirmed that aggregating methods improves over unaggregated methods and more reliably identifies parts of the images that are relevant for classification.

### 5.3. Qualitative visual evaluation

We show heatmaps for individual examination for each of the methods in fig. 1 and compare qualitatively. While visual evaluation of explanations for neural networks can be misleading, there is no better way available of checking whether any given explanation method agrees with intuitive human understanding (Adebayo et al., 2018). Additionally, we compute alignment between human-annotated images and the explanation methods in the appendix, using the human benchmark for evaluation introduced in (Mohseni & Ragan, 2018).

*AGG-Var* combines features of the aggregated methods by attributing relevance on the classified object as a whole, but considering smaller details such as the face of an animal as more relevant. It is a combination of the detail-oriented and shape-oriented methods. Compared to SmoothGrad, which concentrates on one isolated feature, the relevance is more evenly distributed and aligned with our human intuition that classification takes context into account and does not rely on e.g. only the beak of a bird. We conclude that combining explainability methods provides a meaningful visual improvement over single methods.

### 5.4. Evaluation with Sensitivity-*n* on low-dimensional input

To quantitatively compare explanation methods on a low-dimensional input we use Sensitivity-*n* (Ancona et al., 2018). The exact procedure is described in section 2. We compare on MNIST (LeCun & Cortes, 2010) and FashionMNIST (Xiao et al., 2017), two low-dimensional dataset with a basic CNN<sup>2</sup> (architecture in appendix). We follow the procedure suggested in (Ancona et al., 2018) and test on a hundred randomly sampled subsets for 1000 randomly sampled test images. The number of pixels in the set *n* is chosen at fifteen points logarithmically spaced between 10 and 780 pixels.

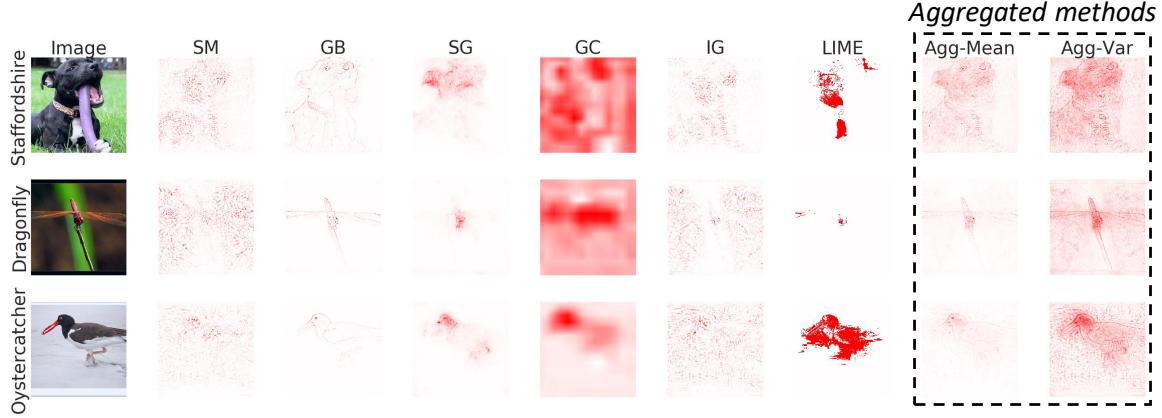
As described in section 2, for a range of *n* (between 1 and the total number of inputs) a hundred subsets of the input features are removed. For each *n*, the average Pearson Correlation Coefficient (PCC) between the decrease in output and the relevance of the removed output features is reported. The result is a curve of the PCC dependent on the removed percentage.

We show results in fig. 2. *AGG-Mean* and *AGG-Var* perform in range of the best methods. For the FashionMNIST CNN *AGG-Mean* and *AGG-Var* perform better than unaggregated methods. For the MNIST CNN Guided Backpropagation and *AGG-Mean* perform best. Considering that MNIST has a nearly binary data representation (black/white), 'removing' a pixel by setting it to black is informative in itself, making Sensitivity-*n* less reliable for networks trained on this particular dataset. For both networks (trained on FashionMNIST and MNIST respectively), SmoothGrad and GradCAM perform considerably worse than the other methods.

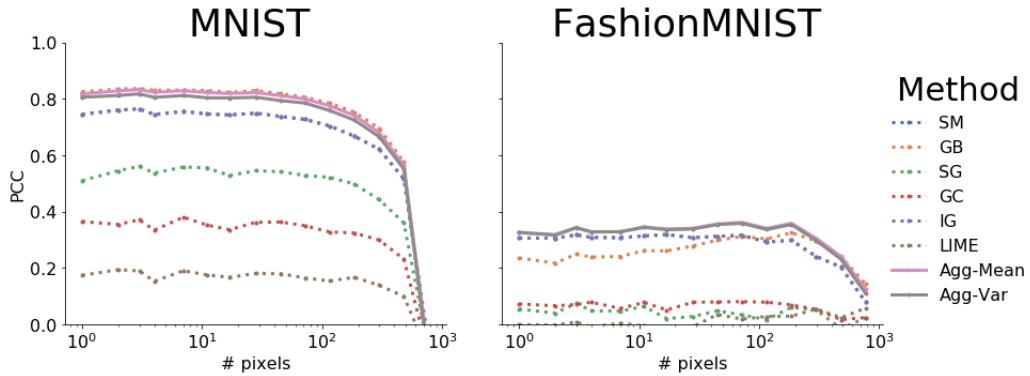
In summary, aggregation seems to not be as beneficial when applied to a low-dimensional, "easier" tasks such as MNIST as it is for ImageNet, performing in range of the best unaggregated method. We hypothesize that this is because there is less epistemic uncertainty in explanations for less complex tasks and network architectures.

<sup>2</sup>Model and code retrieved from [https://github.com/keras-team/keras/blob/master/examples/mnist\\_cnn.py](https://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py).

## Aggregating explanation methods



**Figure 1.** Example images from Imagenet and the heatmaps produced by different methods on VGG19. Aggregated methods combine features from all methods. Too heavy focus on one feature by SmoothGrad is smoothed away by the aggregation.



**Figure 2.** Sensitivity- $n$  for explanation methods. Higher is better. The proposed methods, AGG-Mean and AGG-Var perform better or equally good as all other methods.

### 5.5. Attacking explanation methods

We evaluate how robust aggregated methods are against adversarial attacks, compared to unaggregated methods. In all cases we assume that the attacker has full knowledge of the network architecture and weights (white box attack) but cannot change them. However, the attacker has full control over the input.

Following (Dombrowski et al., 2019) we run experiments on a pretrained VGG16 (Simonyan & Zisserman, 2014) and consider Layerwise Relevance Propagation (LRP), Saliency Mapping (SM), Guided Backprop and Integrated Gradients. The latter was not used in the aggregation. Unless otherwise noted we followed (Dombrowski et al., 2019) in the choice of hyperparameters for attacking explanation methods. Since the ReLU function used in neural networks is not twice differentiable, we replace it with a differentiable approximation, SoftPlus for the iterative creation of the adversarial input. The final manipulated heatmaps are cre-

ated with the ReLU non-linearity. Further details about the experiments are in the appendix.

We consider the two scenarios introduced in section 5.5. In all cases, the objective of the attacker is to make the explanation of input  $E_{x'} \approx E_{\hat{x}}$  while keeping  $x' \approx x$ . To do this, he manipulates  $x'$ .

We visually confirmed that the adversarial images look similar to the input images and provide examples in the appendix. We measure the difference between the start explanation  $E_x$  target explanation  $E_{\hat{x}}$  and adversarial explanation  $E_{x'}$  with the MSE (Mean Square Error), the PCC (Pearson Correlation Coefficient) and the top- $k$  intersection with  $k$  being ad-hoc set to 10% (Dombrowski et al., 2019; Ghorbani et al., 2017).

In all metrics, explanations obtained with different methods have different 'base' values (similarity between the explanations of two randomly chosen images) due to structural

differences between explanation methods. To account for this, we consider for each similarity metric  $m_{\text{sim}}$  the difference  $m_{\text{sim}}(E_{\hat{x}}, E_{x'}) - m_{\text{sim}}(E_{\hat{x}}, E_x)$ , i.e. how much *more* similar the attack makes  $E_{x'}$  look to  $E_{\hat{x}}$ . For the MSE, this results in a negative score, since the difference between the target and the attack is less than between the target and the starting point. For all metrics, the ideal score is 0, i.e. the attack did not change the explanation at all. Thus, for MSE a high value is desirable, for PCC and top- $k$  union a low value is desirable.

**Transferability of attacks on explanation methods** We consider a case where the attacker does not know what explanation method is used, i.e. we attack a different explanation method than the one that is used. If the attack translates well, i.e. the image manipulation fools both methods, similarity metrics should be similar for both explanation methods.

In fig. 3 we provide results for attacking Guided Backpropagation and extracting an explanation with LRP. For a hundred samples we visualize for each sample the respective similarity metrics for both explanation methods in fig. 3. If the attack translates well, the points should lie on the identity line in fig. 3. Samples below the identity line for PCC and topK and above for MSE indicate that the attack does not generalize to other explanation methods.

As visible in fig. 3 and anecdotally in fig. 4 (red data point in fig. 3), attacks perform much worse on other methods (here LRP) than the targeted one (here GB). We provide statistics for other combinations in the appendix.

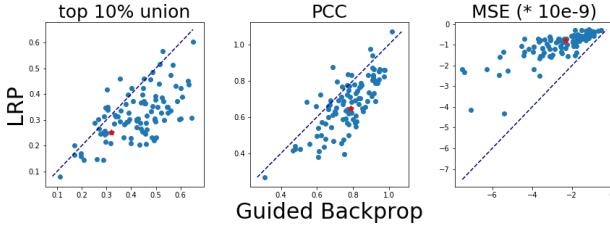


Figure 3. Attacking one method does not translate to attacks on the other methods. Similarity metrics (topK and PCC) should be low, MSE should be high. The red dot is the sample visualized in fig. 4

The lack of transferability results are in line with the findings of (Zhang et al., 2018b).

**Attacking aggregations of explanation methods.** In the second scenario the attacker knows that the explanations are aggregated and attacks the aggregation. We aggregate LRP, GB and SM and compare against those methods as well as Integrated Gradient. IG was not included in the aggregation as it requires sampling for each step, making it computationally much more expensive than the other methods.

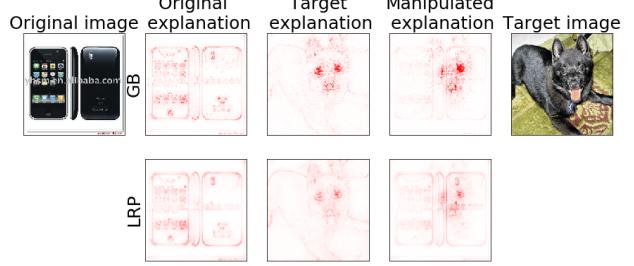


Figure 4. Example showing the transferability of attacks. The adversarial input was calculated to attack GB (upper row). We then extracted the explanation with LRP (lower row). The attack does not transfer well to LRP. To visualize details better we clipped values at the 99th percentile.

In table 2 we provide metrics averaged over a hundred samples. Agg-Mean outperforms unaggregated methods. We also provide a direct comparison to GuidedBackprop in fig. 5. To give an intuition on what differences int the metrics look like, we visualize a sample (red dot in fig. 5) in fig. 6. We see that Agg-Mean opposed to the unaggregated methods largely preserves the original heatmap after the attack. More examples are provided in the appendix.

The resilience of the aggregate to attacks can be understood in terms of averaging induced smoothness. In (Dom-browski et al., 2019) the beneficial effects of averaging in the SmoothGrad method are described. As noted in (Dom-browski et al., 2019) SmoothGrad is computationally expensive. We conjecture that the diversity of the methods involved in the present aggregate implies that smoothing can be achieved at less computational effort.

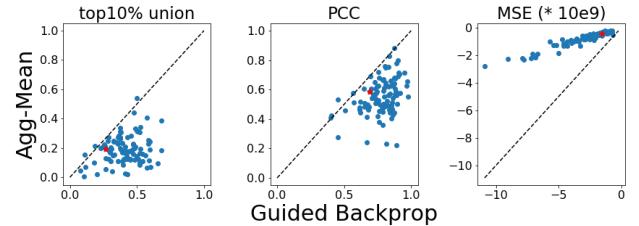
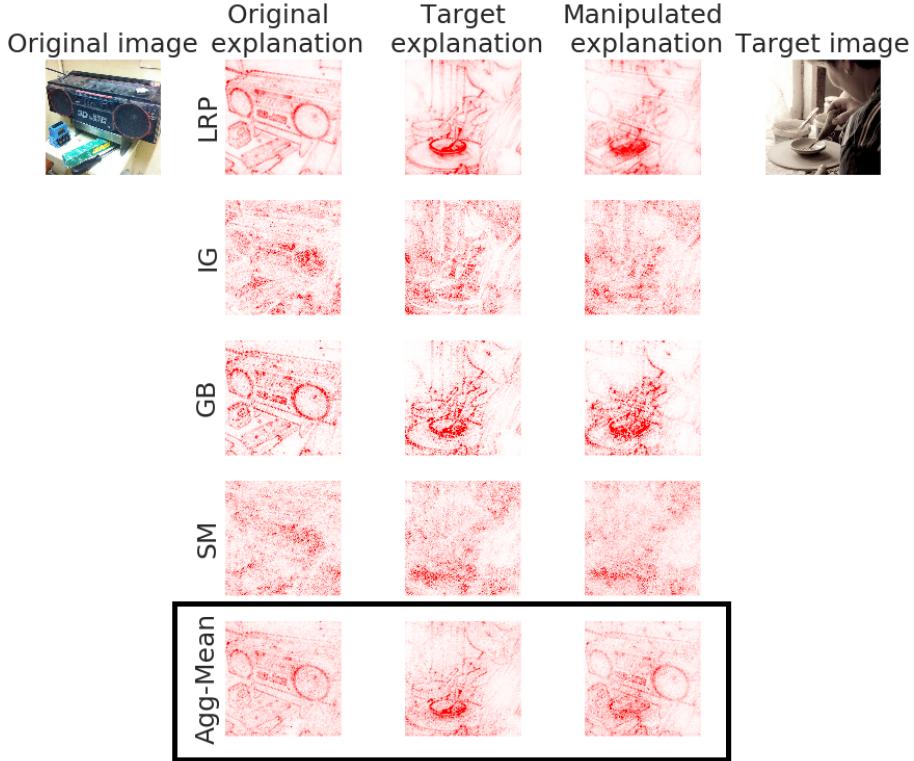


Figure 5. Visualizing the success of an attack on AGG-Mean compared to unaggregated method (Guided Backprop). Similarity metrics (topK and PCC) should be low, MSE should be high for less similarity between target and adversarial. In the majority of cases, AGG-Mean is more robust than Guided Backprop. The red dot is the sample visualized in fig. 6

## 6. Conclusion

The explanation problem for object detection using neural networks is ill-posed: Many schemes have been proposed leaving considerable epistemic uncertainty. We propose to

### Aggregating explanation methods



*Figure 6.* Attacking explanation methods. We provide the adversarial input images in the appendix. AGG-Mean visibly preserves original explanation best, thus being the most resistant to adversarial attacks. To visualize details better we clipped values at the 99 percentile.

*Table 2.* Evaluation scores across methods and architectures on a hundred samples. *AGG-Mean* surpasses all considered methods in all metrics. Deviation is SE.

METHOD	MSE (*10E-9)	PCC	TOP 10% UNION
SM	-0.92 ± 0.00	0.74 ± 0.01	0.40 ± 0.01
GB	-3.25 ± 0.02	0.77 ± 0.01	0.42 ± 0.01
LRP	-1.45 ± 0.01	0.81 ± 0.01	0.49 ± 0.01
IG	-1.76 ± 0.01	0.82 ± 0.01	0.47 ± 0.01
AGG-MEAN	<b>-0.89 ± 0.01</b>	<b>0.54 ± 0.01</b>	<b>0.24 ± 0.01</b>

mitigate this by aggregating multiple explanation methods. This serves multiple purposes, reducing variance representing the epistemic uncertainty, reducing bias by extending the range of features that an explanation can highlight and reducing vulnerability to adversarial attacks.

We gave a simple proof that aggregating explanation methods will perform at least as good as the typical individual method. In practice, we found evidence that aggregating methods outperforms any single method. This evidence was substantiated across quantitative metrics. While our results show that different vanilla explanation methods perform best on different network architectures, an aggregation supersedes all of them on any given architecture.

Attacks on explanation has received considerable recent attention. We provided experimental evidence that aggregations are a more robust to adversarial manipulation than individual explanation methods. In (Dombrowski et al., 2019) arguments are presented that the observed vulnerability is due to non-smoothness of contemporary networks. It is also argued that averaging as in SmoothGrad increases robustness. We conjectured that the averaging of the diverse set of explanation methods involved in the aggregate creates similar smoothness. We noted that in contrast to (Dombrowski et al., 2019), the aggregate does not require modification (smoothing) of the network.

## References

- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9525–9536, 2018.
- Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. Fairwashing: the risk of rationalization. *arXiv preprint arXiv:1901.09749*, 2019.
- Ancona, M., Ceolini, E., Oztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.
- Anonymous. Irof: A low resource evaluation metric for explanation methods. 2020.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
- Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. Explaining image classifiers by counterfactual generation. 2018.
- Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. IEEE, 2017.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.
- Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, 2019.
- Geman, S., Bienenstock, E., and Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.
- Ghorbani, A., Abid, A., and Zou, J. Interpretation of Neural Networks is Fragile. oct 2017.
- Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.
- Hansen, L. K., Nielsen, F. Å., Strother, S. C., and Lange, N. Consensus inference in neuroimaging. *NeuroImage*, 13(6):1212–1218, 2001.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Heo, J., Joo, S., and Moon, T. Fooling neural network interpretations via adversarial model manipulation. *arXiv preprint arXiv:1902.02041*, 2019.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2673–2682, 2018.
- Kindermans, P.-J., Hooker, S., Adebayo, J., Brain, G., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un)reliability of saliency methods. In *Proceedings Workshop on Interpreting, Explaining and Visualizing Deep Learning (at NIPS)*, 2017.
- LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.
- Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.
- Mohseni, S. and Ragan, E. D. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*, 2018.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.
- Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.
- Rieger, L., Chormai, P., Montavon, G., Hansen, L. K., and Müller, K.-R. Structuring Neural Networks for More Explainable Predictions. pp. 115–131. Springer, Cham, 2018.

- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Muller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature, 2019.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and 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. IEEE, 2017.
- Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153, 2017.
- Sigurdsson, S., Philipsen, P. A., Hansen, L. K., Larsen, J., Gniadecka, M., and Wulf, H.-C. Detection of skin cancer by classification of raman spectra. *IEEE transactions on biomedical engineering*, 51(10):1784–1793, 2004.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Simonyan, K., Vedaldi, A., and Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. dec 2013. URL <http://arxiv.org/abs/1312.6034>.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. 06 2017.
- Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2014.
- Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.
- Vedaldi, A. and Soatto, S. Quick shift and kernel methods for mode seeking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5305 LNCS, pp. 705–718, Berlin, Heidelberg, oct 2008. Springer Berlin Heidelberg. ISBN 3540886923. doi: 10.1007/978-3-540-88693-8-52.
- Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. aug 2017. URL <https://arxiv.org/abs/1708.07747>.
- Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.
- Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.
- Zhang, Q., Nian Wu, Y., and Zhu, S.-C. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2018a.
- Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., and Wang, T. Interpretable deep learning under fire. *arXiv preprint arXiv:1812.00891*, 2, 2018b.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.
- Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. In *ICLR*, pp. 3, 2017.

## A. Appendix

### A.1. Aggregating explanation methods to reduce variance - detailed derivation

All currently available explanation methods have weaknesses that are inherent to the approach and include significant noise in the heatmap (Kindermans et al., 2017; Adebayo et al., 2018; Smilkov et al., 2017). A natural way to mitigate this issue and reduce noise is to combine multiple explanation methods. Ensemble methods have been used for a long time to reduce the variance and bias of machine learning models. We apply the same idea to explanation methods and build an ensemble of explanation methods.

We assume a neural network  $F : X \mapsto y$  with  $X \in \mathbb{R}^{m \times m}$  and a set of explanation methods  $\{e_j\}_{j=1}^J$  with  $e_j : X, y, F \mapsto E$  with  $E \in \mathbb{R}^{m \times m}$ . We write  $E_{j,n}$  for the explanation obtained for  $X_n$  with method  $e_j$  and denote the mean aggregate explanation as  $\bar{E}_n$  with  $\bar{E}_n = \frac{1}{J} \sum_{j=1}^J E_{j,n}$ . While we assume the input to be an image  $\in R^{m \times m}$ , this method is generalizable to inputs of other dimensions as well.

We define the error of an explanation method as the mean squared difference between a hypothetical 'true' explanation and an explanation procured by the explanation method, i.e. the MSE. For this definition we assume the existence of the hypothetical 'true' explanation  $\hat{E}_n$  for image  $X_n$ .

For clarity we subsequently omit the notation for the neural network.

We write the error of explanation method  $j$  on image  $X_n$  as  $err_{j,n} = \|E_{j,n} - \hat{E}_n\|^2$  with

$$\text{MSE}(E_j) = \frac{1}{N} \sum_n err_{j,n}$$

and  $\text{MSE}(\bar{E}) = \frac{1}{N} \sum_n \|\bar{E}_n - \hat{E}_n\|^2$  is the MSE of the aggregate. The typical error of an explanation method is represented by the mean

$$\begin{aligned} \overline{\text{MSE}} &= \frac{1}{N} \sum_n \frac{1}{J} \sum_j \|\hat{E}_n - E_{j,n}\|^2 \\ &= \frac{1}{NJ} \sum_{n,j} \|\hat{E}_n - E_{j,n} + \bar{E}_n - \bar{E}_n\|^2 \\ &= \frac{1}{NJ} \sum_{n,j} \|(\hat{E}_n - \bar{E}_n) + (\bar{E}_n - E_{j,n})\|^2 \\ &= \frac{1}{NJ} \sum_{n,j} \|\hat{E}_n - \bar{E}_n\|^2 + \|\bar{E}_n - E_{j,n}\|^2 + \frac{1}{NJ} \sum_{n,j} \left( 2\text{Tr} \left[ (\hat{E}_n - \bar{E}_n)(\bar{E}_n - E_{j,n}) \right] \right) \\ &= \frac{1}{N} \sum_n \|\hat{E}_n - \bar{E}_n\|^2 + \frac{1}{NJ} \sum_{n,j} \|\bar{E}_n - E_{j,n}\|^2 + 2 \frac{1}{N} \sum_n \text{Tr} \left[ (\hat{E}_n - \bar{E}_n) \left( \frac{1}{J} \sum_j (\bar{E}_n - E_{j,n}) \right) \right] \\ &= \frac{1}{N} \sum_n \|\hat{E}_n - \bar{E}_n\|^2 + \frac{1}{NJ} \sum_{n,j} \|\bar{E}_n - E_{j,n}\|^2 + 2 \frac{1}{N} \sum_n \text{Tr} \left[ (\hat{E}_n - \bar{E}_n) \underbrace{\frac{1}{J} \sum_j (\bar{E}_n - E_{j,n})}_{=0} \right] \\ &= \frac{1}{N} \sum_n \|\hat{E}_n - \bar{E}_n\|^2 + \frac{1}{NJ} \sum_{n,j} \|\bar{E}_n - E_{j,n}\|^2, \end{aligned}$$

hence,

$$\overline{\text{MSE}} = \text{MSE}(\bar{E}) + \frac{1}{NJ} \sum_{n,j} \underbrace{\|\bar{E}_n - E_{j,n}\|^2}_{\text{epistemic uncertainty}} \geq \text{MSE}(\bar{E})$$

The error of the aggregate MSE( $\bar{E}$ ) is less than the typical error of the participating methods. The difference - a ‘variance’ term - represents the epistemic uncertainty and only vanishes if all methods produce identical maps.

### A.2. Comparing aggregate of two methods

In appendix A.1 we showed theoretically that the average MSE of two or more explanation methods will always be higher than the error of the averaged of those methods. Empirically, we test this for the IROF score first proposed in with combinations of any two methods for ResNet101 and show the results in fig. 7. For any two methods, the matrix shows the ratio between the aggregate method IROF and the average IROF of the aggregated methods. The aggregate IROF is always lower, confirming our theoretical results.

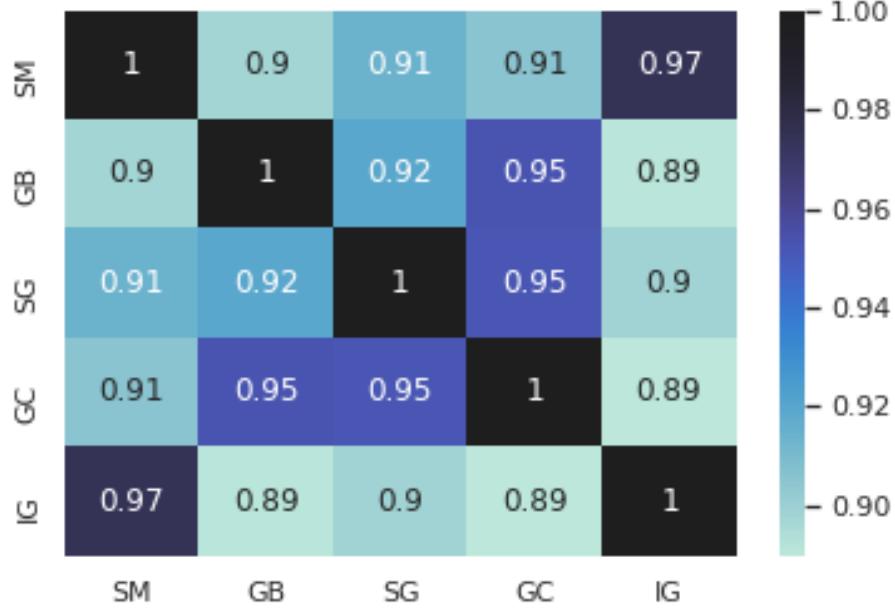


Figure 7. Ratios between the aggregate IROF and the averaged IROF of two methods. Aggregation always improves on the results, as all values outside of the diagonal are below one.

### A.3. Experimental setup

#### A.3.1. GENERAL

We use SLIC for image segmentation due to availability and quick run time (Achanta et al., 2012). Preliminary experiments with Quickshift showed similar results (Vedaldi & Soatto, 2008). SLIC was chosen over Quickshift due to the quicker run time. The number of segments was set to 300 ad hoc. For a detailed description with motivation and evaluation of the evaluation metric used we refer to (Anonymous, 2020)

For AGG-Var, we add a constant to the denominator. We set this constant to 10 times the mean std, a value chosen empirically after trying values in the range of [1, 10, 100] times the mean.

Evaluations were run with a set random seed for reproducibility. SE were reported either for each individual result or if they were non-significant in the caption to avoid cluttering the results.

All experiments were done on a Titan X.

#### A.3.2. MNISTs

The training for both models was equivalent. The architecture was as follows:

(input)-(conv(32,3,3))-(conv(64,3,3))-(maxPool(2,2))-(dropout(0.25))  
-(fully connected(128))-(dropout(0.5))-(output(10))

ReLU was used as a non-linearity for both. All networks were trained with Adadelta and early stopping on the validation set (patience of three epochs) (Zeiler, 2012). The final accuracy for MNIST was 99.21%.: The final accuracy on FashionMNIST was 92.46%.

### A.3.3. IMAGENET

We tested our method on five network architectures that were pre-trained on ImageNet: VGG19, Xception, Inception, ResNet50 and ResNet101 (Deng et al., 2009; Simonyan & Zisserman, 2014; He et al., 2016; Chollet, 2017; Szegedy et al., 2016). We used the pre-trained networks VGG19, Xception and Inception, obtained from the keras library and did not change the networks in any way. (Deng et al., 2009; Szegedy et al., 2016; Chollet, 2017; Simonyan & Zisserman, 2014).

We downloaded the data from the ImageNet Large Scale Visual Recognition Challenge website and used the validation set only. No images were excluded. The images were preprocessed to be within  $[-1, 1]$  unless a custom range was used for training (indicated by the preprocess function of keras).

### A.3.4. DETAILS ABOUT ATTACKING EXPLANATION METHODS

For a range of explanation methods we chose to compare against LRP, Gradient, Guided Backpropagation and Integrated Gradients, a range of well-known and well-established explanation methods (Sundararajan et al., 2017; Bach et al., 2015; Springenberg et al., 2014; Simonyan et al., 2013). Since Integrated Gradients is thirty times more computationally expensive than other methods, we did not include it in the aggregation as it would have slowed down experiments considerably.

Unless otherwise noted, all metrics are computed as the average of a hundred data samples with mean and SE. Informally, we also found that the MSE does not align well with perceived changes in the explanations, likely due to it being susceptible to outliers.

We used a pretrained VGG16 for all experiments attacking explanation methods (Simonyan & Zisserman, 2014).

#### A.4. Alignment between human attribution and explanation methods

We want to quantify whether an explanation method agrees with human judgement on which parts of an image should be important. While human annotation is expensive, there exists a benchmark for human evaluation introduced in (Mohseni & Ragan, 2018). The benchmark includes ninety images of categories in the ImageNet Challenge (ten images were excluded

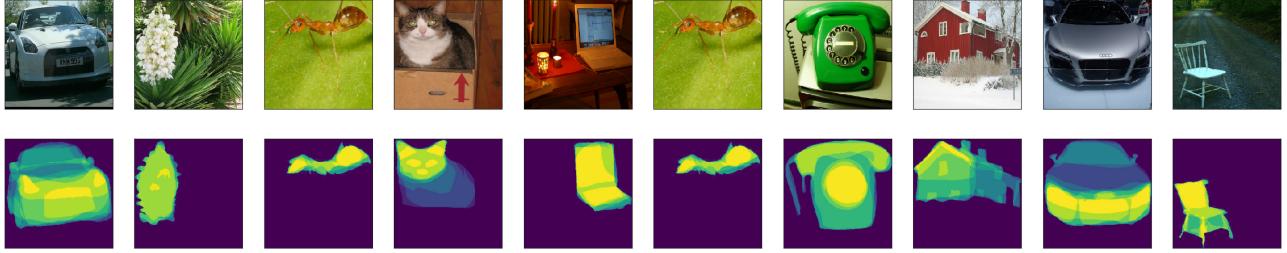


Figure 8. Example images from (Mohseni & Ragan, 2018) with human-annotated overlays.

due to the category not being in the ImageNet challenge) and provides annotations of relevant segments that ten human test objects found important. Example images are shown in fig. 8.

While human evaluation is not a precise measure, we still expect some correlation between neural network and human judgement.

To test the alignment, we calculate the cosine similarity,

$$\text{similarity}(e_j) = \frac{\sum_{n=1}^N A_n E_{j,n}}{\sqrt{\sum_{n=1}^N A_n^2} \sqrt{\sum_{n=1}^N E_{j,n}^2}}$$

between the human annotation and the explanations produced by the respective explanation methods.  $A_n$  is the human annotation of what is important for image  $X_n$

Since the images in this dataset are 224x224 pixel large, we only compute the cosine similarity for the network architectures where pretrained networks with this input size were available.

We see that *AGG-Mean* and *AGG-Var* perform on-par with the best methods (SmoothGrad and GradCAM). While the aggregated methods perform better than the average explanation method, they do not surpass the best method.

When we combine the two best-performing single methods, SmoothGrad and GradCAM, we surpass each individual method. We hypothesize that this is because the epistemic uncertainty is reduced by the aggregate.

Table 3. Cosine similarity between heatmap and human annotated benchmark. All SE below 0.05

METHOD	RESNET101	RESNET50	VGG19
AGG-MEAN	0.63	0.66	0.64
AGG-VAR	0.66	0.68	0.67
GB	0.42	0.49	0.47
GC	0.60	0.62	0.60
IG	0.45	0.45	0.47
MEAN(SG+GC)	<b>0.69</b>	<b>0.70</b>	<b>0.65</b>
SG	0.63	0.65	0.59
SM	0.45	0.45	0.47

### A.5. Details about attacking explanation methods

**Choice of explanation methods** We focused on explanation methods that have previously been shown to be susceptible to adversarial attacks. As such, we did not include GradCAM in the experiments, neither as a comparison or in the aggregation.

Different explanation methods have different computational loads. Notably, SmoothGrad and IntegratedGradients involve the sampling of many explanations for a single pass, increasing computation times by the number of samples () and were not included in the aggregation but as a comparison.

**Choice of hyperparameters** We followed (Dombrowski et al., 2019) for the choice of hyperparameters in learning rate and beta growth. For AGG-Mean we chose a learning rate of  $10^{-3}$  and 1500 iterations for the attack. We heavily use their code, retrieved from the repository of the authors<sup>3</sup> (Dombrowski et al., 2019).

#### A.5.1. MORE EXAMPLES

We provide more examples showing different explanation methods being attacked in figs. 9 to 12. An abridged version of fig. 9 is also shown in the main text.

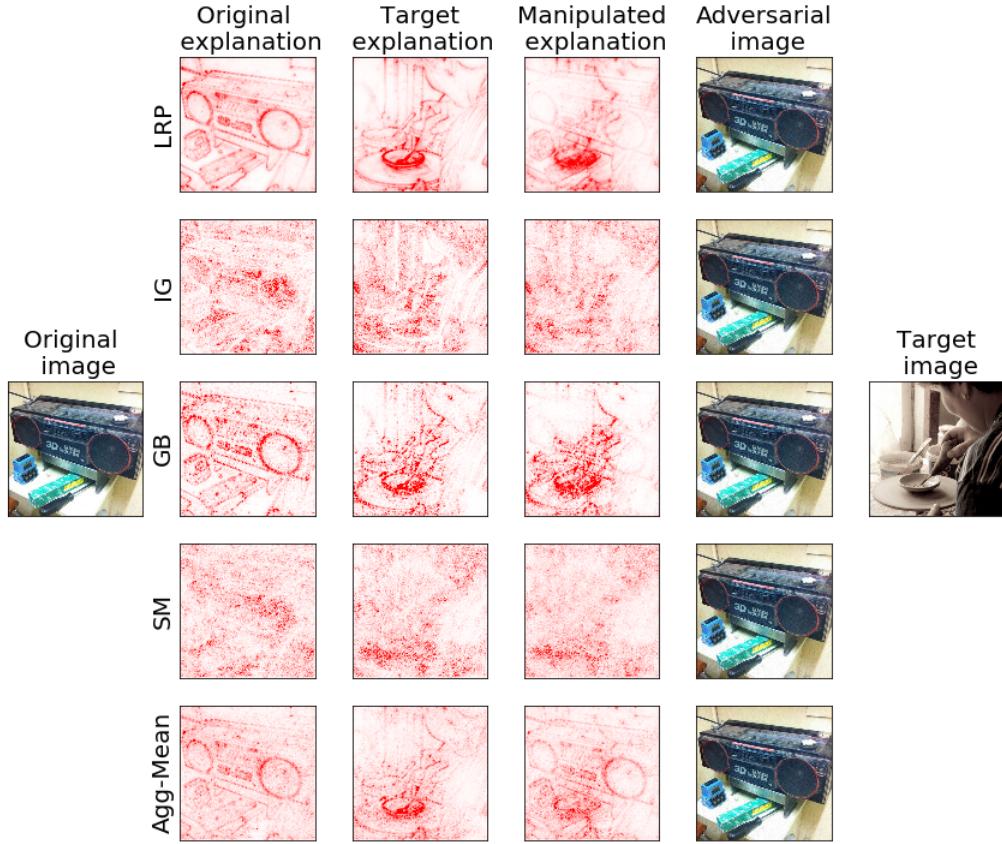
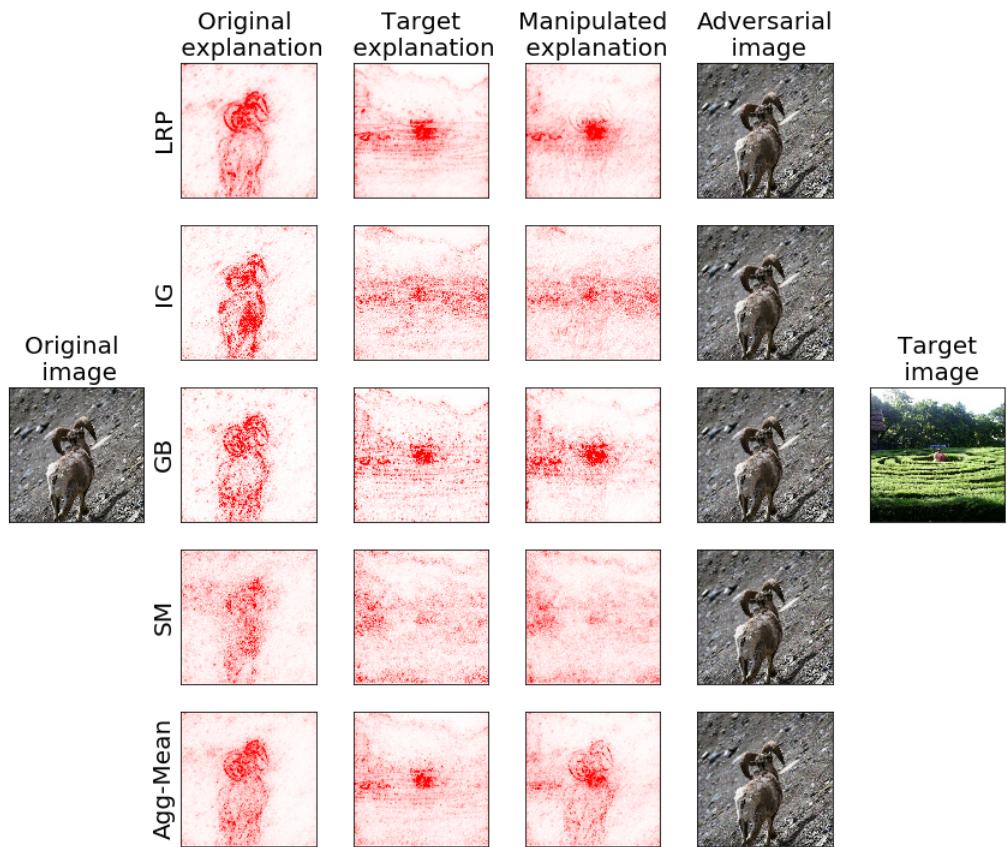
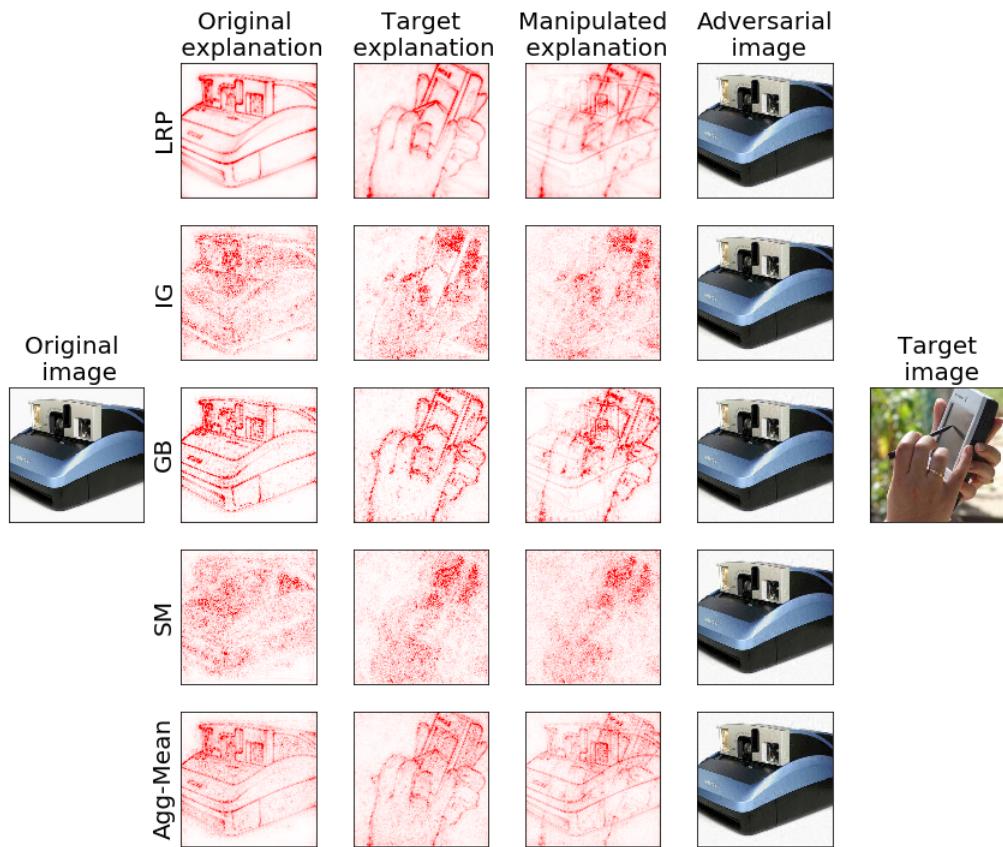


Figure 9. Attack shown in the main text, including the adversarial input images. There are no visual differences for any of the adversarial inputs.

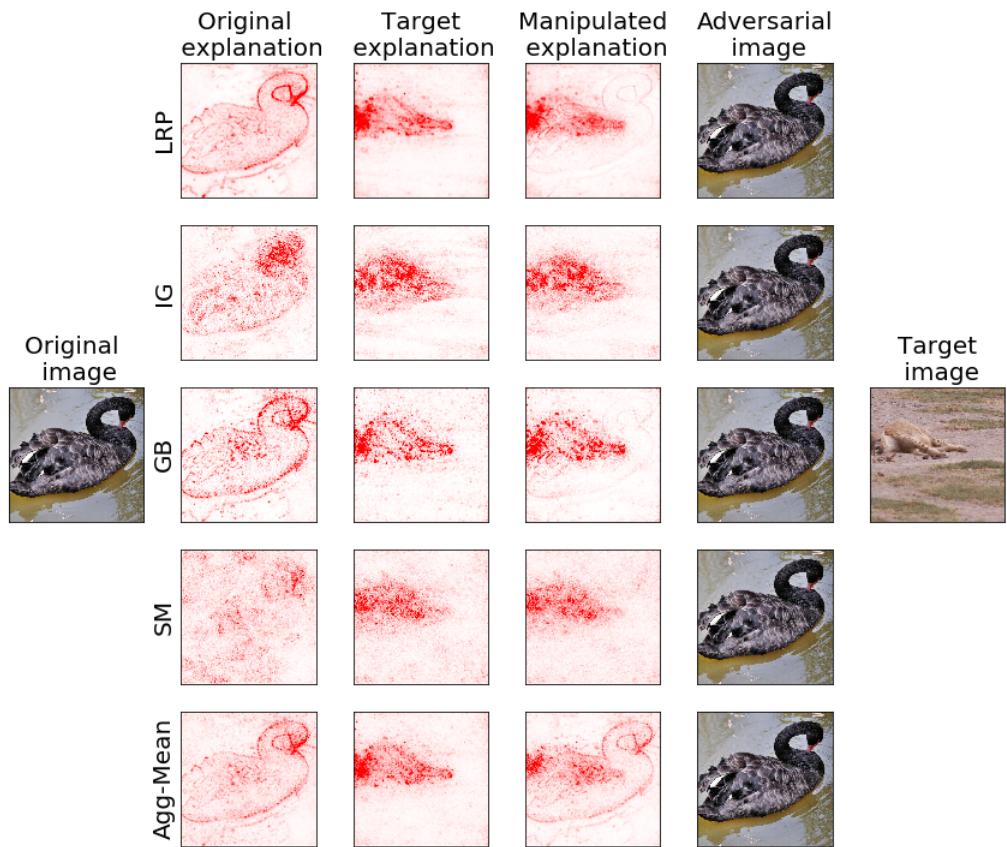
<sup>3</sup>[https://github.com/pankessel/explanations\\_can\\_be\\_manipulated](https://github.com/pankessel/explanations_can_be_manipulated)



*Figure 10.* Appendix example 1. Aggregation is more robust against attack. There are no visual differences for any of the adversarial inputs.



*Figure 11.* Appendix example 2. Aggregation is more robust against attack. There are no visual differences for any of the adversarial inputs.



*Figure 12.* Appendix example 3. Aggregation is more robust against attack. There are no visual differences for any of the adversarial inputs.

### A.5.2. TRANSFERABILITY OF ATTACKS

In the main text we show similarity metrics differences between the method being attacked and not being attacked for Guided Backprop and LRP. Here we provide scatter plots for the rest of the considered methods in figs. 13 to 16:

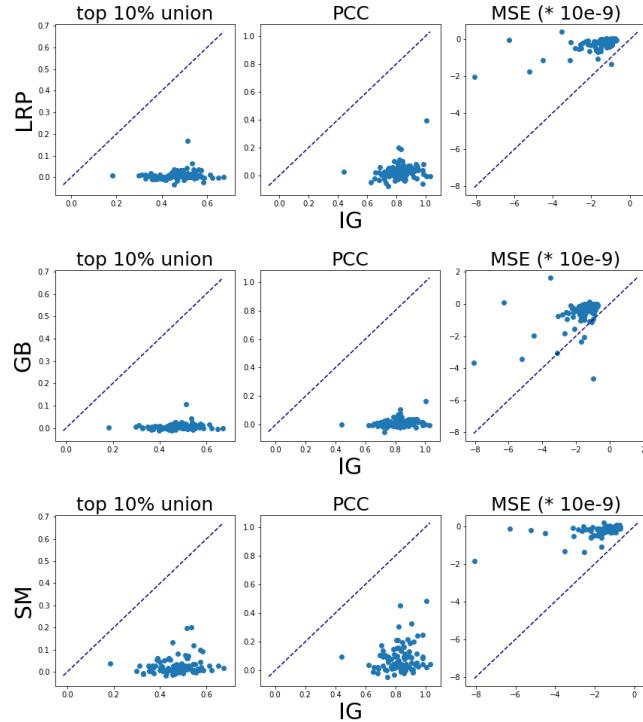


Figure 13. Integrated Gradient as starting method

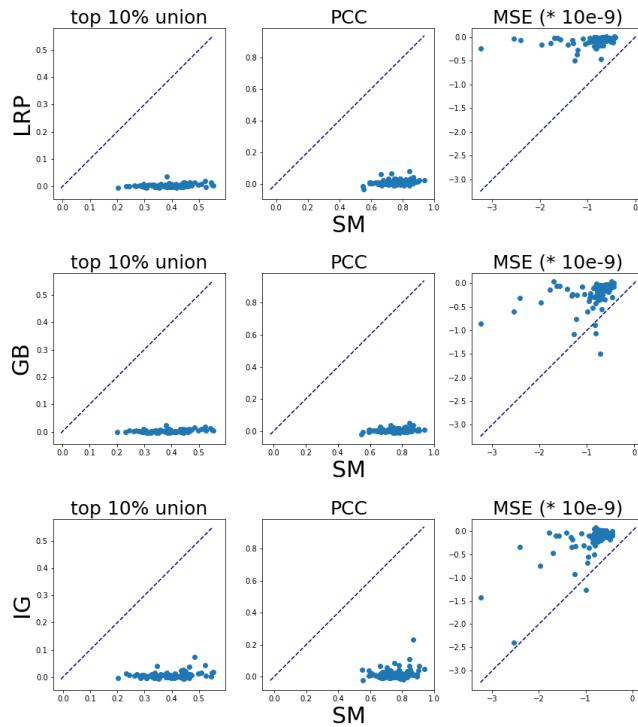


Figure 14. Gradient as starting method

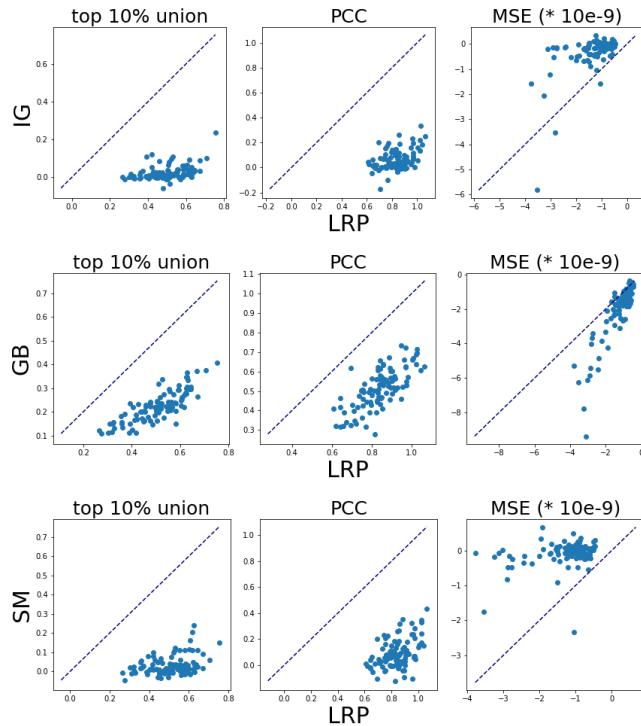


Figure 15. LRP as starting method

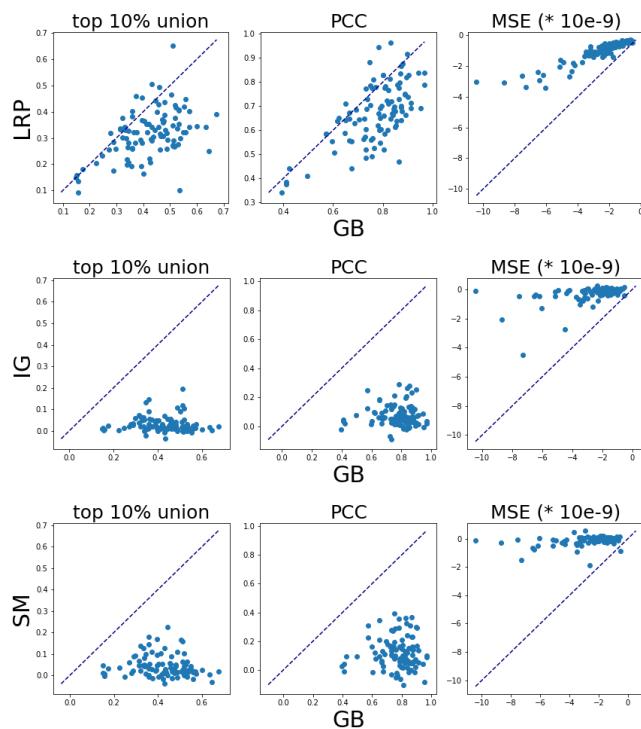


Figure 16. GuidedBackprop as starting method

## A.5.3. SIMILARITY OF THE ATTACKED IMAGES TO THE STARTING IMAGES

We provide the average distance of the adversarial images to the original images in table 4 (calculated in RGB space, average over all pixels). As can be seen in figs. 9 to 11 and 12, there is no visual difference to the input images for any of the attacks. Attacking *AGG-Mean* has the smallest distance to the input image, supporting our hypothesis that aggregating explanation methods removes vulnerabilities to adversarial manipulation.

*Table 4.* Evaluation scores across methods and architectures on a hundred samples, including similarity of the resulting image to the starting image. Deviation is SE.

METHOD	MSE $\Delta$ (*10E-9)	PCC	TOP 10% UNION	MSE ON IMAGES
SM	-0.92 ± 0.00	0.74 ± 0.01	0.40 ± 0.01	0.0027 ± 0.0002
GB	-3.25 ± 0.02	0.77 ± 0.01	0.42 ± 0.01	0.0110 ± 0.0025
LRP	-1.45 ± 0.01	0.81 ± 0.01	0.49 ± 0.01	0.0047 ± 0.0006
IG	-1.76 ± 0.01	0.82 ± 0.01	0.47 ± 0.01	0.0102 ± 0.0022
AGG-MEAN	<b>-0.89 ± 0.01</b>	<b>0.54 ± 0.01</b>	<b>0.24 ± 0.01</b>	0.0013 ± 0.0001

#### A.5.4. OTHER ATTACKS

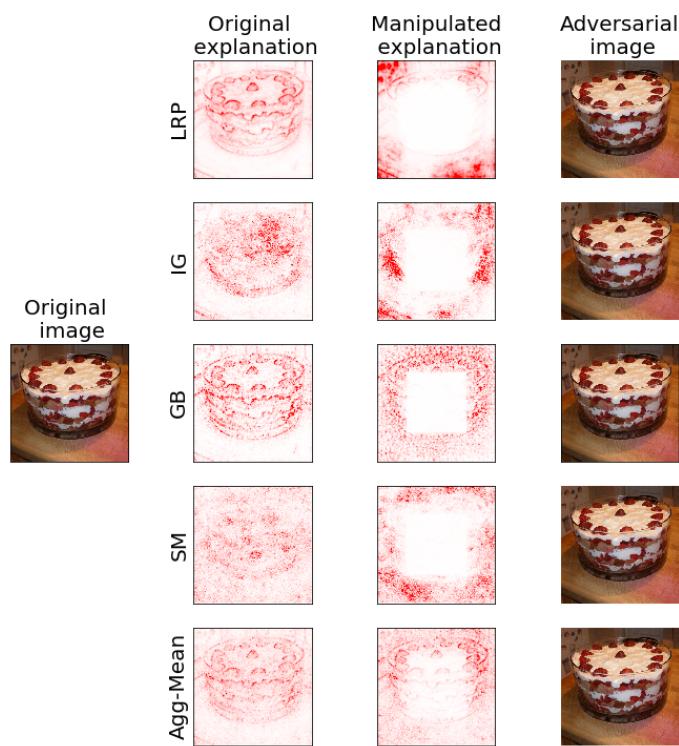
In the main text we mainly concern ourselves with reproducing a pre-specified target explanation, as this is a use case where the motivation of an attacker is apparent. However, as introduced in (Ghorbani et al., 2019) other attack objectives are also conceivable.

We show results when following the objective of making a specified area of the explanation not relevant, i.e. a blank space in the explanation as introduced in (Ghorbani et al., 2019). A square (in size a quarter of the image) centered on the middle should not contain any relevance for the explanation. Size and position of the blank space were chosen ad-hoc, we assume that the center of the image generally contains useful information for the classification. We show quantitative results in table 5, computing how much percentage of the original relevance is preserved and qualitative results in figs. 17 and 18. Since the goal of the attack is to entirely remove the relevance in the area, measuring how much relevance is left effectively measures how robust the explanation method is against the adversarial attack.

While an aggregation is not completely robust to the attack, far more of the original explanation is preserved, supporting the results in the main text.

*Table 5.* Manipulating explanations to show a blank (irrelevant) square. Aggregating explanation methods preserves far more of the original explanation than any single method.

METHOD	START_PERCENTAGE	END_PERCENTAGE	PRESERVED
SM	0.09	1.84E-02	0.20
GB	0.11	2.58E-03	0.02
IG	0.11	2.31E-02	0.21
LRP	0.11	1.71E-03	0.02
AGG-MEAN	0.10	<b>3.09E-02</b>	<b>0.31</b>



*Figure 17.* Attacking explanation methods to make an area irrelevant. AGG-Mean is most robust.

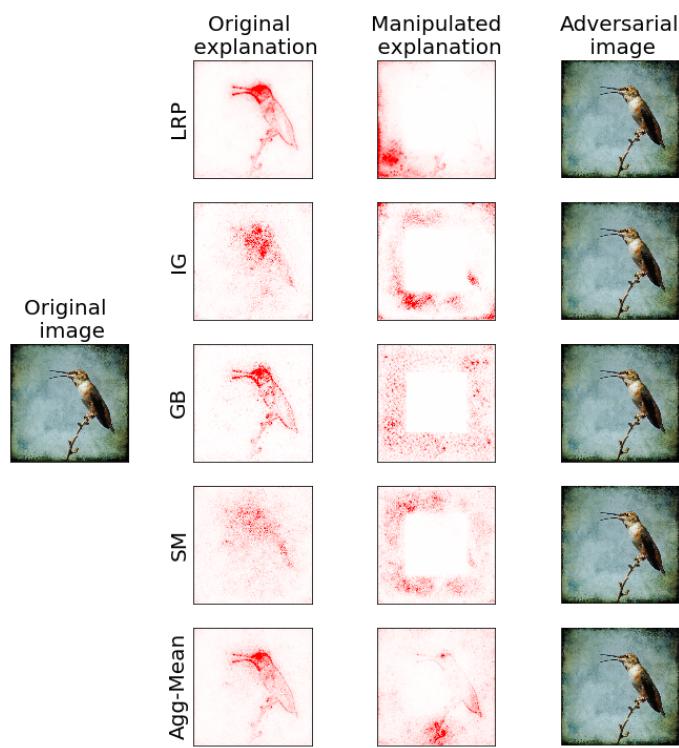


Figure 18. Attacking explanation methods to make an area irrelevant. AGG-Mean is most robust.