Towards Better Visual Explanations for Deep Image Classifiers

Agnieszka Grabska-Barwińska
DeepMind, London, UK
agnigb@google.com

Amal Rannen-Triki
DeepMind, London, UK
arannen@google.com

Omar Rivasplata*
University College London, UK
o.rivasplata@ucl.ac.uk

András György
DeepMind, London, UK
agyorgy@deepmind.com

Abstract

Understanding and explaining the decisions of neural networks is of great importance, for safe deployment as well as for legal reasons. In this paper, we consider visual explanations for deep image classifiers that are both informative and understandable by humans. Motivated by the recent FullGrad method, we find that bringing information from multiple layers is very effective in producing explanations. Based on this observation, we propose a new method, DeepMaps, that combines information from hidden activities. We show that our method outranks alternative explanations with respect to metrics established in the literature, which are based on pixel perturbations. While these evaluations are based on changes in the class scores, we propose to directly consider the change in the network's decisions. Noting that perturbation-based metrics can fail to distinguish random explanations from sensible ones, we propose to measure the quality of a given explanation by comparing it to explanations for randomly selected other images. We demonstrate through experiments that DeepMaps outperforms existing methods according to the resulting evaluation metrics as well.

1 Introduction

Neural networks are well-established tools in a wide range of applications, such as image recognition, natural language understanding, recommendation systems, to name a few. Their deployment, however, is in some cases met with reservation, and comes with its own challenges (e.g. legal or safety-related), given how little guarantees are available about their generalisation properties, or about their robustness to changes in the input data. Therefore, there is an increasing demand for explanations (Goodman & Flaxman, 2017), especially urgent where machine learning algorithms support human decisions, e.g., in medical diagnosis, law, banking, among other domains.

In this paper, we add to the explainable AI effort by focusing on explaining decisions of deep image classifiers. More precisely, we consider trained networks rather than the problem of image classification itself (cf. Rieger & Hansen, 2019; Kindermans et al., 2017), and we study instance-based explanations (Baehrens et al., 2010), i.e. explanations for a given image. Supplying such visual explanations together with network decisions might increase trust of the end user, or, contrarily, help to spot deficiencies in the network's decisions. Thus, our work is relevant for deciding which networks are safe to deploy (cf. Adebayo et al., 2020).

[*] During this work, O. Rivasplata was an intern at DeepMind.

1st Workshop on eXplainable AI approaches for debugging and diagnosis (XAI4Debugging@NeurIPS2021).

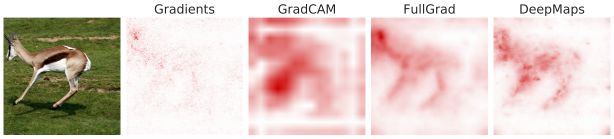

Figure 1: Saliency maps. Example visual explanations were produced by one of the first methods (GRADIENTS), a popular one (GRADCAM), a recent one (FULLGRAD) and ours (DEEPMAPS).

What we call a *visual explanation* takes the form of an image that matches the shape (array of pixels) of the given input image. Such explanations can be easily visualised and even overlaid over the input image, boosting interpretability. Specifically, we focus on *saliency maps*, which are visual explanations that highlight pixels that are most relevant for the network's decision, such as shown in Fig. 1. As one can see, different saliency maps range from scattered points to big blobs through shapes which highlight the most relevant content in the image.

The first attempts to produce visual explanations were based on input gradients (see, e.g. Fig. 1, Gradients), showing the local sensitivity of the class score, which is rather hard to interpret (Baehrens et al., 2010; Simonyan et al., 2014). To address this issue, Selvaraju et al. (2016) proposed to analyse the information obtained from the last convolutional layer rather than the input. Their method yields explanations of lower resolution, which, while desirable, also tends to remove the fine details of the object shape (Fig. 1, GradCAM). We advocate for explanations that are of sufficiently low resolution to be localised around the object of interest, and at the same time sufficiently high resolution to keep the distinctive details of its shape. Recent work by Srinivas & Fleuret (2019) goes in this direction, and argues that methods for extracting visual explanations should take bias parameters into account (see also Kindermans et al., 2017). Applying their method to the original VGG-16 network (Simonyan & Zisserman, 2014) results in a vast improvement over the input-based saliency maps (Fig. 1, FullGrad). However, an analysis of the behaviour of VGG-16 shows that the bias parameters contribute very little to the class scores and have negligible effect in the final decisions. Thus, the success of the FULLGRAD method of Srinivas & Fleuret (2019) must be for some other reason. We hypothesise that it is rather due to aggregating explanations over different depths, and propose a new method, DEEPMAPS, that produces saliency maps based on hidden activities instead of biases.

Apart from producing better explanations, on which many studies have focused, developing evaluations for network explanations remains challenging. We often rely on visual comparison to the ground truth images as a way to rank methods. While such an evaluation is easy to understand, it is biased towards human perception and results in the introduction of many heuristics for smoothing or sharpening the saliency maps. It is therefore hard to tell which part of the resulting explanations is related to the network behaviour and which part is related to the method heuristics, and so developing quantitative measures that are unambiguous as well as principled is of high interest.

Popular quantitative measures are built on image perturbation as a means for evaluating saliency maps. In the most common pixel-removal approach, image regions are masked while the change in the class scores is monitored (Samek et al., 2016). The masks are chosen according to the given saliency map, either selecting the most or least important pixels. However, existing approaches have multiple drawbacks: they care about class scores instead of the actual classification performance, and they are not always able to distinguish sensible explanations from obscure ones; for example, high frequency random perturbations are likely to confuse state-of-the-art neural networks (Samek et al., 2016; Fong & Vedaldi, 2017; Dabkowski & Gal, 2017), and hence can achieve high scores as explanations when considering the removal of the most important pixels.

To overcome the above difficulties, we propose new evaluation approaches. First, rather than monitoring the score change while removing pixels, we monitor the change in decisions. In addition, we propose a new evaluation method that is constructed to distinguish meaningful explanations from random ones. This new evaluation method measures the quality of a given explanation by comparing it to a set of other explanations obtained by the same explanation method for random images from the dataset of interest.

We analyse the performance of DEEPMAPS through extensive experiments and compare it to FULLGRAD and other existing methods (see also Appendix E). Arguably, it produces explanations that are visually more convincing, as they particularly highlight the fine details of the object shape that are important to the predicted class (see, e.g. Fig. 1, DeepMaps). We also demonstrate that in VGG-16, DEEPMAPS outperforms previous explanations with respect to evaluation metrics commonly used in the literature, and the new ones proposed in this paper. We also analyse DEEPMAPS for ResNet-50 (He et al., 2016), showing a competitive performance to previous methods.

To summarize, our contributions come in two angles: (i) We challenge the use of biases for visual explanations and propose a new method that aggregates information from hidden activations instead. (ii) We rethink the evaluation of visual explanations and propose new metrics, and a novel evaluation paradigm, which have intuitively desirable properties. The rest of the paper is organised as follows. Previous work is presented in §2. The analysis leading to the development of DEEPMAPS and the description of the method are presented in §3. Existing and new quality metrics are described in §4, together with an in-depth analysis of the performance of the different methods according to the different metrics.

2 Notation and background

We consider feed-forward (convolutional) ReLU networks for image classification, as sketched in Fig. 2 (top). A ReLU network composed of L layers takes an input image $\mathbf{x} \in \mathbb{R}^{m \times n \times \Phi}$ of width m, height n, and composed of Φ colour channels, and progressively maps this image into subsequent hidden layers, which gives the "images" $\mathbf{x}^l$ for $l \in \{1, \ldots, L-1\}$ defined by $\mathbf{x}^l = \left[W^l \mathbf{x}^{l-1} + b^l\right]_+$, where $\mathbf{x}^0 = \mathbf{x}$ is the original input, $\mathbf{x}^l \in \mathbb{R}^{m_l \times n_l \times \Phi_l}$ for suitable dimensions, W^l is a weight matrix of appropriate size, b^l is a tensor with the same size as $\mathbf{x}^l$ (although, in practice, bias values are often shared across space), and $[\cdot]_+$ is the positive part function (applied entrywise to a tensor).[2] The network computes class scores f_c for all possible classes $c \in \{1, \ldots, C\}$ by $f_c = w_c^L \mathbf{x}^{L-1} + b_c^L$. The predicted label for $\mathbf{x}$ is given by the class corresponding to the highest score. Throughout the paper, each entry of a vector corresponding to a neuron in a layer (such as $\mathbf{x}^l$) is indexed by its spatial position (i, j) and its channel ϕ.

2.1 From gradient-based to attribution-based explanations

The first attempts to produce explanations were based on the study of the gradients, coinciding with the weights for the case of linear models. Gradient visualisations provide insight into how the classifier works locally, by quantifying how each class weighs every pixel in a small neighborhood of the input image. This is especially true for ReLU networks, as the class scores are piecewise linear functions of the input $\mathbf{x}$.

The earliest methods for producing visual explanations based on this concept of gradient back-propagation are due to (Baehrens et al., 2010; Simonyan et al., 2014). Such methods simply compute the gradients $\frac{\partial f_c(\mathbf{x})}{\partial x_{ij\phi}}$ (note that they exist in ReLU networks for almost all images, apart from a set of Lebesgue measure zero), and produce a saliency map $\mathbf{s}$ of the same size as the input image $\mathbf{x}$, by taking, for example, $\mathbf{s}_{ij}(\mathbf{x}) = \sum_\phi \left| \frac{\partial f_c(\mathbf{x})}{\partial x_{ij\phi}} \right|$. We refer to this method as GRADIENTS. In a similar vein, a large family of attribution methods were developed, devising rules for decomposing the class score and back-propagating the decomposition to the input space (Bach et al., 2015; Selvaraju et al., 2016; Sundararajan et al., 2017; Smilkov et al., 2017; Shrikumar et al., 2017; Lundberg & Lee, 2017). Of those, particularly relevant to this work are gradient×input attributions (Srinivas & Fleuret, 2019), defined as

$$a_{cij}(\mathbf{x}) = \sum_{\phi=1}^{\Phi} x_{ij\phi} \frac{\partial f_c(\mathbf{x})}{\partial x_{ij\phi}} . \tag{1}$$

While these explanations are easy to compute, are intuitively related to the network behaviour, and result in saliency maps showing signal around the objects of interest; they are typically scattered and noisy-looking (similarly to the panel Gradients in Fig. 1). In view of this, from a human-centric perspective, such explanations are hard to interpret.

[2] With a slight abuse of notation, tensors are treated in the equations as their one-dimensional flattened versions.

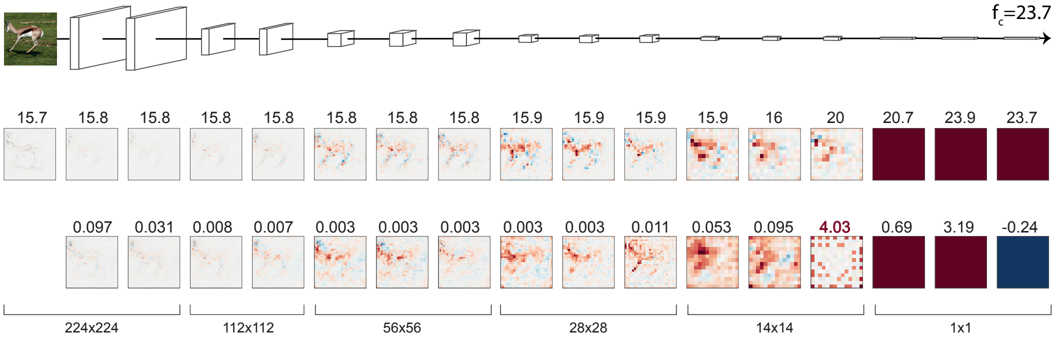

Figure 2: Example attributions in VGG-16. An image of a gazelle is fed into the VGG-16 classifier, which outputs the score $f_c = 23.7$ in favour of that class. Per layer activity attributions ($\mathbf{a}^{h,l}$, middle) and bias attributions ($\mathbf{a}^{b,l}$, bottom) are shown in color (red positive, blue negative), with their gross sum ($A_c^{h,l}$, $A_c^{b,l}$) displayed over each panel. Biases from nearly all convolutional layers contribute less than a percentage of the overall score. The last convolutional layer contributes most value (4.03), but in that layer, attributions lose focus from the gazelle. If the biases were all zero, hidden-layer attributions would sum to the same value in every single layer.

One of the first and most popular methods to produce low-resolution explanations is GRADCAM (Selvaraju et al., 2016), which uses the gradients with respect to the activations in the last convolutional layer. While the saliency maps produced by this method (e.g. Fig. 1, GradCAM) are more centered around the object of interest, they tend to render a low resolution signal, losing the shape of the object.

The concepts humans tend to look for lie in the spectrum of explanations in-between these two extremes. More recently, Srinivas & Fleuret (2019) proposed the FULLGRAD method that incorporates information from the biases. Their method is derived by looking at complete decompositions of the class scores, i.e. attributions that sum up to the score. For a network with zero-biases, due to the piecewise linearity of the output, this is obtained from gradient×input attributions:

$$f_c(\mathbf{x}) = \sum_{i,j,\phi} x_{ij\phi} \frac{\partial f_c(\mathbf{x})}{\partial x_{ij\phi}} = \sum_{i,j} a_{cij}(\mathbf{x}). \tag{2}$$

In general, a complete decomposition of a class score requires taking bias parameters into account. Defining bias attributions as

$$a_{cij\phi}^{b,l} = b_{ij\phi}^{l} \frac{\partial f_c(\mathbf{x})}{\partial b_{ij\phi}^{l}}, \tag{3}$$

Srinivas & Fleuret (2019) noted that $f_c(\mathbf{x}) = \sum_{i,j} a_{cij}(\mathbf{x}) + \sum_{l,\phi} \mathbf{a}_{c\phi}^{b,l}$. Mapping the biases back to a single image of the original size ($m \times n$), they proposed the saliency map

$$\mathbf{s}_{ij} = \Psi(\mathbf{a}_c)_{ij} + \sum_{l=1}^{L} \sum_{\phi \in \Phi^l} \Psi(\mathbf{a}_{c\phi}^{b,l})_{ij}\,, \tag{4}$$

where $\mathbf{a}_c = (a_{cij})_{i,j}$ is the matrix of input attributions (Eq. (1)). Srinivas & Fleuret (2019) recommended using

$$\Psi(\mathbf{a}) = \mathrm{bilinearUpscale}(\mathrm{normalize}(|\mathbf{a}|)) \tag{5}$$

where $\mathrm{bilinearUpscale}$ is a typical image resizing operation mapping its input image to size $m \times n$, and $|.|$ and $\mathrm{normalize}$ are applied entrywise with $\mathrm{normalize}(a_{ij}) = (a_{ij} - \min(\mathbf{a}))/(\max(\mathbf{a}) - \min(\mathbf{a}))$ for a matrix $\mathbf{a}$.

The resulting explanation is shown in Fig. 1 (FullGrad). There is a clear improvement over saliency maps based on input attributions, which look just like the Gradients, Fig. 1. In the next section, we look into the sources of this improvement, and based on our findings we propose a new explanation method.

3 DeepMaps: rethinking attributions

We illustrate the decomposition of the class score for an example image of a gazelle in Fig. 2 for a VGG-16 network. Note that our network was trained without layer normalisations, so there are no other parameters we need to take into account beyond the biases b^l. For each layer l, we sum up bias attributions over all channels, $a^{b,l}_{cij} = \sum_\phi a^{b,l}_{cij\phi}$. We show the bias attributions $a^{b,l}_{cij}$ in the bottom row of Fig. 2, with their sum $A^{b,l}_c = \sum_{i,j} a^{b,l}_{cij}$ shown above the corresponding images. From left to right, attributions decrease in size, reflecting the size of the corresponding convolutional layers (sketched on top), but are shown upscaled to the original size, similarly to the operation in Ψ above.

In the middle row of Fig. 2, we plot input attributions (left), as well as attributions of hidden activations in every layer, defined as

$$a^{h,l}_{cij} = \sum_{\phi \in \Phi^l} x^l_{ij\phi} \frac{\partial f_c}{\partial x^l_{ij\phi}} = \sum_{\phi \in \Phi^l} a^{h,l}_{cij\phi} \tag{6}$$

with $a^{h,0}_{cij} = a_{cij}$. Similarly to the biases, above the attributions of each layer we note their sum

$$A^{h,l}_c(\mathbf{x}) = \sum_{i,j,\phi} x^l_{ij\phi} \frac{\partial f_c(\mathbf{x})}{\partial x^l_{ij\phi}} = \sum_{i,j} a^l_{cij}(\mathbf{x}). \tag{7}$$

Similarly to (2), in a bias-free network we have $f_c(\mathbf{x}) = A^{h,l}_c(\mathbf{x})$, that is, the class scores can be decomposed using the hidden-layer attributions. For such a network, the numbers $A^{h,l}_c$ over each $\mathbf{a}^{h,l}$ panel would be identical and equal to the network output (top right).

From Fig. 2, we can make two important observations, which are generally true for other images, as well: (i) The noisy-looking input attribution has a significantly higher value than the bias attributions, and the sum of attributions for each hidden layer is approximately the same, showing that the biases over these layers have little effect. The most important bias attribution comes from the last convolutional layer and does not make sense from a visual perspective. (ii) All the hidden-layer attributions focus on or around the gazelle, while the bias attributions seem less focused in the last layers. In general, the attributions from the hidden layers seem to be of superior quality.

To summarise, for the studied VGG-16, the information conveyed by the biases seems to have a low impact on the class scores and the final decision. Despite this little impact on the score, FULLGRAD explanations seem visually convincing (see Fig. 1). Therefore, it is natural to ask: *What is the key operation that makes the strength of* FULLGRAD*?* Our hypothesis is that, rather than relying on biases, the crucial contribution of FULLGRAD is the aggregation of information from all the hidden layers of the network. Hidden activations in deeper layers depend on correlations in larger input patches and hence shed some light on the non-linear structure of the network. Furthermore, attributions derived from the hidden activations seem to be visually more convincing than those derived from the biases (as in Fig. 2). Based on these observations, we propose DEEPMAPS, a new method to generate explanations.

DeepMaps: combining attributions from multiple layers: Inspired by the superior quality of $\mathbf{a}^{h,l}$ over $\mathbf{a}^{b,l}$ attributions and the advantages of taking explanations from multiple layers (i.e. of multiple resolution), DEEPMAPS aggregates hidden-layer attributions as

$$\mathbf{s}_{ij} = \sum_{l=l_0}^{L} \Psi\left(\mathbf{a}^{h,l}_c\right)_{ij} \tag{8}$$

where the transformation Ψ is defined in Eq. (5).

Transforming attributions into saliency maps is mostly governed by heuristics. Other than the transformation Ψ, there are also different choices for the level at which it is applied. Note that we apply $\Psi(\cdot)$ to $\mathbf{a}^{h,l}$, rather than to every feature map as in FULLGRAD (cf. Eq. (4)). Thus, our approach emphasises attributions per layer, rather than per feature. Commonly, the number of features grows for spatially smaller layers (Fig. 2 top), which means FULLGRAD effectively overemphasises low-resolution visual explanations. Note that in Fig. 2, hidden-layer attributions seem to have a similar scale ($A^{h,l}_c(\mathbf{x})$) to the input attribution. DEEPMAPS then corresponds, for this example, to giving explanations with equal weight on every resolution, which in turn leads to an explanation dominated by the important resolutions. In Appendix D, we compare single-layer explanations generated by these two approaches.

4 Evaluating explanations

In this section, we analyse the explanations produced by DEEPMAPS and compare them to previous methods. We first extensively analyse the original VGG-16 architecture Simonyan & Zisserman (2014), for which we observed a limited contribution of the biases. We then extend our analysis to ResNet-50, a network in which biases have a more important role (Appendix A). In the sequel, all the visualizations and quantitative evaluations use images from the validation set of ImageNet. All the metrics are evaluated over 5000 images taken from classes 0..99 (50 images per class). Whenever we plot confidence intervals, they refer to one standard deviation of the class-wide mean ($n = 50$).

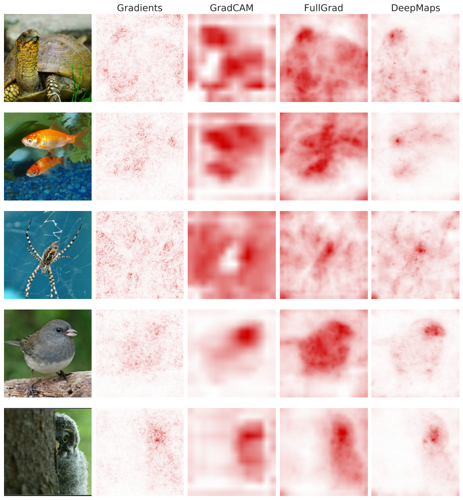

Figure 3: Saliency maps for VGG-16.

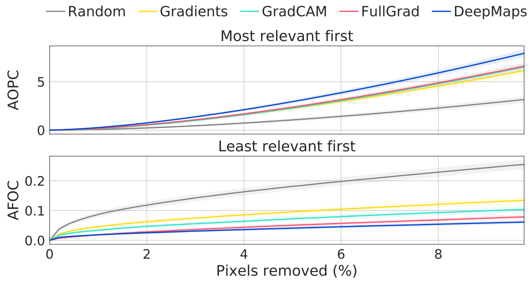

Figure 4: Existing metrics. Average AOPC and AFOC values as a function of the amount of pixels removed (shaded areas show one standard deviation of the mean over classes).

We start with visually inspecting some outputs of DEEPMAPS as well as some of the previous explanation methods for VGG-16. Fig. 3 shows saliency maps obtained by computing the input gradients, GRADCAM (Selvaraju et al., 2016), FULLGRAD (Srinivas & Fleuret, 2019) and DEEPMAPS.[3] Those panels confirm the tendency of the gradients to yield very scattered explanations and of GRADCAM to produce blobby saliency maps. FULLGRAD and DEEPMAPS are both a compromise between the noisy local explanations and the rough delineation of the object of interest. They both highlight accurately the edges and the fine details of the shapes. However, if we look, for example, at the bird example in the fourth row, it appears that FULLGRAD shows an almost uniform signal all over the bird, while DEEPMAPS emphasizes more the beak and the eyes and at a lower level the rest of the animal. Considering the number of birds categories in ImageNet (59), it is expected that the network focuses on finer details than the body to distinguish between the different species, while the body is important to distinguish birds from other super-categories. In addition, in the second row, FULLGRAD seems to pick up some of the background with the fishes, while DEEPMAPS focuses more precisely on the latter, with again a higher emphasis on the eyes. In general, DEEPMAPS seems to produce saliency maps that are easy to analyse and that highlight the distinctive details of the object of interest. These observations, however, are subjective and need to be supported by quantitative evaluations, which we provide below.

4.1 Score-based pixel removal metrics

In the literature, explanations are usually evaluated using image perturbations, that is, measuring how sensitive the network is to modifying the input image according to the ranking provided by a given saliency map (Samek et al., 2016; Srinivas & Fleuret, 2019; Tomsett et al., 2020). To introduce a mild perturbation (and avoid introducing adversarial effects), selected pixels are usually masked out with their values (over all input channels) replaced with an average value computed over the dataset, in effect "removing" the pixel from the input image. Pixel-removal strategies can be divided into two families (Samek et al., 2016): (i) strategies where the most salient pixels are removed,

[3] While arguably there are more saliency methods in the literature we could compare to, such as Layerwise Relevance Propagation (Bach et al., 2015) and its numerous variations, Gur et al. (2021) demonstrated that FULLGRAD is one of the best methods (see their Fig. 3), thus our comparisons are meaningful.

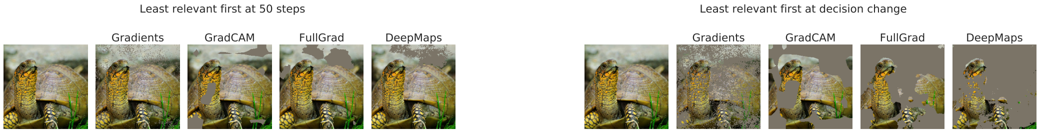
Least relevant first at 50 steps
Gradients
GradCAM
FullGrad
DeepMaps
Least relevant first at decision change
Gradients
GradCAM
FullGrad
DeepMaps

Figure 5: Comparing the standard LeRF approach (removing up to 10% of the image) with our new approach of perturbing until the classification is changed.

designated as *Most Relevant First* (MoRF); (ii) strategies where the least salient pixels are removed. designated as *Least Relevant First* (LeRF).

The effect of pixel-removal is most often studied through the change of the class scores. To this end, we first consider two such metrics for evaluation purposes. The area over the perturbation curve (AOPC), defined by Samek et al. (2016), evaluates the average change in the score of the predicted class as the most relevant pixels are removed progressively (Appendix B). If an explanation succeeds at detecting important regions of the input, the decrease in the class score is expected to be high. AOPC is then expected to be higher for better explanations.

The LeRF absolute fractional output change ($\mathrm{AFOC}_{\mathrm{LeRF}}$) is defined as the change in the score of the originally predicted class relative to the original score when the pixels are removed in a LeRF fashion. This metric is used by Srinivas & Fleuret (2019), who advocate for the LeRF approach as it lessens the effect of the artifacts introduced by pixel removal, which are particularly troublesome when removing random pixels from the salient parts of the image. The decrease in the class score is expected to be lower for the more successful methods. Therefore, for $\mathrm{AFOC}_{\mathrm{LeRF}}$, lower values are better. Note that other variants of these metrics are possible, for example, a LeRF version of AOPC or a MoRF variant of AFOC.

In Fig. 4, we compare DEEPMAPS to random pixel removal, GRADIENTS, GRADCAM and FULLGRAD with respect to average AOPC (top) and $\mathrm{AFOC}_{\mathrm{LeRF}}$ (bottom). The results demonstrate that our method outperforms competitors in detecting both the most important pixels on which the network is relying for its prediction (AOPC), and the least important pixels that can be ignored with minimal effect on the predicted class score ($\mathrm{AFOC}_{\mathrm{LeRF}}$).

4.2 Rethinking the evaluation of saliency maps

While the previous metrics focus on the class scores, they ignore the main purpose of our networks, i.e. the classification itself. In this section, we introduce two new pixel removal evaluations that are based on the decision shift, rather than the change in the class scores.

One way to measure the importance of the selected pixels is to determine what proportion of the pixels can be removed before the predicted class changes.[4] When removing the most relevant pixels first, an explanation is deemed better if the number of pixels n_{MoRF} to be removed is less, while in case of removing the least relevant pixels first, we want the number of pixels n_{LeRF} which can be removed without changing the decision high. A smaller n_{MoRF} value means a more concise explanation. However, as the MoRF perturbation procedure targets the most sensitive region of the input, it is expected to be noisy (albeit informative). This can be complemented by looking at the n_{LeRF} value, which shows how good an explanation is in selecting the irrelevant parts of an image.

In Fig. 5, we show images when a fixed number of pixels are removed, as introduced for LeRF procedures by Samek et al. (2016) (Fig. 5, top row) and the one we propose (bottom row). This figure shows that continuing with the perturbations until the prediction changes, results in more sensible masks. They cover a much larger part of the input, which are not essential for the network's classification.

While the decision-based pixel-removal metrics provide more direct measures than the score-based ones, they are not robust to the effects of unnatural image perturbation, which are known to affect state-of-the-art image classifiers. We therefore propose a new evaluation procedure that directly addresses this issue. The idea is to compare the quality of an explanation to the average quality of

[4] The proportion of pixels here is an arbitrary choice. Any other quantity measuring the amount of change, e.g. change in total energy, can be applied similarly.

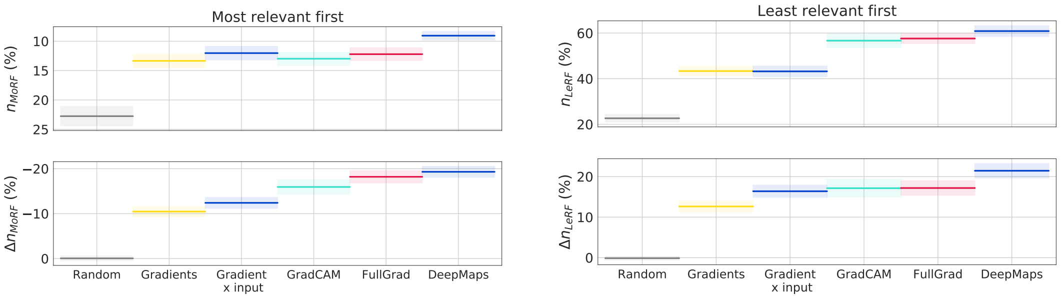

Figure 6: Comparison of saliency methods in VGG-16 using decision-based pixel-removal metrics. The data were collected using images from the ImageNet validation set, for classes 0..99, 50 images each. The coloured boxes depict one standard deviation of the mean across classes ($n = 50$).

explanations obtained with the same method for other images. When the explanations are specific, we expect that the explanation ($\mathbf{s}$) of the underlying image ($\mathbf{x}$) will have a much higher quality than the explanation coming from another image. For random or generic explanations (e.g. selecting the middle region of an image), the quality of explanation coming from any image should be similar. Therefore, for any pixel removal metric σ (where $\sigma(\mathbf{x}, \mathbf{e})$ measures the quality of explanation $\mathbf{e}$ on image $\mathbf{x}$), independently of its category (LeRF or MoRF), we define

$$\Delta\sigma(\mathbf{x}, \mathbf{s}) = \sigma(\mathbf{x}, \mathbf{s}(\mathbf{x})) - \mathbb{E}[\sigma(\mathbf{x}, \mathbf{s}(\boldsymbol{\xi}))] \tag{9}$$

where $\mathbf{s}(\cdot)$ is some saliency method and the expectation is taken over images $\boldsymbol{\xi}$ selected uniformly at random. In our experiments, the expectation is replaced by $\frac{1}{K}\sum_{k=1}^{K} \sigma(\mathbf{x}, \mathbf{s}(\boldsymbol{\xi}_k))$ for $K = 5$ images $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)$ selected uniformly at random from our dataset for every image considered.

Note that all pixel-removal metrics are naturally sensitive to the size of the region of interest, which varies widely between images. Therefore, averaging yields conclusions biased towards smaller or larger objects (depending on the evaluation method). This leads to considering *pairwise tests*, i.e. to compare the effectiveness of the methods per image. For example, one can look at the *win*-statistics (Chattopadhyay et al., 2017), i.e. the proportion of images when one saliency method outperforms another. Naturally, the disadvantage of such a comparison is that it does not take into account how much better the different explanations are for individual images.

4.3 Decision-based evaluation

In Fig. 6, we compare random pixel removal, Gradients, gradient×input attributions (Eq. (1)), gradCam, FullGrad, and DeepMaps with respect to decision-based pixel-removal metrics (note that the direction of the y-axes is chosen so that higher means better). The top panels (in the MoRF and the LeRF section, resp.) show the average impact of pixel-removal procedures. When removing pixels randomly, about 20% of the image needs to be removed for the decision to change. This number drops to about 10% for DeepMaps when removing the most relevant pixels first. Interestingly, n_{MoRF} for gradCam and FullGrad are similar to those reported for Gradients and input attributions, and are somewhat worse than for DeepMaps (about 12-14%). Considering Δn_{MoRF}, the comparison to alternative explanations, changes the picture (Fig. 6, bottom left). First note that the random method gets score zero. This is because all random saliency maps are (on average) equally good in changing network decisions. Gradients, on the other hand, yield explanations that are clearly image-specific: On average, 10% image pixels less need to be removed to change the network decision. For DeepMaps, the difference in the number of pixels removed by alternative explanations ($\boldsymbol{\xi}_k$) and by the true one is almost 20%. The metric also shows the advantages of FullGrad and gradCam versus the simpler input-gradient-based methods.

A complementary picture is revealed when targeting the least relevant pixels (Fig. 6, right). According to n_{LeRF}, there is a clear benefit from using smooth explanations. Similarly to the findings of Srinivas & Fleuret (2019), FullGrad seems even more accurate than gradCam in delineating the irrelevant regions. With DeepMaps, a whooping 60% of the image can be removed (on average) without changing the network's decision. In this case, the Δn_{LeRF} metric finds DeepMaps, FullGrad, and gradCam less advantageous compared to the input-based measures than n_{LeRF}. This is because if one can remove over half of the image without changing the network's decision,

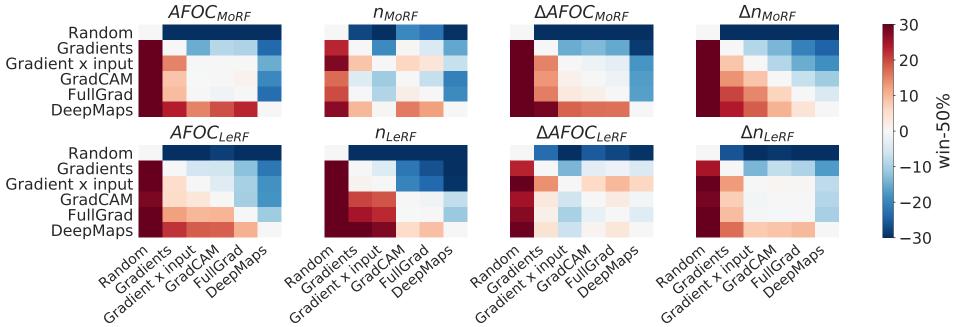

Figure 7: Pairwise tests for different explanation methods applied to VGG-16. The heat maps show the frequency (-50%) of the images for which the method displayed on the left has a better performance than the method displayed on the bottom as measured by the metric displayed on the top of the panel.

it is more likely that such a vast region will overlap with an explanation (involving a similarly large region) for some other image. This metric is then less informative than its MoRF counterpart.

Finally, in Fig. 7, we provide a statistical summary for the results shown in Fig. 6 as well as AFOC. The results are presented in the form of pairwise comparisons, and also include the Δ-metric (which is applicable to any base metrics). Each panel of Fig. 7 shows the frequencies (after subtracting 50%) at which the methods displayed on the left of the heat map yielded values better than the methods displayed at the bottom. The name of the metric used for the comparison is indicated on the top of each heat map. All metrics, except for $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$, rank DeepMaps the best among the tested methods. We noted our reservation towards Δn_{LeRF} already. For $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$, the comparison with alternative explanations makes even less sense, as the perturbation removes a mere 10% of the irrelevant pixels. In lots of images, there could be a lot of such regions that are equally valid. This metric can therefore be misleading.

In Appendix C we provide further analysis about the behavior of the evaluation metrics, demonstrating the statistical significance of our findings above, as well as looking into the consistency of the induced rankings of the different saliency methods.

4.4 Explaining ResNet

So far, we extensively analysed the VGG-16 architecture, showing interesting properties and a clear advantage for DeepMaps. However, DeepMaps yields interesting explanations for other architectures as well. In Appendix F, we show detailed experiments conducted to analyse ResNet-50 following a protocol similar to our previous VGG-16 experiments. A slight advantage for DeepMaps is noticeable in sample images (see Appendix F), with again a better focus on details. Regarding quantitative metrics, GradCAM, FullGrad, and DeepMaps all seem to have comparable quality with respect to the LeRF methods, while DeepMaps seems slightly better for the MoRF metrics. Furthermore, it is notable that DeepMaps has a comparable quality to FullGrad while using information from a significantly lower number of layers. In our experiments, DeepMaps computes attributions from the outputs of the 17 residual blocks of the model, while FullGrad requires the biases of all the 53 layers (including layer normalisation parameters). To conclude, this experiment confirms that the hidden-layer attributions are a source of information that is worth considering to explain the decisions of ResNets as well, showing a more efficient use of information with comparable (or better) results than previous methods.

5 Conclusion

We considered visual explanations of deep image classifiers. We demonstrated that hidden activities in the different layers are a better source of information than biases recommended by earlier work, and proposed a new method, DeepMaps, to generate saliency maps from hidden activities. We also introduced new measures to evaluate saliency maps, which take into account the decisions of the classifier rather than the class scores, and are also able to identify very generic or random explanations. We demonstrated experimentally that DeepMaps clearly outperforms competing explanation methods on VGG-16, and has advantages in explaining the decisions of ResNet-50.

Acknowledgements

We would like to thank Razvan Pascanu and the anonymous reviewers for their insightful comments, which helped improve the presentation of the manuscript. Omar Rivasplata gratefully acknowledges funding from the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/W522636/1. This work was done while Omar was a research scientist intern at DeepMind.

References

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations, 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.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010.

Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *arXiv preprint arXiv:1710.11063*, 2017.

Dabkowski, P. and Gal, Y. Real time image saliency for black box classifiers. In *NIPS*, 2017.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Grill, T. Krippendorff-alpha python package, 2017. URL `https://github.com/grrrr/krippendorff-alpha/blob/master/krippendorff_alpha.py`.

Gur, S., Ali, A., and Wolf, L. Visualization of supervised and self-supervised neural networks via attribution guided factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13), 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un)reliability of saliency methods. In *Explainable AI*, 2017.

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.

Peter, J. P. Reliability: A review of psychometric basics and recent marketing practices:. *Journal of Marketing Research*, 16:6–17, 1979.

Rieger, L. and Hansen, L. K. Aggregating explainability methods for neural networks stabilizes explanations. *arXiv preprint arXiv:1903.00519*, 2019.

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*, 28:2660–2673, 11 2016. doi: 10.1109/TNNLS.2016.2599820.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. Grad-cam: Why did you say that? Visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL `http://arxiv.org/abs/1610.02391`.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153, 2017.

Silberman, N. and Guadarrama, S. Tensorflow-slim image classification model library, 2016. URL `https://github.com/tensorflow/lucid`.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

Srinivas, S. and Fleuret, F. Full-gradient representation for neural network visualization. In *NeurIPS 2019*, 2019.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A. Sanity checks for saliency metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:6021–6029, 04 2020. doi: 10.1609/aaai.v34i04.6064.

A The role of bias parameters in ImageNet classifiers

We surveyed networks trained on ImageNet and available via the Lucid package (Silberman & Guadarrama, 2016) to estimate how bias parameters contribute to prediction accuracy.

In Fig. 8, the networks are arranged by the number of bias parameters. Best performing networks do not require the highest number of biases (e.g. InceptionV2$_{\text{slim}}$).

Removing biases destroys the performance of most classifiers considered (Fig. 8, bottom), except for those that do not use batch normalization: AlexNet, CaffeNet, VGG16/VGG19, InceptionV1 (GoogleNet) and Inception V1$_{\text{caffe}}$. In fact, without biases, only CaffeNet suffers a drop in performance larger than 1%.

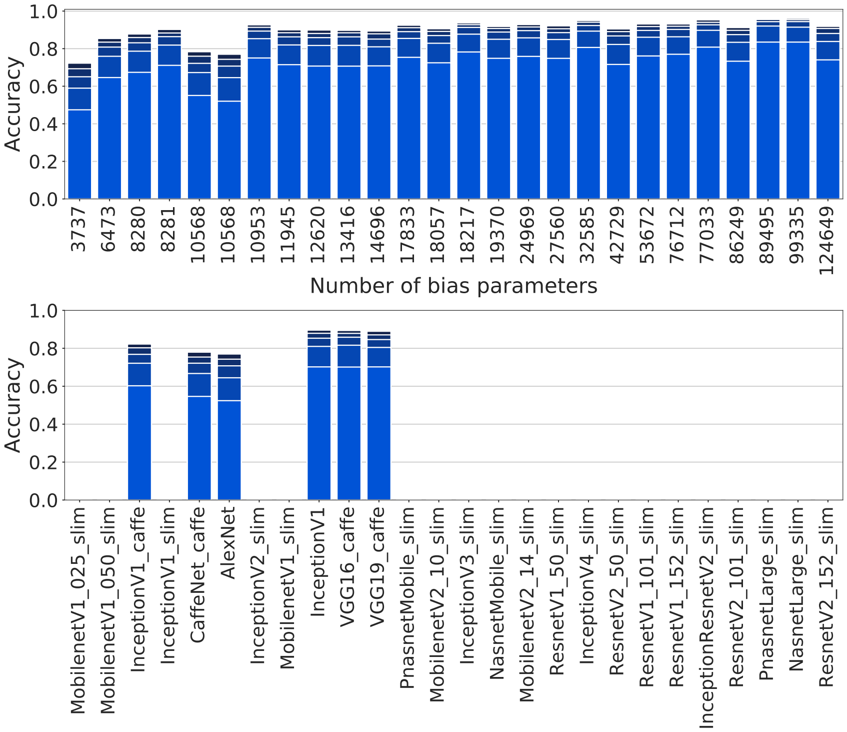

Figure 8: Networks available in the Lucid package. Test accuracy of networks available in the Lucid package before (top) and after setting bias parameters to zero (bottom). The models are arranged according to the number of bias parameters. The total length of the bar denotes top-5 accuracy, while the segments separated with white lines indicate top-1 to top-4 accuracy. All networks sensitive to bias removal use Batch Normalization, and achieve near 0 accuracy after the biases are set to 0. Larger networks achieve higher accuracy.

B AOPC/AFOC metrics

Samek et al. (2016) propose to measure the *area over the perturbation curve*: for an input image $\mathbf{x}$, they define

$$\mathrm{AOPC}(\mathbf{x}) = \frac{1}{P+1} \sum_{k=0}^{P} f_c(\mathbf{x}) - f_c(\mathbf{x}_+^{(k)}), \tag{10}$$

where $c(\mathbf{x}) = \arg\max_{c'} f_{c'}(\mathbf{x})$ is the predicted class for image $\mathbf{x}$, and $\mathbf{x}_+^{(k)}$ is obtained by replacing the k pixels having the highest values in the saliency map with the mean RGB values over the training set (MoRF), and P is the maximum number of pixels removed this way. The AOPC is then obtained as the average value over all images in the dataset (Samek et al., 2016). In our experiments, P is set to remove 10% of the pixels and we use 5000 test images spanning classes 0–99.

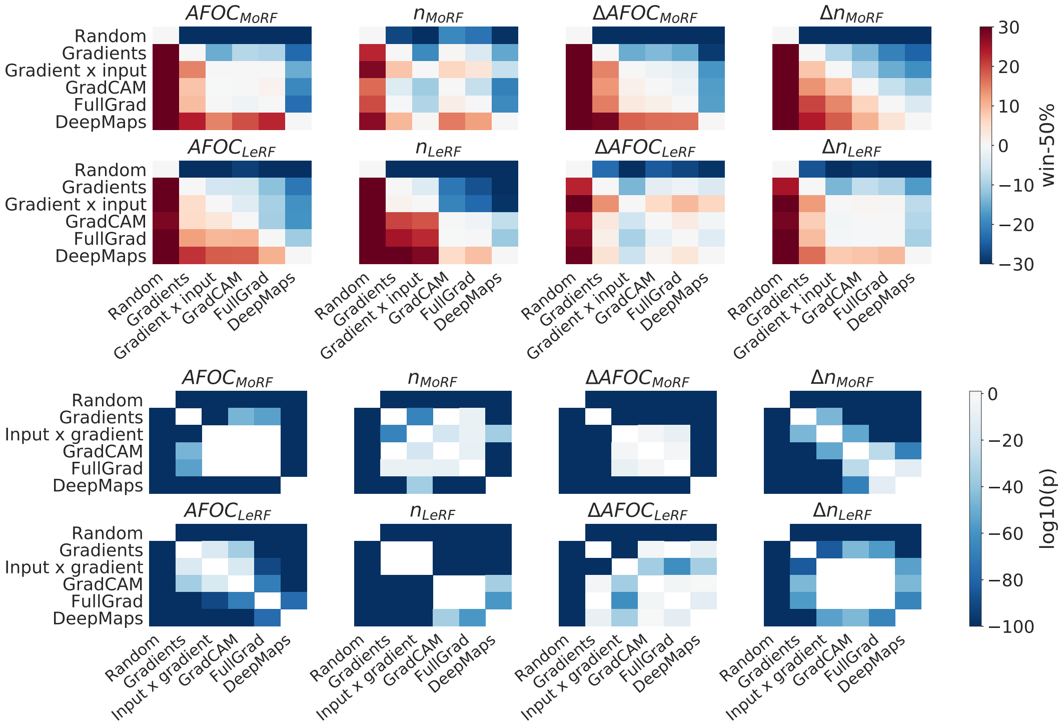

Figure 9: Statistical evaluation of saliency methods in VGG-16. For convenience, we show again the win statistic (top). Wilcoxon signed-rank tests were performed on 5000 pairs of images for each method comparison, the resulting p-values are shown on a logarithmic scale (see colorbar). For example, the first row/column in each heatmap communicates that random choice of pixels is highly significantly ($p < 10^{-100}$) different to saliency maps from any other method. All metrics except for $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$ and Δn_{MoRF} attest to a highly significant benefit of using DEEPMAPS to extract visual explanations from VGG-16 ($p < 10^{-30}$). See the text for details.

If an explanation succeeds at detecting important regions of the input, the decrease in the class score is expected to be high. $\mathrm{AOPC}(\mathbf{x})$ is then expected to be higher for better explanations.

Srinivas & Fleuret (2019) consider the *absolute fractional output change* for input $\mathbf{x}$ defined as

$$\mathrm{AFOC}(\mathbf{x}) = \frac{|f_{c(\mathbf{x})}(\mathbf{x}) - f_{c(\mathbf{x})}(\mathbf{x}_{-}^{(k)})|}{f_{c(\mathbf{x})}(\mathbf{x})}, \tag{11}$$

where $\mathbf{x}_{-}^{(k)}$ denotes the image after replacing its k least-relevant pixels with the average pixel over the training set (LeRF). The other quantities are defined as for AOPC. Srinivas & Fleuret (2019) remove up to 10% of the image.

Note that both metrics can have other variants, for example a MoRF version of AFOC or a LeRF version of AOPC. When both variants are used, they are distinguished by appropriate subscripts. All the statistical evaluations (Fig. 7, Fig. 9 and Fig. 16) are based on AFOC measured for both the most and the least salient perturbations (removing 10% of the image).

C Further empirical analysis of the evaluation metrics

In this section we provide further analysis of the behavior of the metrics, including statistical significance of our findings as well as the consistency of the saliency metrics.

C.1 Pairwise tests for VGG-16

In Fig. 9, we present p-values from the Wilcoxon signed-rank test ($\mathrm{scipy.stats.wilcoxon}$, `SciPy.org`), performed on data shown in Fig. 7, which is repeated in Fig. 9 for convenience. Colours are chosen such that white corresponds to the null hypothesis (metric values derived from two saliency methods are statistically indistinguishable). The darkest colour corresponds to $p < 10^{-30}$; note, however, that many of the tests yield values as low as 10^{-200}, or even zero.

All metrics except for $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$ depict DEEPMAPS as the best method for extracting visual explanations from VGG-16 at a highly statistically significant level ($p < 10^{-30}$). See below for our hypothesis why $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$ is of a limited use. $\mathrm{AFOC}_{\mathrm{LeRF}}$, on the other hand, appears much more informative, distinguishing between every single explanation method with high significance (the least significant p-values are on the order of 10^{-10}).

While $\mathrm{AFOC}_{\mathrm{LeRF}}$ seems to be much better metric than $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$, the opposite is true for the MoRF metrics based on network's decisions: While n_{MoRF} overestimates the quality of gradient×input attributions (see below), the Δn_{MoRF} metric ranks saliency methods in the same, intuitive way as $\mathrm{AFOC}_{\mathrm{LeRF}}$. Higher significance (lower p-values) are achieved with Δn_{MoRF} than with n_{MoRF} and with $\Delta\,\mathrm{AFOC}_{\mathrm{MoRF}}$ than with $\mathrm{AFOC}_{\mathrm{MoRF}}$ in most pairwise comparisons (Fig. 9 bottom).

In summary, Δ comparisons refine the message carried by n_{MoRF}, and they provide a complementary information to n_{LeRF}. For VGG-16, they have a small effect on $\mathrm{AFOC}_{\mathrm{MoRF}}$ and corrupt the message carried by the $\mathrm{AFOC}_{\mathrm{LeRF}}$ metric, as described below.

On gradient×input attributions Both the $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$ and n_{MoRF} metrics indicate a supreme quality of gradient×input attributions. This might be an unfortunate side effect of our choice of image perturbations, which replace pixels with their expected value — grey. As a result, "removing" grey pixels amounts to no change. Now, gradient×input saliency maps attribute the lowest scores to grey pixels (see Fig. 3). This is because grey corresponds to zero value across all colour channels, curtailing the gradient×input score. Thus, gradient×input saliency maps will get an automatic boost in all LeRF metrics.

In a symmetrical manner, we can consider the most extreme pixel values, corresponding to black and white. Removing these amounts to the strongest possible change to the input. On the other hand, black and white are favoured by gradient×input scores. Thus, gradient×input saliency maps would induce stronger perturbations in the MoRF procedures, leading to a boost of the corresponding metrics. We believe this automatic, input-induced mechanism underlies the apparent benefit of gradient×input attributions method according to $\Delta\,\mathrm{AFOC}_{\mathrm{LeRF}}$ and n_{MoRF} metrics. We would advise against using these metrics.

C.2 Consistency of evaluations

Tomsett et al. (2020) suggested that good metrics to evaluate explanation methods should be consistent in the sense that they should rank different explanation methods similarly on individual images. Following their work, in this section we analyze the consistency of the evaluation metrics considered in the paper in ranking the different explanation methods, as well as the similarity of the different metrics (as baseline, we only present results for AFOC and not for AOPC, since the former is consistently slightly better and otherwise they behave similarly). We consider two measures from Tomsett et al. (2020), taken from the psychometric literature (Peter, 1979): inter-rater reliability and internal consistency reliability.

Inter-rater reliability is measured by Krippendorff's α coefficient, which quantifies the consistency of ranking over multiple images: a value of 1 (the maximum value) means that all explanation methods are ranked in the same way on every image and 0 that the rankings on different images are independent. We used the Krippendorff-alpha package (Grill, 2017), and applied it to the ranks assigned by a given metric to the six saliency methods considered (Random, GRADIENTS, gradient×input, GRADCAM, FULLGRAD, DEEPMAPS). The average α coefficients, computed over 500 randomly selected images (out of 5000) and repeated 50 times, are shown in Fig. 10 for VGG-16. One can observe that our n_{LeRF} and Δn_{MoRF} saliency metrics are more consistent than the traditional metrics. Moreover, our modification $\Delta\sigma$ (cf. Eq. (9)) to compare to the alternative explanations clearly improves the consistency for any MoRF metric σ. On the other hand, this modification decreases the

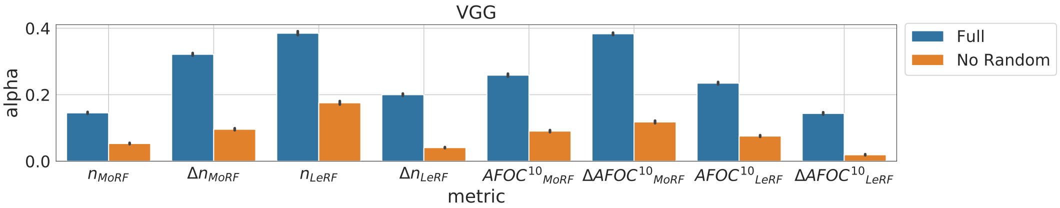

Figure 10: Inter-rater reliability: Krippendorff's α coefficient of consistency of the evaluation metrics for the five saliency methods GRADIENTS, gradient×input, GRADCAM, FULLGRAD, DEEPMAPS (”No Random”) and for six methods also including Random (”Full”). The average α is shown for 500 randomly selected images (out of 5000) repeated 50 times. The error bars represent 95% confidence intervals evaluated over the 50 random draws. While not presented here, we get very similar results for the other saliency maps considered. For ResNet-50, see Fig. 17.

consistency for LeRF metrics. A potential reason for this is that our images are often centered, thus the unimportant regions of the images may overlap quite a bit, hence alternative LeRF explanations can work reasonably well (see the paragraph ”On gradient×input attributions” in Appendix C.1 for another reason behind $\Delta\sigma$ being less effective in the LeRF setup).

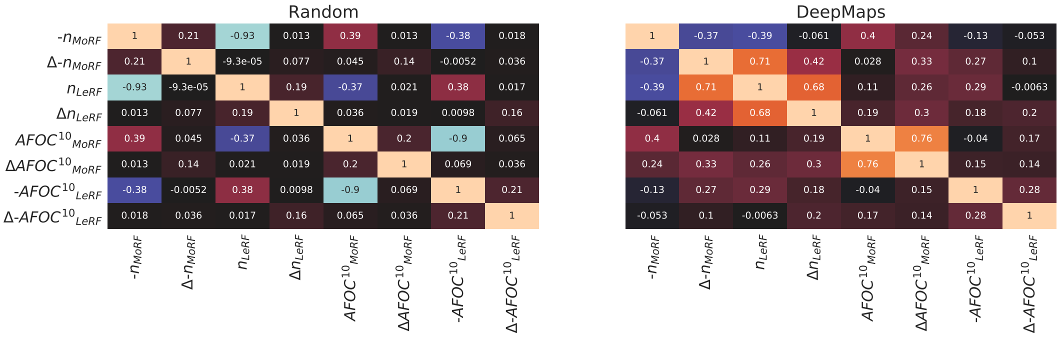

Figure 11: Internal consistency reliability, as measured by Spearman's correlation between metric values assigned to 5000 visual explanations for VGG-16. All other methods of extracting visual explanations showed a very similar picture to the one shown here for DEEPMAPS.

The internal consistency reliability, computed by Spearman's rank correlation ρ, measures pairwise similarity between the saliency metrics. Fig. 11 presents pairwise correlations between all evaluation metrics (we use the negative of the n_{MoRF} and $\mathrm{AFOC}_{\mathrm{LeRF}}$ metrics since for these, larger is better), for two saliency methods, Random and DEEPMAPS (computed over our dataset of 5000 images). Results on the Random saliency map confirm that applying $\Delta\sigma$ instead of the original evaluation method σ reduces the similarity in evaluating a meaningless (random) saliency map: while n_{MoRF}, n_{LeRF}, and $\mathrm{AFOC}_{\mathrm{MoRF}}$, $\mathrm{AFOC}_{\mathrm{LeRF}}$ are highly correlated,[5] their Δ-transformed versions are not. On the other hand, the Δ-transformation has much smaller impact on sensible visual explanations, such as DEEPMAPS, for the evaluation metrics n_{LeRF} and $\mathrm{AFOC}_{\mathrm{MoRF}}$ ($\rho = 0.68$ and $\rho = 0.76$, respectively). Thus, Δn_{LeRF} and $\Delta\,\mathrm{AFOC}_{\mathrm{MoRF}}$ measure similar qualities of explanations as intended with the original metrics. These conclusions are similar to those we drew by comparing win statistics (Fig. 9).

Taken the consistency results into account, if one wants to use multiple metrics to evaluate saliency methods as recommended by Tomsett et al. (2020), we suggest to use the metrics $\Delta\,\mathrm{AFOC}_{\mathrm{MoRF}}$ (or Δn_{MoRF}) and n_{LeRF}, as they are the most consistent while measuring different aspects of the visual explanations, with $\rho(\Delta\,\mathrm{AFOC}_{\mathrm{MoRF}}, n_{\mathrm{LeRF}}) = 0.26$ ($\rho(\Delta n_{\mathrm{MoRF}}, n_{\mathrm{LeRF}}) = 0.42$).

While all the results presented in this section were shown for VGG-16, we get very similar results for ResNet-50 (see Appendix F).

[5] Correlations between these pairs of metrics are negative, because they measure similar qualities, but favour a different result of the perturbation: MoRF procedures prefer images that are easy to perturb, while the opposite is true for the LeRF metrics.

D Attributions from single layers

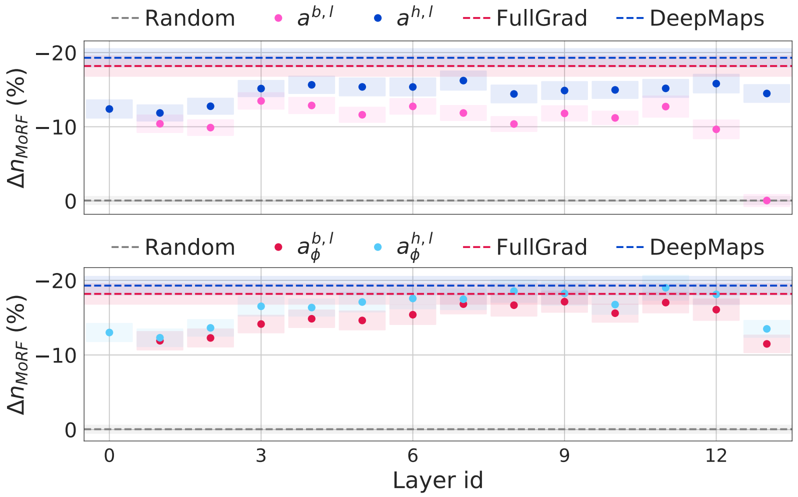

Figure 12: Single-layer saliency maps evaluated with Δn_{MoRF}. Explanations are generated for each network layer, the label indicates where the Ψ transformation is applied: per layer (top, resulting in $\Psi(\sum_{\phi} \mathbf{a}_{\phi}^{b,l})$, $\Psi(\sum_{\phi} \mathbf{a}_{\phi}^{h,l})$ saliency maps) or for each channel (bottom, yielding $\sum_{\phi} \Psi(\mathbf{a}_{\phi}^{b,l})$ and $\sum_{\phi} \Psi(\mathbf{a}_{\phi}^{h,l})$). Random, FullGrad and DeepMaps methods are included for reference.

We showed that DeepMaps is competitive as a method for producing visual explanations. Here, we take a step back to evaluate what makes the difference. In Fig. 12, we use one of the most informative metrics we introduced, Δn_{MoRF}, to evaluate the quality of visual explanations produced by each layer of VGG-16. In the top panel of Fig. 12, our evaluation relates to per-layer attributions $\mathbf{a}^{h,l}, \mathbf{a}^{b,l}$ (such as the ones shown in Fig. 2). Our metrics confirm the visual assessment: for every layer, hidden activities yield a better explanation than the corresponding biases. The quality of the last convolutional layer (the problematic one, with the highest impact on the final score, layer 13) is zero. The bottom panel in Fig. 12 relates to the alternative way of aggregating explanations: while DeepMaps first aggregates attributions by channels (summing over ϕ in Eq. (6)) and then normalises with Ψ, FullGrad normalises every feature attribution (such as $\mathbf{a}_{\phi}^{b,l}$) and then sums these for all ϕ. In the bottom panel the normalisation for each layer is done per channel. Comparing the two figures, one can see that summing bias-explanations (bottom, red) rather than attributions (top, pink) improves the quality of the saliency maps, especially for layer 13 (recall that bias-attributions were dominated by this single layer). On the other hand, the order of normalisation and aggregation has little effect on explanations from hidden activities.

Here, both methods aggregate over all layers in the network. Given the variable quality of layer attributions, it is possible that aggregating a subset of layers would do a better job. This requires the development of a proper procedure and is left for future work.

Using matrix factorisation, as done by Gur et al. (2021), is another idea of combining layer- and channel-specific explanations, which is likely to benefit approaches such as FullGrad (based on input and bias attributions) or DeepMaps (hidden layer attributions).

E SmoothGrad

SmoothGrad, by Smilkov et al. (2017), is another approach to increase human readability of saliency maps. The trick is to aggregate saliency maps for a bunch of samples from a local neighbourhood of a given image. Average over samples appears less noisy than any sample and is often easier to decipher by eye, but is it more faithfully depicting the network's sensitivity?

We combined SmoothGrad with the methods studied in our paper. While all saliency maps appear "smoothed" by this method, the metrics improve only for Gradients, and the methods based on deeper activities remain better, with DeepMaps without smoothing remaining the best, as shown in Fig. 13.

Thus, according to our metrics, SmoothGrad does not produce more faithful saliency maps. On the other hand, we could apply the smoothing idea of SmoothGrad to the metrics as well (by performing perturbations on local neighbours of a given image). This would prefer explanations which are more robust to input perturbations. Analysing such metrics is left for future work.

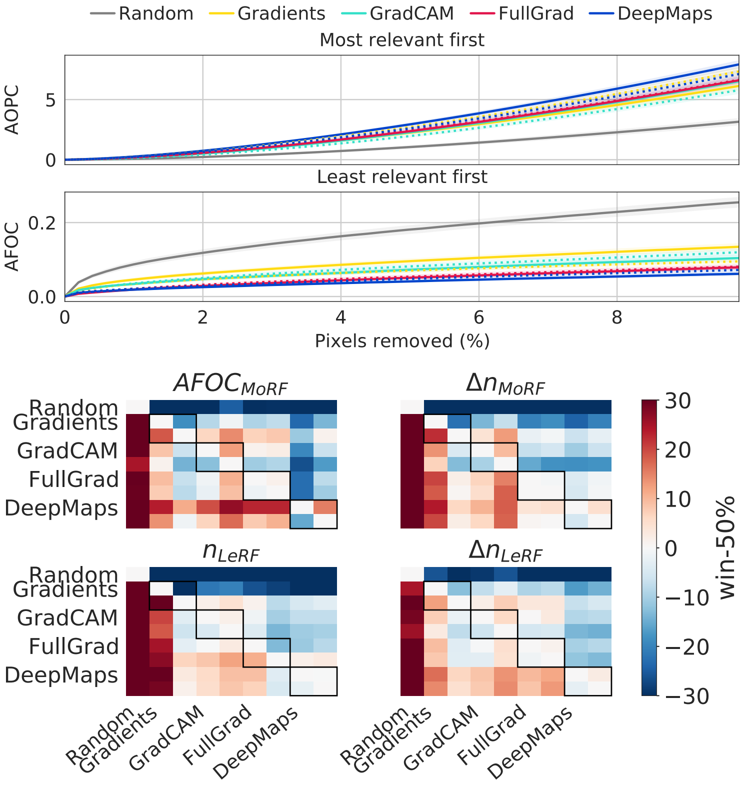

Figure 13: Top: SmoothGrad (dotted lines) compared to the original methods (solid lines); Bottom: Win statistics with our recommended choice of metrics. Every second row/column corresponds to the SmoothGrad version of each method (black boxes highlight comparison of each method to their SmoothGrad version).

F ResNET

Details of the ResNet-50 architecture are given by He et al. (2016). In short, the network consists of 1 + 16×3 blocks of convolutions. After the initial convolution (which reduces the output size from 224×224 to 112×112), the convolutions come in blocks of three, each triplet accompanied by a residual connection and followed by a ReLU nonlinearity (see Fig. 3 of He et al., 2016). In our implementation of DeepMaps, we extract explanations from ReLU layers that follow the initial convolution, and each of the triplets, i.e. every time the parallel routes of computation come together again. Altogether, these amount to 17 hidden-activity attributions ($\mathbf{a}^{h,l}$).

In the implementation of FullGrad, we had to extract all parameters due to the batch normalisation (see below). These amounted to 53 blocks, and included parallel computational routes. The only regular bias parameters used in this network are assigned to the output layer, and as such, they do not enter the computation of the FullGrad saliency maps.

Batch normalisation Batch normalisation (Ioffe & Szegedy, 2015) results in the following additive component:

$$\mathrm{BatchNorm}(x_k) = \beta - \gamma \frac{\mathrm{mean}(x_k)}{\sqrt{\mathrm{var}(x_k) + \epsilon}}\,, \tag{12}$$

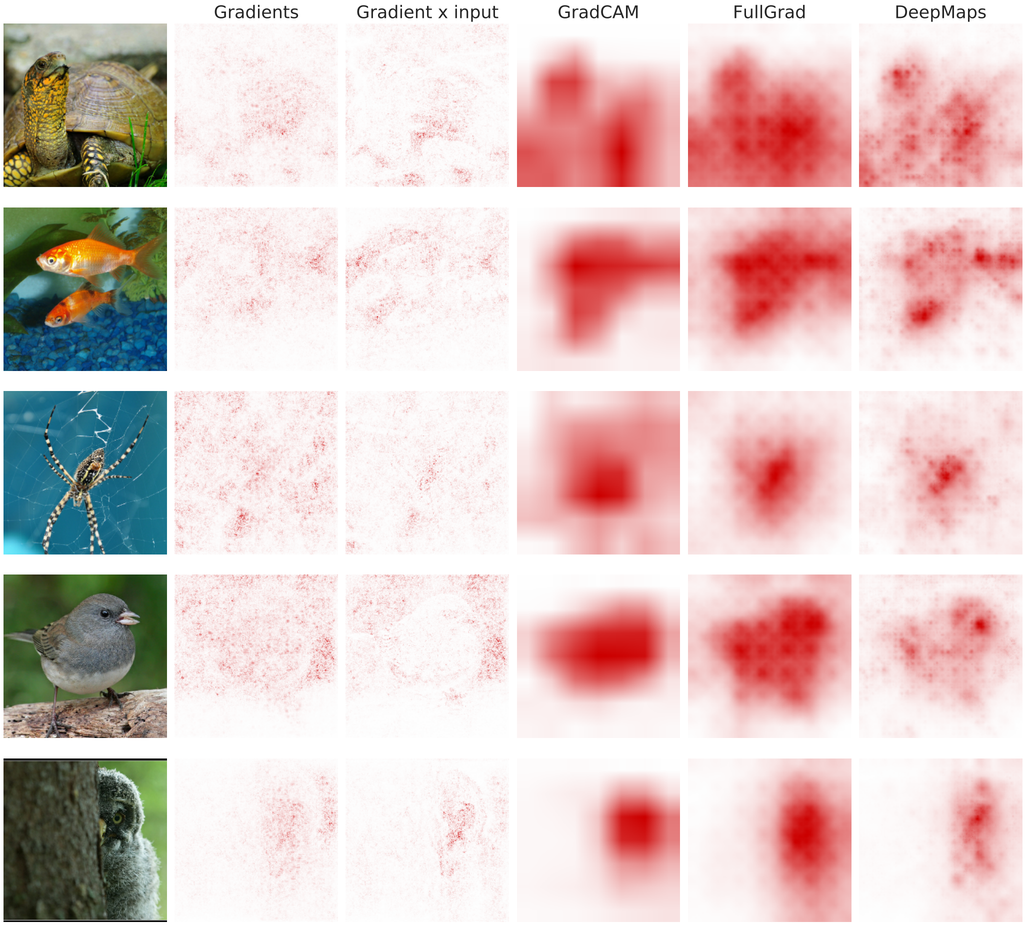

Figure 14: Example saliency visualisations for ResNet-50. As in Fig. 3, DEEPMAPS offers explanations with a finer resolution detail than the alternative methods. Specifically, it highlights eyes, beaks, heads of the animals, features that offer strong clues about the animal species. A chequered pattern shows in FULLGRAD saliency maps. gradient×input attributions devalue grey pixels (such as jaybird's belly) and attribute high importance to black and white pixels.

where β and γ are trainable parameters and x_k is the activity of a given neuron (or a convolutional kernel). During training, activity statistics are collected online, however, at test time, these act as parameters. [6]

Saliency maps In Fig. 14, we show example saliency maps for ResNet-50. Similarly to VGG-16, the explanations extracted by DEEPMAPS contain finer details than those extracted with FULLGRAD and GRADCAM. All three are more interpretable than the GRADIENTS explanations. A chequered pattern appears in FULLGRAD saliency maps, it is much less prominent in DEEPMAPS and absent from GRADCAM, indicating that it is not due to the top convolutional layer.

Evaluation In Fig. 15, we present results for ResNet-50 using the same metrics as for VGG-16 in Fig. 6. On average, random perturbations have the same effect in both networks — around 23% of the image needs to be removed for the classification to change. However, when guided by any other saliency map, it takes more pixels to change the decisions of ResNet-50 than of VGG-16, both for the MoRF and LeRF perturbations.

[6] To pull out the relevant values from a saved TensorFlow v.1 model was a technical challenge; we decided against open-sourcing this model-specific source code as we would not be able to maintain it given the fast

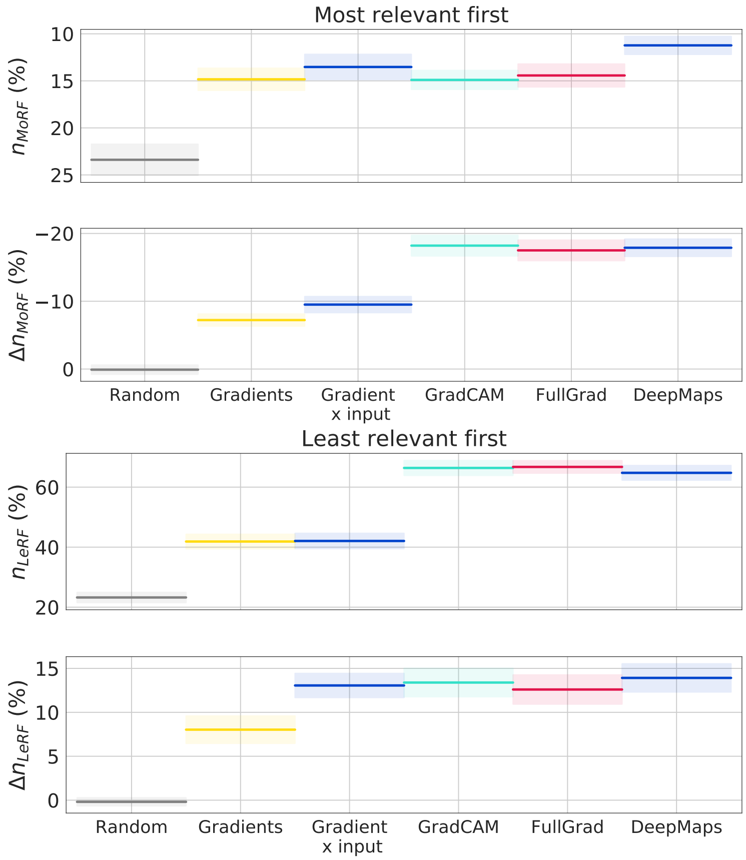

Figure 15: Comparison of saliency methods in ResNet-50 using decision-based pixel-removal metrics. The coloured boxes depict one standard deviation of the mean across classes ($n = 50$).

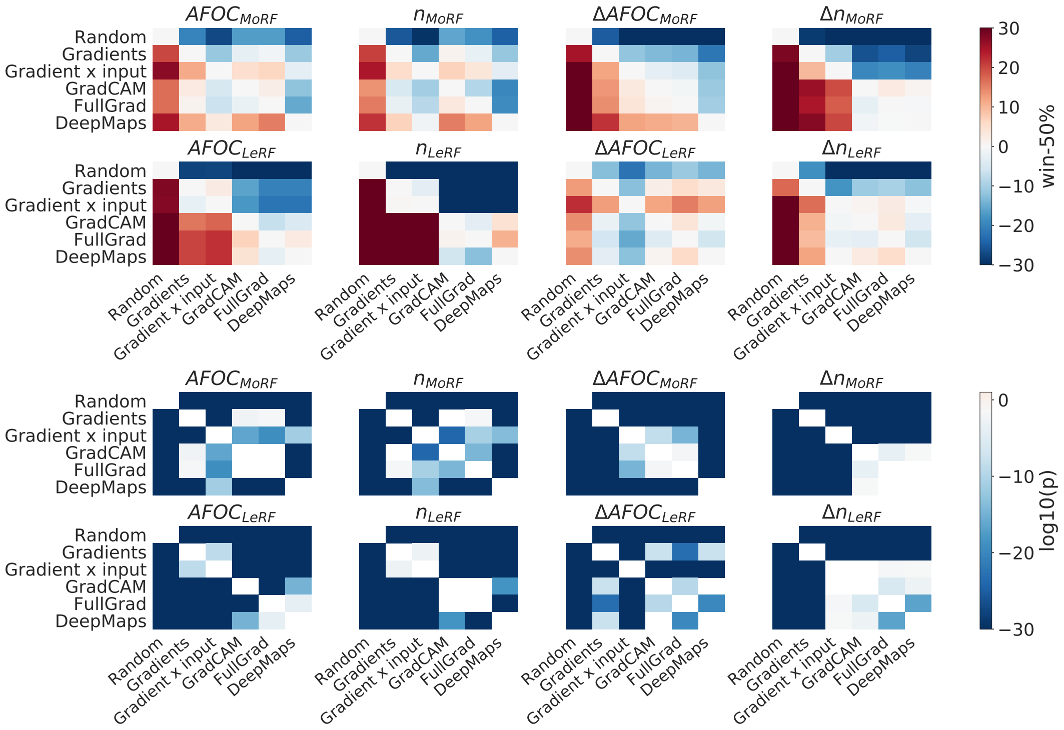

Figure 16: ResNet-50 win statistic and p-values.

There is a striking similarity between the shape of n_{MoRF} for the two networks, with identical ranking of the explanation methods according to this metric. Here again, gradient×input attributions appear competitive to the other, more interpretable methods, indicating that this metric is not faithfully reflecting our explainability prior (see Appendix C.1).

When comparing Δn_{MoRF} in Fig. 6 and Fig. 15, there is a clear difference for GRADCAM, which in ResNet-50 appears to be as effective as FULLGRAD and DEEPMAPS.

A similar picture is derived from the n_{LeRF} metric: both GRADCAM and FULLGRAD allow to remove more pixels in LeRF fashion than in the case of VGG-16.

Pairwise comparisons (Fig. 16) confirm these observations: FULLGRAD offers more concise explanations (highest n_{LeRF} count, albeit not significant, with $p > 0.1$ when compared to GRADCAM); LeRF depiction is more image-specific in DEEPMAPS (highest Δn_{LeRF} count, $p < 0.002$, win=52.7% compared to GRADCAM). GRADCAM appears most accurate in image-specific delineation of the most important regions (lowest Δn_{MoRF} count), albeit the difference to the other methods is small ($p < 0.04$, win=51.1% compared to DEEPMAPS). DEEPMAPS is the best according to most of the MoRF metrics, although for $\text{AFOC}_{\text{MoRF}}$ and n_{MoRF}, they compete with gradient×input attributions (the problem we deliberated on in Appendix C.1). According to Δn_{MoRF} and LeRF metrics, they are very similar to FULLGRAD and GRADCAM.

In summary, for ResNet-50, there is no clear choice for the explanation method. Depending on whether we care about the change in classification confidence (as approximated by AFOC) or classification itself (measured by the number of pixels), we might choose DEEPMAPS, FULLGRAD or GRADCAM. However note that in DEEPMAPS, we have the flexibility of choosing which maps to aggregate. For example, limiting the aggregation to the top layers is likely to help against the chequered pattern evident in Fig. 14 for FULLGRAD and DEEPMAPS, but not for GRADCAM, which focuses on the top layer.

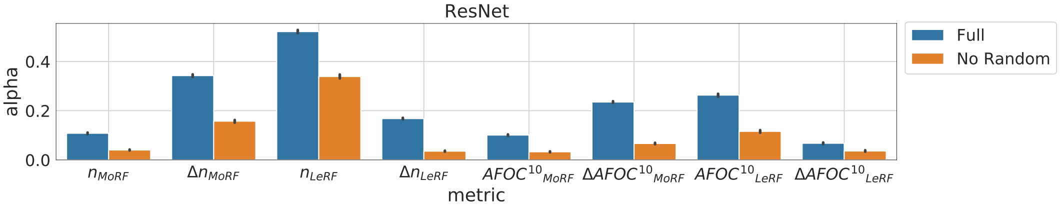

Figure 17: Inter-rater reliability: Krippendorff's α coefficient of consistency of the evaluation methods on ResNet-50. See Fig. 10 for details.

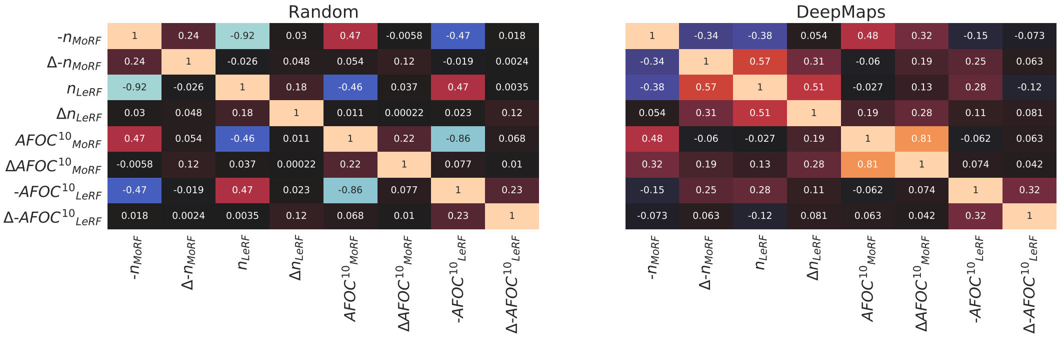

Figure 18: Internal consistency reliability, as measured by Spearman's correlation between metric values assigned to 5000 visual explanations for ResNet-50. See Fig. 11 for details.

pace of upgrades to TensorFlow. We verified that all the input and bias attributions sum up to the class score, $f_c(\mathbf{x}) = \sum_{i,j} a_{cij}(\mathbf{x}) + \sum_{l,\phi} \mathbf{a}^{b,l}_{c\phi}(\mathbf{x})$.