Aggregating explainability methods for neural networks stabilizes explanations

Preprint · March 2019

DOI: 10.48550/arXiv.1903.00519

Laura Rieger
Technical University of Denmark

Aggregating explainability methods for neural networks stabilizes explanations

Laura Rieger [1] **Lars Kai Hansen** [1]

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. In fact, most works rely on manually assessing the explanation to evaluate the quality of a method. This injects uncertainty in the explanation process along several dimensions: Which explanation method to apply? Who should we ask to evaluate it and which criteria should be used for the evaluation? 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. Our findings show that the aggregation is more robust, well-aligned with human explanations and can attribute relevance to a broader set of features (completeness). Second, we propose a novel way of evaluating explanation methods that circumvents the need for manual evaluation and is not reliant on the alignment of neural networks and humans decision processes.

1. Introduction

While neural networks have achieved great success in classic visual recognition problems, explaining the networks' decisions remains an open research problem. 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. The challenge of explaining neural network decisions is amplified by the fact that there is no agreement on what a sufficient explanation is and accordingly, 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 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 epistemological uncertainty of the explanation task.

Our contribution is motivated by a key insight in machine learning: Ensemble models can potentially 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 and thus reducing epistemic uncertainty in neural networks. For further discussion of neural network uncertainty see e.g. (Kendall & Gal, 2017)

First, we test the hypothesis that ensembles of multiple explanation methods are more robust than single explanation methods. We analyze this idea theoretically and evaluate it 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. Second, we explore a new approach to quantitatively evaluate explanation methods without relying on human evaluation. We circumvent the problem of high correlation between neighbor pixels and human bias that is present in current evaluation methods.

2. Related Work

The new-found interest into explainability is reflected in a lot of recent work into the problem (Kindermans et al., 2017; Selvaraju et al., 2017; Bach et al., 2015; Zhang et al., 2018; 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). In this paper, we focus on gradient-based explanation methods (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).

Simonyan et al. (2013) is to our knowledge the earliest work that proposed backpropagating the output onto the input to gain an understanding of a neural network decision. They propose visualizing the saliency of a specific class output in regards to the input. Springenberg et al. (2014) extended on that idea by introducing guided backpropagation

[1]DTU Compute, Technical University Denmark, Lyngby, Denmark. Correspondence to: Laura Rieger <lauri@dtu.dk>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

with applying ReLU non-linearities on the backpropagation. This removes noise as compared to a regular saliency map. Selvaraju et al. (2017) proposes an explanation method specifically for convolutional neural networks called Grad-CAM. By backpropagating relevance through the dense layers and up-sampling the evidence for the convolutional part of the network, they obtain rough activation maps that show relevant parts of the input image. They also proposed guided Grad-CAM, the product of guided backpropagation and Grad-CAM. Compared to guided backpropagation it reduces noise, since all the relevance concentrates in the area marked relevant by Grad-CAM. Bach et al. (2015) proposed backpropagating relevance with specific attribution rules. The general principle is the same as in Simonyan et al. (2013); Springenberg et al. (2014) but the way relevance is propagated from one layer to the next is different.

Integrated Gradients is a method proposed in 2017 by Sundararajan et al. (2017) that sums up the gradients from linearly interpolated pictures between a baseline (f.e. a completely black image) and the actual image as a way to circumvent noise and highly correlated features. SmoothGrad is a recently proposed modification of vanilla explanation methods. They propose filtering out noise from a basic saliency map by creating many samples of the original input with Gaussian noise (Smilkov et al., 2017). They obtain the final saliency map by averaging over all samples. This leads to a clearer heatmap, since much of the noise is filtered out.

In recent years there has also been more work addressing theoretical flaws or empirical fail-modes within explanation methods. (Adebayo et al., 2018; Kindermans et al., 2017; Ghorbani et al., 2017) The evaluation of explanation methods is a relatively new topic. Recently Adebayo et al. (2018) showed that using visual assessment to evaluate explanation methods can be misleading. In particular, they show that several widely used explanation methods are visually similar to simple edge detection. Bau et al. (2017) introduced a large dataset of semantically annotated images that can be used to measure the alignment between natural human-known concepts and hidden layers in the neural network. They used this dataset to measure the effect of different training strategies and algorithms on disentanglement and interpretability.

Bach et al. (2015) proposed a novel way to evaluate an explanation method without relying on visual evaluation. By flipping pixels to their opposite in descending order of relevance and measuring how fast the output score goes down, it is possible to measure how well the explanation method reflects features deemed important by the machine learning algorithm. Ancona et al. (2018) proposed Sensitivity-n, the notion that for every n number of inputs, the decrease in output score, when those inputs are canceled out, should be equal to the relevance of those inputs. This is fulfilled completely if and only if all inputs contribute linearly and independently to the output. Hooker et al. (2018) followed a similar approach to ours in measuring how much accuracy degrades when part of the input is replaced with zero-information input. In contrast to our method, they first require retraining the same architecture from scratch on the degraded dataset.

3. Aggregating explanation methods allows for more expressive representations

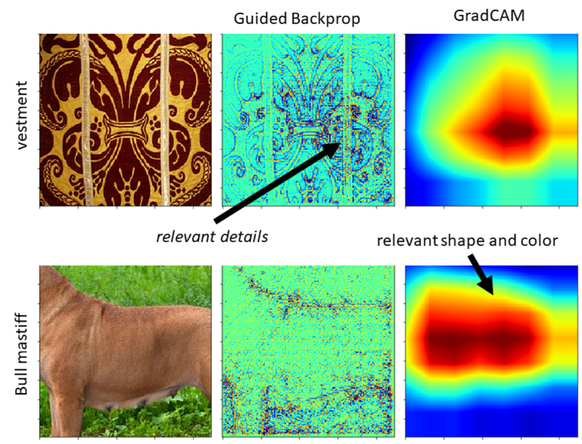

Figure 1. Different explanation methods fail to pick up different feature types. Upper row: Details and structure are important for classification. Lower row: Large segment is important for classification.

A deficiency that has not been addressed yet is that different explanation methods will display different kinds of features such as structure, color or shapes. Therefore a given method will fail to adequately explain any given classification if the true explanation is a mixture of features. We demonstrate this for two well-known methods, Guided Backpropagation and Grad-CAM in Fig. 1. Grad-CAM highlights general areas, whereas Guided Backpropagation highlight edges and details. As a result, the former cannot adequately explain the upper image, which is composed of fine details. The latter cannot adequately explain the lower image, where a large shape and color are important. Guided Grad-CAM, the product of the two, would also fail to explain the lower row since it would only highlight details within the important area and fail to reflect the whole feature.

By combining explanation methods that can highlight different kinds of features, the explanation can display a richer combination of features than any method on its own. This becomes especially necessary for more complicated classification tasks, where a complex explanation is needed.

4. Aggregating explanation methods reduces noise

As has already been explored in literature, all currently available explanation methods based on backpropagation have weaknesses inherent in 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 already have been used for a long time to reduce variance and bias of machine learning models. We apply the same idea to explanation methods and build an ensemble of explanation methods. Since we initially do not have any information about which method performs better, it is desirable for all methods to have the same weight. We average over all available methods.

We show that the error of the average of multiple methods will be lower than the typical error of the individual methods used. For this, we assume the existence of a hypothetical 'true' explanation y_n for the output of a neural network for image n. We further define $\mu_{j,n}$ the explanation obtained with method $j \in [1, ..., J]$ for the output of a given neural network for image n, and mean aggregate explanation $m_n = \frac{1}{J}\sum_{j=1}^{J} \mu_{j,n}$.

We define the error of explanation method j on image n as

$$E_{j,n} = ||\mu_{j,n} - y_n||^2$$

and $E_m = \frac{1}{N}\sum_n ||m_n - y_n||^2$ is the mean error of the aggregate for N images. The typical error of an explanation method is represented by the mean

$$\begin{aligned}\overline{E} &= \frac{1}{J}\sum_j \frac{1}{N}\sum_n ||\mu_{j,n} - y_n||^2 \\ &= \frac{1}{J}\sum_j \frac{1}{N}\sum_n \left(||\mu_{j,n} - m_n||^2 + ||y_n - m_n||^2\right),\end{aligned}$$

hence,

$$\overline{E} = \frac{1}{J}\sum_j \frac{1}{N}\sum_n ||\mu_{j,n} - m_n||^2 + E_m \geq E_m$$

We see that the error of the aggregate E_m is less than the typical error of the participating models. The difference - a 'variance' term - represents the epistemic uncertainty and only vanishes if all methods produce identical maps.

5. Considering uncertainty in the explanation

Hence, by a simple average over all available explanation methods, we reduce the variance of the explanation compared to using a single method. We subsequently refer to this approach as *AGG-Mean*. This estimator does not take into account the estimate of the local epistemic uncertainty provided by the pixel-wise variance over methods. A straightforward route to incorporate this information would be to form an 'effect size' map by normalizing the mean aggregate locally by its standard deviation. Intuitively, this will assign less relevance to segments with high disagreement between methods. However, since it is not guaranteed that there is a non-zero standard deviation, we need to stabilize the uncertainty estimate. We consider two stabilizers. One possibility is to only normalize by the standard deviation when the mean value is above a certain threshold. By setting the threshold appropriately low, we ensure that variance between methods is only taken into account for segments where a certain level of relevance is present in the ground truth. We refer to this scheme as *AGG-Clipped*. Finally, we consider regularizing the estimate of local variance by an additive constant. This could either reflect a priori information regarding epistemic or aleatoric uncertainties or simply be considered a smoothing regularizer. This variant will be called *AGG-Posterior*. In Section 7 we compare the three methods against non-aggregated methods.

6. Evaluating explanation methods in a quantitative way with segment flipping

A recurring problem with new explainability methods is how to evaluate them in an objective way. (Bach et al., 2015) proposed using 'pixel-flipping', where pixels with high importance are consecutively flipped to the opposite color. Unfortunately, this only works for low-dimensional input. For images, this translates to low-resolution images, where there is a low correlation between neighboring pixels.

For high-dimensional pictures, features are composed of a non-linear combination of pixels in close distance. If one pixel is flipped, this will not change the overall feature and should therefore not result in a changed output score. As a result, the relevance values of single pixels are not indicative of the feature's importance as a whole.

To remedy this problem, we propose using a conventional segmentation algorithm to sub-divide the image into meaningful segments. In this work, we used Quickshift for this task (Vedaldi & Soatto, 2008). For explainability methods, that provide the importance of each pixel, the importance of each segment is the mean of the values for each pixel. By sorting the segments according to their importance, we get a ranking of how relevant each segment is. We identify segments in decreasing order of their relevance. Each segment is replaced with the mean pixel value over the entire dataset. The process is shown in Fig. 2. After a segment is replaced with the mean pixel value, hereby referred to as greying out, the resulting image is classified again with the neural network and the resulting softmax score for the correct class is reported. The result is a curve of the softmax class score dependent on how many segments of the image are grayed out, such as in Fig. 4.

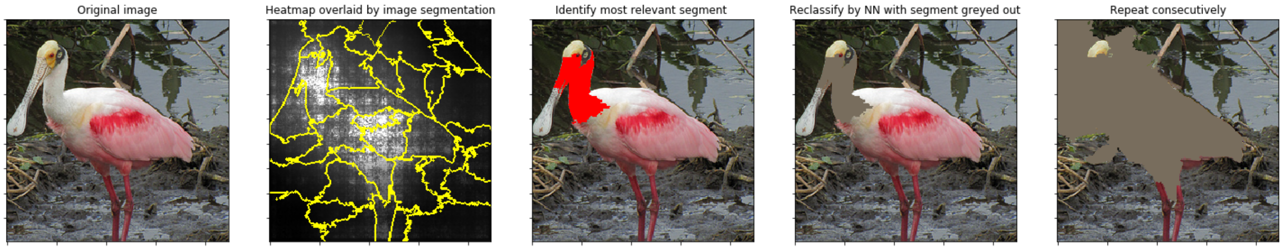

Figure 2. Quantitative evaluation: Relevant segments as identified by an explanation method get consecutively replaced by the mean. The decrease of the class score over the number grayed segments is reported.

If an explanation method works well, it will identify relevant segments with high accuracy. As a result, the softmax score for the correct class will go down faster because relevant segments are grayed out earlier in the process. We can compare two methods by checking for which method the softmax score is reduced quicker. We can use this as a quantitative comparison of two or more explainability methods that does not rely on human evaluation. Apart from being able to quickly compare two methods, we also bypass any human bias in this way. The approach only measures how well segments relevant for classification by the network are identified and does not rely on supplied prior knowledge about the classes. As a result, it can also be used with tasks where there is insufficient knowledge about what the neural network should mark as relevant.

In addition to introducing a quantitative measure of explanation methods, we open up the future possibility of comparing two networks by comparing how quickly the class score by one network goes down when replacing segments that are important for the other network. This could offer another valuable metric for measuring the stability of network performance. If graying out segments that one network marks as relevant decreases the score for the other network significantly, both networks learned stable features.

7. Experiments and discussion

We evaluated our proposed method with a number of experiments and metrics. As discussed in Section 6, the evaluation of an explanation method is difficult due to the fact that no definitive ground-truth for an explanation exists. Therefore we use multiple evaluation metrics that measure different aspects of explainability and consider the performance in all of these metrics. In Sections 7.2 and 7.3 we measure how accurately the explanation reflects the decision process of the network. In Sections 7.4 and 7.5 we measure the agreement of the explanation with human evaluation as a proxy for understandability by humans. We tested our method on well-known networks that were pre-trained on ImageNet: Inception, VGG19 and Xception (Deng et al., 2009; Simonyan & Zisserman, 2014; Szegedy et al., 2016; Chollet, 2017). We did not fine-tune the networks in any way and obtained the pre-trained weights from the Keras library (Chollet, 2015). Keras and Tensorflow were used as frameworks (Chollet, 2015; Abadi et al., 2015).

Some of the methods result in positive and negative evidence. We only considered positive evidence for the ImageNet tasks, so that methods could be compared. To ascertain, that this does not corrupt methods significantly, we compared the methods that do contain negative results against their filtered version. We found negligible to no difference between the two versions of a method in the metrics we are using. For aggregation, the explanation heatmaps are first processed by clipping relevance values at the 99th percentile to remedy outliers as done in (Smilkov et al., 2017). We then normalize all heatmaps to sum up to one. Intuitively this means that all methods assign the same total amount of relevance onto the input.

We compared the aggregation methods against saliency, guided backpropagation, SmoothGrad and Grad-CAM and guided Grad-CAM to have a selection of backpropagation-based methods. With the exception of guided Grad-CAM these methods were also used for the aggregation. Guided Grad-CAM was excluded from the aggregation, because it does not contain any information not in Grad-CAM or guided backpropagation and lead to a disproportionate weighing of the two.

7.1. Evaluating the evaluation

A good evaluation method should be able to reject the null hypothesis (a given meaningful explanation method is no better than random choice) with high confidence. To check whether our newly proposed evaluation method is accurate, we calculated paired t-test p-values of SmoothGrad on the Xception network against random choice for segment and pixel flipping. We chose SmoothGrad since it is an established explanation method where we can be sure, that it should be better than random noise.

For segment flipping, we set the 50% most relevant segments on a hundred images to the mean value over the dataset. For pixel flipping, we set the equivalent number of pixels to the mean value. We compared this to either setting the same number of random pixels or segments to the mean. If the difference between random choice and chosen between the explanation method is high, we are more confident that the explanation method reports meaningful information. Since we compare p-values for the same explanation method but different evaluation methods, the p-values contain information about how meaningful the evaluation method is.

The resulting p-values for this are reported in Table 1. For pixel-flipping we reported results on flipping relevant pixels, setting it to zero and setting it to the average. With the exception of setting the pixel to zero, all evaluation methods can reject the null hypothesis with $p < 0.05$. Segment flipping for the same explanation method can reject the null hypothesis, that the same explanation method does not contain any information with much higher confidence for the same number of total pixels. We can conclude that our evaluation method, segment flipping, is a better way to evaluate explanations for neural networks.

Table 1. t-test p-values of explanation methods

	T-STATISTIC	P-VALUES
PIXEL FLIPPING (MEAN-VALUE)	2.7	7.8 10E-1
PIXEL FLIPPING (BLACK)	0.3	1.4 10E-3
PIXEL FLIPPING (FLIP)	4.4	3.4 10E-5
SEGMENT FLIPPING	**6.4**	**3.5 10E-9**

7.2. Evaluation with Sensitivity-n on low-dimensional input

We use Sensitivity-*n* proposed by Ancona et al. (2018) to evaluate explanation methods on MNIST and FashionMNIST. Both are low-dimensional datasets with an image size of 28x28 pixels in black and white (Xiao et al., 2017; LeCun & Cortes, 2010). For each, we trained a basic convolutional neural network to 99% accuracy on MNIST and 92% accuracy on FashionMNIST respectively.

Since Sensitivity-*n* randomly samples subsets of pixels and does not take correlation between neighboring pixels into account, we only report results on low-dimensional images where correlation between neighboring pixels is low. 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. With only 784 pixels in the input, nearly all information in an image is lost at the end point.

Results are shown in Fig. 3. *AGG-Mean* and *AGG-Posterior* perform in range of the best methods. For MNIST, they are better than the best single methods. For FashionMNIST, they are within range of the best single method. The noticeably worse correlation for FashionMNIST is likely due to the fact that a lot of the images have highly relevant black pixels as part of the object, where setting the pixels to 0 does not change the image. When we inspected sample images, even deleting 50% or more of the pixels still left an easily distinguishable image, which was not the case with MNIST.

In alignment with the results in the original paper, correlation tapers off with increasing number of pixels in the set (Ancona et al., 2018). Averaging over all methods (*AGG-Mean*) as well as dividing by the standard deviation plus a prior (*AGG-Posterior*) perform on par with the best methods. Clipping the standard deviation if the mean is too low (*AGG-Clipped*) does not perform well.

7.3. Evaluation with segment flipping

To quantitatively compare the quality of the explanation methods, we use the procedure introduced in Section 6 on ImageNet. In Table 2 we report the final average decay after 50% of the image have been grayed out. At 50% the image still contains enough information that classification does not become spurious and dependent on random details. We note that SmoothGrad, another method that averages over multiple decisions by applying noise to the network input, performs better than the vanilla version, Saliency.

For all pretrained networks considered, averaging over explanation methods (*AGG-Mean*) and dividing by the regularized standard deviation (*AGG-Posterior*) results in lower class scores than any of the vanilla methods. Since this means, that more relevant segments are correctly identified, we can deduce that aggregating explanation methods represents a valid way of improving explanations for neural networks compared to using a single explanation method. In Fig. 4 we show the decline of the class score with the increasing number of grayed out segments. We observe that all explanation methods perform significantly better than randomly choosing segments to gray out.

Clipping the standard deviation before normalizing (*AGG-Clipped*) does not produce robust results. On inspection of the produced heatmaps we observed, that clipping led to a collapse of the heatmap where most of the relevance is focused on few pixels as seen in Fig. 8. This is likely the reason for *AGG-Clipped* performing worse.

We note that the final class scores for VGG19 are higher for any of the given methods, indicating that identifying relevant areas is more difficult for this neural network.

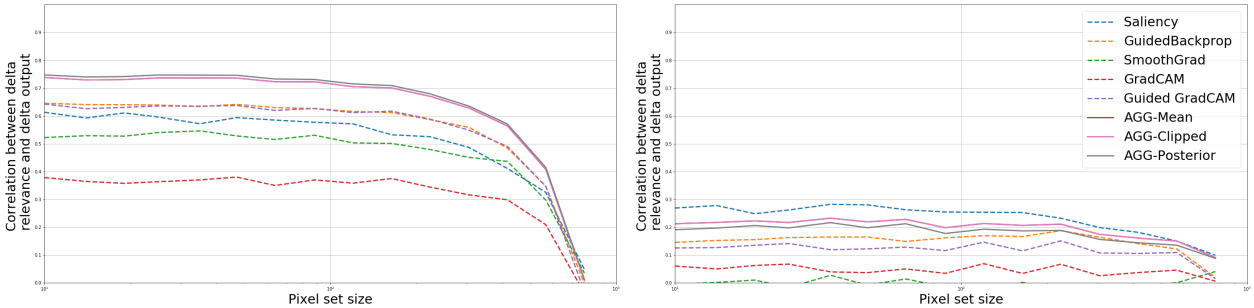

Figure 3. Quantitative evaluation: Sensitivity-*n* for MNIST (left) and FashionMNIST (right). *AGG-Mean* is covered by *AGG-Posterior*. Legend is for both plots.

Table 2. Remaining class score after 50% of image segments grayed out. Lower is better.

	VGG19	XCEPTION	INCEPTION
SALIENCY	0.14 ± 0.01	0.39 ± 0.02	0.25 ± 0.01
GUIDED BACKPROP	0.00 ± 0.00	0.35 ± 0.02	0.20 ± 0.01
SMOOTHGRAD	0.13 ± 0.01	0.35 ± 0.02	0.19 ± 0.01
GRAD-CAM	0.09 ± 0.00	0.35 ± 0.01	0.22 ± 0.01
GUIDEDGRAD-CAM	0.09 ± 0.00	0.35 ± 0.01	0.20 ± 0.01
AGG-MEAN	$\mathbf{0.08 \pm 0.00}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.14 \pm 0.01}$
AGG-POSTERIOR	$\mathbf{0.08 \pm 0.00}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.14 \pm 0.01}$
AGG-CLIPPED	0.14 ± 0.01	0.45 ± 0.02	0.27 ± 0.01

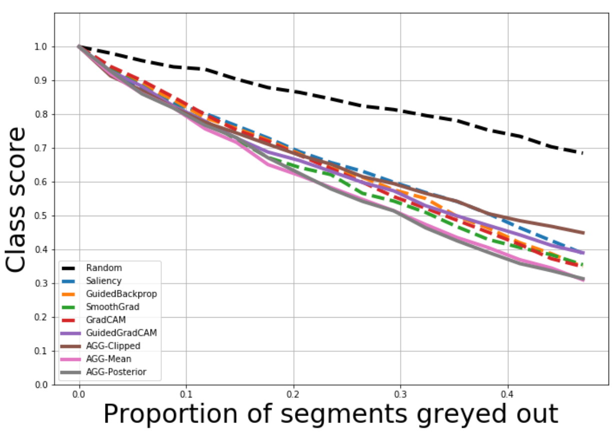

Figure 4. Quantitative evaluation: Decay of class scores with segment flipping for pretrained VGG19.

7.4. Evaluation with ML Interpretability Benchmark

We use the benchmark introduced in (Mohseni & Ragan, 2018) to test the alignment between human evaluation and explanation methods. The benchmark consists of a hundred images of twenty-five different categories. Since ’man’ and ’woman’ are not present as categories in the traditional ImageNet task, we filter out these categories and evaluate on ninety images. For each image, the benchmark provides a heatmap of segments that ten human subjects considered important. We show example images of the dataset in Fig. 5. For reasons previously discussed in Section 2, human evaluation does not necessarily provide a ground truth in explaining decisions by neural networks. However, it is reasonable to assume that there should be a correlation between relevance assigned by human and machine. Therefore we also evaluate the alignment of the methods with human judgement.

The cosine similarity is reported for three pre-trained networks on Imagenet in Fig. 6. Saliency, Integrated Gradients and Guided Backprop perform significantly worse. Our proposed aggregation, averaging over methods (*AGG-Mean*) and using a prior while incorporating variance (*AGG-Posterior*) perform on-par with the best methods. Additionally, in Table 3 we reported the Jaccard index between human and explanation methods for VGG19. For humans, we marked areas selected by at least two human reviewers as relevant. For the explanation methods, we marked areas that had at least the average relevance over the whole picture as relevant. The results show that Grad-CAM, AGG-Posterior and AGG-Mean perform best, indicating high alignment between human and machine.

7.5. Qualitative visual evaluation

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 intu-

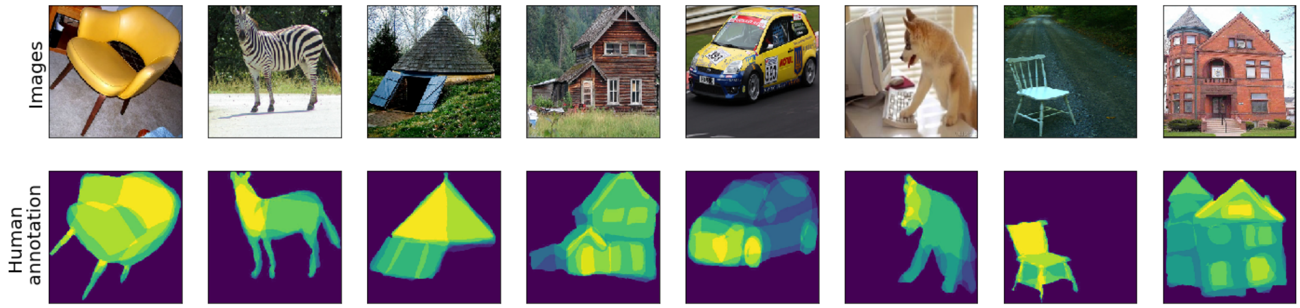

Figure 5. Example images and human-annotated heatmaps from (Mohseni & Ragan, 2018)

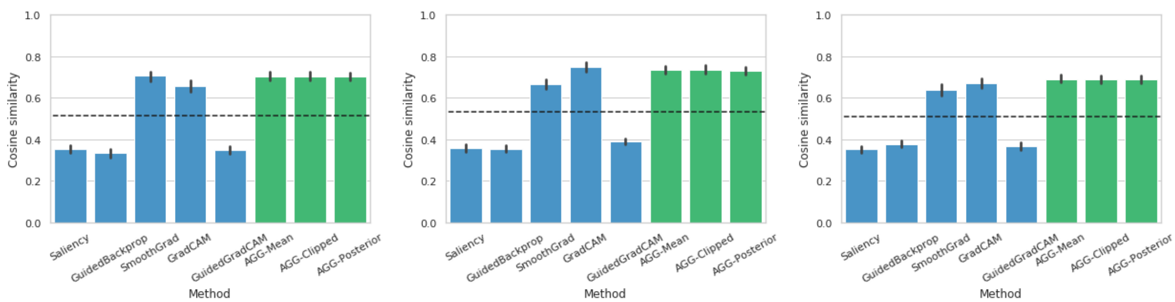

Figure 6. Averaged cosine similarity between human-assigned relevance and explanation methods reported on Inception(left), Xception (middle) and VGG19 (right). Aggregated methods in green. Dashed line is the average over all methods.

Table 3. Jaccard index between heatmap and human annotated benchmark on VGG19.

	JACCARD
SALIENCY	0.20 ± 0.01
GUIDEDBACKPROP	0.19 ± 0.01
SMOOTHGRAD	0.27 ± 0.01
GUIDED GRAD-CAM	0.17 ± 0.01
GRAD-CAM	$\mathbf{0.31} \pm 0.01$
AGG-MEAN	$\mathbf{0.32} \pm 0.01$
AGG-POSTERIOR	$\mathbf{0.32} \pm 0.01$
AGG-CLIPPED	0.30 ± 0.01

itive human understanding (Adebayo et al., 2018). Therefore we also examined the heatmaps manually for individual images, some of which can be seen in Fig. 7.

We see that *AGG-Mean* and *AGG-Posterior* combine features of the used 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. As such, it is a combination of the detail-oriented and shape-oriented methods. *AGG-Posterior* distributes relevance onto more area compared to *AGG-Mean*. For the class *chain mail* in the lower row, there is disagreement between methods on whether the collar is relevant or not. This is reflected in the explanation map for *AGG-Posterior* where more relevance is attributed to the actual chain mail compared to *AGG-Mean*. We can visually confirm that combining explainability methods provides a meaningful improvement over single methods.

7.6. Discussion

We first analyzed whether our newly proposed evaluation method gives meaningful results. On the same explanation method, segment flipping gives better results than the alternative, pixel flipping.

We proceeded to evaluate aggregated methods against unaggregated methods in multiple metrics, measuring alignment with the network that is being explained as well as human judgment. We reported results for three different metrics. For all of them, aggregated methods perform at least as good as the best unaggregated method. It is worth noting, that the best unaggregated method varies, dependent on the metric. As an example, GradCAM performs on top for the human benchmark but worst for Sensitivity-n. We can summarize, that *AGG-Mean* and *AGG-Posterior* always performed at least as good as the best method going into the aggregation. This is very interesting, since it indicates that aggregating methods only ever leads to a net benefit in reduction of the bias. Looking at visualizations of the heatmaps, we have observed that the aggregated heatmaps are a compromise

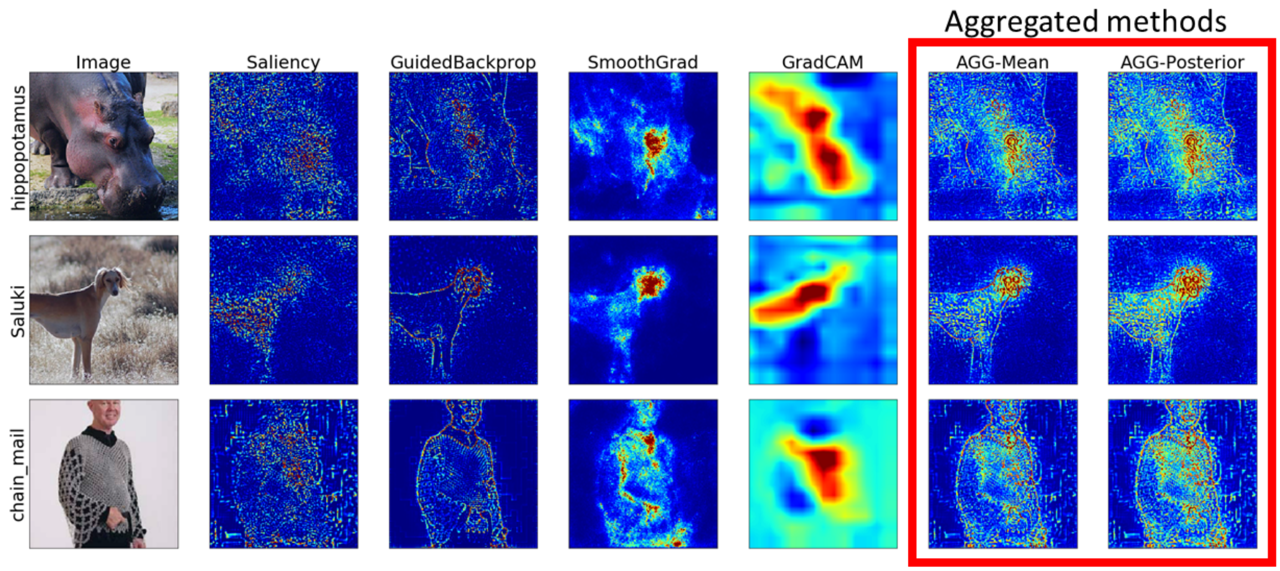

Figure 7. Example images from Imagenet and the heatmaps produced by different methods on VGG19. Aggregated methods give more complete explanations.

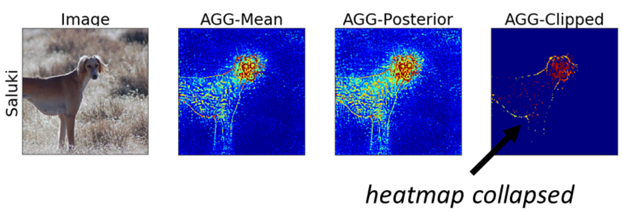

Figure 8. Imagenet image and the heatmaps produced by aggregated methods. Clipping the standard deviation collapses the heatmap onto very few pixels.

between relevance focused only on few areas and a very uniform distribution of relevance onto the object being classified. *AGG-Mean* and *AGG-Posterior* perform equivalently in the quantifiable metrics. Looking at the visual examples however, *AGG-Posterior* distributes relevance more evenly onto the entire object. As such, they are both as aligned with human judgment as the most aligned single method, while reflecting the neural network decision at least as well as the best methods.

Across all metrics, *AGG-Clipped* does not perform as well as the two other variants and in fact performs worse than the average of all methods. We suspect that this is because clipping the standard deviation leads to all of the relevance concentrated on very few pixels as shown in Fig. 8.

8. Conclusion

The explanation problem for object detection using neural networks is fundamentally ill-posed: Many schemes have been proposed leaving considerable epistemic uncertainty. We propose to mitigate this by aggregating multiple explanation methods. This serves multiple purposes, reducing variance representing the epistemic uncertainty and reducing bias by extending the range of features that an explanation can highlight.

We gave a simple proof that aggregating explanation methods will perform at least as good as the typical individual method. In practice, however, we found evidence that aggregating methods will perform as well as *a*ny single method. We found this evidence substantiated across three quantitative metrics. In addition, we examined examples manually and found further visual evidence. Thus aggregated explanation methods will be a viable alternative to choosing a single method in the future.

Additionally we proposed a novel way of evaluation for explanation methods that circumvents the problem of high correlation between pixels and does not rely on visual inspection by humans, an inherently misleading metric. Having a viable and scalable way of evaluating explanation methods will speed up research into explainability, as methods become more comparable.

Future work will include a better categorization of the currently available explanation methods and improved aggregation methods based on these characterizations.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

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.

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.

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.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327. IEEE, 2017.

Chollet, F. Keras, 2015. URL https://keras.io.

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.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of Neural Networks is Fragile. oct 2017.

Hansen, L. K., Nielsen, F. Å., Strother, S. C., and Lange, N. Consensus inference in neuroimaging. *NeuroImage*, 13 (6):1212–1218, 2001.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. Evaluating feature importance estimates. *arXiv preprint arXiv:1806.10758*, 2018.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

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.

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.

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., Vigas, 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.

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. 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, 2018.

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.