BENCHMARKING THE ATTRIBUTION QUALITY OF VISION MODELS

Alicja Zabicka az4g24@soton.ac.uk
Piotr Cieslik pc1u24@soton.ac.uk
Daniel Burgess djb1g20@soton.ac.uk

1 INTRODUCTION

1.1 RESEARCH BACKGROUND

As attribution methods continue to be well established as interpretability tools for vision models, the area is heavily researched from different perspectives. The paper for which we reproduced experiments (Hesse et al., 2024), proposes a new evaluation methodology for attribution measures and demonstrates its use on 23 attribution methods. They also share insights into the impact of design choices on the quality of the attribution.

The authors decided to build on top of incremental-deletion score (IDS) (Samek et al., 2017)[1], which measures how incrementally removing patches of an image (replacing them with random noise, local blurring or other interventions), starting from those with the highest attribution score, impacts the model's output. IDS is in turn an extension of the pixel flipping protocol (Lapuschkin et al., 2015) which uses the idea that interventions on the pixels with high importance should have bigger impact on the model's output than those made on pixels with lower importance.

They criticise IDS for having an issue of class information leakage; a phenomenon where the location of the removed patch in the image can give away some hints about the correct class. They also point out two main flaws of IDS:

- The *out of domain issue*, a problem that arises from using attribution methods on images with removed data, i.e. images with white patches, local blurs or other artificially created modifications does not belong to the domain the model was trained on, therefore, we can no longer rely on attribution methods.
- IDS does not support straightforward inter-model comparisons. They point out this limitation, which does not allow to use IDS as a tool for assessing the impact of model design choices on attribution quality. The reason why this limitation exists in IDS is that it is dependent directly on the model output, therefore any change in the model, even calibration, will lead to a different IDS score for the same attribution method. That makes comparison of attribution quality between models impossible.

They created a new protocol for evaluating attribution methods, *in-domain single-deletion score* (IDSDS), which, they claim, solves the mentioned limitations of IDS.

1.2 HOW DOES IDSDS WORK

IDSDS works by deleting patches from images and observing the decrease in logits for the target class. They treat the decrease in the logit for the target class as a "ground truth for importance" of the area covered by the patch. The final IDSDS value is computed as a Spearman rank-order correlation between the logit drops and the sums of attribution maps for the respective patches. This way, they measure whether the attribution is a monotonic function of what the authors treat as important.

1.3 PAPER CLAIMS

The authors claim that fine-tuning the models with the images with deleted patches *solves* the out-of-domain (OOD) issue. Although it is clear that such fine-tuning *mitigates* the OOD issue, the claim that it is entirely solved, strikes us as overconfident.

They also say that the train and test domains are *exactly* aligned, which is only true if the previous point holds.

Another claim is that their protocol solves the class information leakage. We can expect that the class information leakage is indeed solved, as the position of the deleted patches does not depend on the attribution maps produced for a particular image, and therefore does not give any hints about the target class.

Their next claim is that their protocol allows for inter-model comparisons. It is clear that the fundamental cause of IDS not allowing for inter-model comparisons – dependence only on the model's output – has been overcome, although they make no effort into giving any arguments as to why the comparison of IDSDS scores for different models gives meaningful results. For instance, different models might have different logit drops in response to equally important information. That claim is not justified enough; however, it would be rather difficult to disprove it. A claim connected to that, is that the ranking of the attribution methods remains stable across different models.

They also make a series of claims about how the design choices influence attribution quality, for instance, that smaller depth or switching off of Batch Normalisation (BN) layers improves the quality of attribution maps. In order to produce those claims, they use the assumption that different models can be compared with their protocol, which, as we have seen, they do not give enough justification for.

1.4 WEAKNESSES OF IDSDS

These are the weaknesses that we have identified:

- **Using deletion / modifications of the input** One of the main weaknesses, in our opinion, of IDSDS is the same as any deletion-based protocol: deleting parts of the image can introduce information that the model might interpret as a different class (a truck being a well-known example).
- **Out-of-domain** We like that they tried to mitigate the OOD issue by fine-tuning the models on deleted patches before evaluating, although we see their claim that the images with deleted patches are in-domain (ID) after fine-tuning, far fetched and not verified.
- **Change of domain** Fine-tuning models introduces changes to the domain (even if they're not enough to consider images with deleted patches in-domain), therefore it means that attribution maps created by the attribution methods will no longer be the same, as if used with the original domain. That means that we are no longer measuring the quality of the attribution methods but something slightly different. Changing the domain, even slightly, affects the models and, therefore, is likely to change the attribution maps.
- **Implicit assumption that logits drop is the "ground truth for importance"** They say that target class logit drop for an image with a deleted patch is the ground truth of the importance of the data in that patch. They don't give any justification for that. Even though it seems reasonable to say that there is a connection between the logit drop and importance of the information in the image, just assuming that it is the ground truth doesn't seem justified at all, and should that assumption be wrong, that would mean that the entire method is flawed.

2 OUR EXPERIMENTS

All the code for our experiments is located in our repository, `https://github.com/piotrek-cieslik/reproducibility-challenge-benchmarking-attribution-methods`, which we will transfer to `COMP6258-Reproducibility-Challenge` organisation.

2.1 EVALUATION OF ATTRIBUTION METHODS WITH BASELINE COMPARISONS

To evaluate the robustness of the IDSDS protocol, we first reproduced the original evaluation of attribution methods on all selected models. In addition to replicating the original setup, we introduced several *baseline* or *dummy* attribution methods. The rationale for this inclusion is that any robust evaluation

[1]Interestingly, even though (Hesse et al., 2024) cites (Samek et al., 2017) as the source of IDS and that (Samek et al., 2017) describes that score, they do not use name "incremental-deletion score" to refer to their work.

protocol should consistently assign lower scores to these simplistic or semantically ungrounded methods compared to well-established attribution techniques such as GradCAM and Integrated Gradients (IG).

The dummy methods we introduced range from entirely random maps to heuristics loosely related to image semantics. While these methods may, in specific cases, correlate with salient image regions, they are not designed to provide faithful explanations and therefore should not outperform principled attribution techniques under a reliable evaluation protocol.

The baseline attribution methods we evaluated are as follows:

- **Centered Gaussian Blur**: We generated attribution maps using a Gaussian distribution centered in the middle of the image, with 20 different standard deviations, ranging from 1 to 256. This heuristic is based on the observation that objects of interest often appear near the image center. However, if such a method systematically outperforms GradCAM or IG, it would strongly suggest a flaw in the IDSDS protocol.
- **Entropy Proxy**: As a proxy for local entropy, we applied average pooling to squared local deviations from the mean pixel intensity, using seven different kernel sizes. While this does not compute true entropy (which would require probability distributions and logarithmic terms), it provides a rough approximation of local image complexity. The underlying hypothesis is that high-variation regions may correlate with semantic importance.
- **Random Maps** We employed three purely random attribution maps; uniform random values, squared uniform random values and uniform random values offset by 1. These variations allow for minor structural differences in random baselines (relative difference of pixel values), offering more nuanced comparisons.
- **Classical Edge Detectors**: We included four standard computer vision edge detection methods – image gradient (based on pixel intensity differences), Sobel operator, Canny edge detector, and Marr-Hildreth edge detector. These methods do not involve backpropagated gradients and are unrelated to gradient-based saliency methods. Their inclusion serves to test whether edge information alone is disproportionately rewarded by the protocol.
- **Frequency-Based Methods**: We included three attribution maps derived from frequency filtering; high-pass filter (preserving high frequencies), low-pass filter (preserving low frequencies) and band-pass filter (preserving mid-range frequencies). These were motivated by similar considerations as the entropy-based method—namely, that structural or textural variation may loosely correlate with regions of interest.

Although some of these heuristics may occasionally align with meaningful image regions, none are grounded in the model's internal computations or class-specific relevance. If the IDSDS protocol consistently ranks any of these dummy methods above established techniques, it would indicate a fundamental reliability issue in the evaluation framework.

2.2 Consistency of Attribution Method Rankings: Logits vs. Softmax Probabilities

When evaluating the effect of image perturbations on model outputs, attribution protocols may use either the raw logits or the softmax probabilities. It is important to note that a perturbation can decrease a model's logit for the correct class while simultaneously increasing the corresponding softmax probability, due to the normalization effect of the softmax function across all classes.

While logits reflect the raw output of the model, softmax probabilities offer a normalized and bounded measure that is more interpretable and comparable across models and inputs. For this reason, we consider softmax probabilities to be the more reliable choice for quantifying class prediction changes in evaluation protocols.

In the original study, the authors reported that the relative ranking of attribution methods remained largely unchanged whether logits or softmax probabilities were used. However, this analysis was conducted only on a single architecture—VGG-16—which limits the generalizability of the conclusion.

To address this limitation, we extended the comparison across all models included in our experiments. Our aim was to verify whether the attribution method rankings remain consistent across different architectures when switching from logits to softmax probabilities. If the rankings are indeed stable, this would support the use of softmax probabilities in future evaluations without compromising the validity of comparative assessments among attribution techniques.

In this experiment we didn't use Integrated Gradients, because of the large computational cost, as outlined in section 4.

2.3 Evaluation on a Different Dataset and Diverse Architectures

The original study evaluates the IDSDS protocol exclusively on the ImageNet (Deng et al., 2009) dataset. While ImageNet is a widely-used and well-established benchmark in computer vision, relying on a single dataset – regardless of its scale – provides limited evidence for the robustness and general applicability of an evaluation protocol for attribution methods.

To assess the generalizability of the IDSDS protocol beyond ImageNet, we conducted experiments using CIFAR-100 (Krizhevsky, 2012), a standard dataset that differs substantially in both image resolution and content distribution. Our evaluation spanned eight pretrained architectures: ResNet20, ResNet32, ResNet44, ResNet56, as well as four batch-normalized VGG variants (VGG11-BN, VGG13-BN, VGG16-BN, VGG19-BN). The models come from `https://github.com/chenyaofo/pytorch-cifar-models`.

Following the original methodology, we fine-tuned all models prior to evaluation. To accommodate the smaller image size and the different nature of CIFAR-100, we adjusted the training configuration accordingly. Specifically, we trained each model for 30 epochs using mini-batch gradient descent with weight decay of 0.0005, momentum of 0.9, batch size of 128, initial learning rate of 0.005, and a step size of 15 for the learning rate scheduler. For each architecture, we selected the checkpoint with the highest validation accuracy to avoid concerns related to overfitting.

During implementation, we encountered a number of practical limitations in the original codebase. Notably, the authors' implementation hardcoded ImageNet-specific assumptions – such as the 224×224 input resolution – in several parts of the pipeline. This lack of generalization led to erroneous outputs and required considerable manual intervention and debugging to adapt the protocol to different datasets.

In that experiment, we also repeated a sanity check from the original paper, comparing top 1 and top 5 accuracies between the pretrained and finetuned models. We would like to note, that this sanity check is not sufficient to draw any conclusions about network similarity, but we performed it to make sure that our choice of learning parameters does not result in differences in network performances, which would go against the methodology proposed in the original study. For lack of space we can not include the results here, but they are available in our repository in file `experiments/train_and_evaluate_cifar/results_cifar_sanity_check.csv`. The difference is up to 1 percentage point.

2.4 Testing the Claimed Network Similarity

The authors claim that fine tuning does not substantially affect the models' behaviour. In the first section on accuracy, they show that the accuracy is not significantly impacted on uncorrupted inputs and is increased on corrupted ones. A detailed table is provided, outlying the pre-fine-tuned accuracy and post-fine-tuned accuracy for a number of models.

To test that the models' behaviour hasn't changed substantially, the authors perform three experiments. The first one measures the Mean Absolute Difference (MAD) between the original models and the fine-tuned models' target softmax output, positing that a model using different features will result in different output confidences. Therefore a smaller value indicates, according to them, similarity. Unlike the previous section, however, they only provide two examples of a small MAD in the main text of the paper, with no other examples present elsewhere. To validate this claim, we measure the MAD on a number of different models to verify that it holds under a variety of network architectures.

The second experiment involved measuring the MAD between the attribution maps of GradCam. The authors justified its use due to its relative simplicity and low noise compared to other methods. As IDSDS is used to evaluate a broad suite of attribution methods, this again raises some concerns about the lack

of rigour by only testing the similarity of one method. Once again, to test this, we calculated the MAD for a variety of different models using a number of attribution methods.

In the third experiment, the authors randomly select channels from the last convolution layer of the models, and compare the highest activating images between the original model and the fine tuned one. This visualises the "concepts" learned by the channel, and should be similar. Again, only two models are used here. For lack of time, we have not performed this experiment.

3 METHODOLOGY

In our experiments, we evaluated the majority of the attribution methods employed in the original study across most of the models. However, we omitted certain methods to ensure consistent comparison across all models. In the original paper, some attribution methods were not compatible with specific models, which would have hindered direct comparability. Below, we outline the reasons for these exclusions.

- BagNet-33 and Bcos-ResNet-50 do not work with GradCAM family in the code provided
- ViT-B-16 is a vision transformer and lots of the attribution methods covered in the paper works for CNN only
- RISE and RISE-U had missing files in the implementation. They hardcoded file paths in the code and did not provide them for download
- Rollout and CheferLRP works for ViT only
- Bcos attribution method works only for Bcos networks
- BagNet attribution method works only fo BagNet networks

When evaluating, we used the standard validation splits of datasets we were using. For ImageNet, it was 3923 images, for CIFAR-100, it was 10,000 images.

We analysed the results mostly from the perspective of the claims found in the papers 1.3. We also took into account our initial motivations described in section 2. Our conclusions are described in the section .

4 IMPLEMENTATION

Where appropriate, we used the code from the original study repository, `https://github.com/visinf/idsds`. We also included it as a submodule in our repository. Experiments 2.1, 2.2 and 2.3 built on top of the original code, although adding significant amount of effort beyond what was provided. In particular, all the baseline attribution methods are implemented from scratch and integrated into the original code, evaluation and training files are largely rewritten for the purposes of our experiments. For the CIFAR-100 experiment, all ImageNet-specific areas in the code are updated to reflect CIFAR-100 characteristics. For CIFAR-100 models, we selected appropriate target layers for GradCAM. Interestingly, the IDSDS repo used `gradcam_target_layer = 'model.features'` instead of one specific layer.

Code for experiment described in section 2.4 was written from scratch and only the model classes from the original repo were used. This is because they did not provide code for these checks.

As we struggled with Out of Memory exceptions, we decided to fix a known issue of Pytorch being reluctant to release memory. Even with manually deleting variables, clearing Pytorch cache, invoking Pytorch and Python garbage collectors, we could see that the memory was still in use. This could be because references to Pytorch objects might be stored not just in the variables we have access to but also elsewhere, for instance in Pytorch own datastructures or sometimes even in logging libraries. That makes garbage collectors ineffective. Therefore, we decided to implement a solution that encapsulated all model and data loaders creation, training and evaluation within a separate process and were removing those processes once they were finished. That finally proved effective and we could see that the memory was being released. This solution was implemented entirely from scratch.

Some of the attribution methods and model combinations were, even with our solution, so memory demanding, that we could only use batch size of 1 as using batch size of 2 were already causing out of memory exceptions on university GPUs of 48GB or RAM. That was the case with, for instance, Integrated Gradients with SmoothGrad on ResNet-152. Using batch size of 1, made the Integrated Gradients extremely slow to evaluate as it needed around 1 second to evaluate one image. Evaluating 13 models on 7 Integrated Gradients variants, each needing to evaluate 3923 images, totalled to over 100 hours.

5 RESULTS

5.1 EVALUATION OF ATTRIBUTION METHODS WITH BASELINE COMPARISONS

Most of our baseline methods score around zero; the average score is 0.017. However, two of them – Edge-Sobel and Edge-Gradient – consistently score above 0.1, with maximum score 0.1575. They score well on VGG models, but poorly on ResNet models. Notably, both of those methods consistenly outperform Integrated Gradients (zero baseline) + SmoothGrad on VGG models. Entropy baseline methods score above 0.1 on some VGG models, not outperforming Integrated Gradients, but coming close. That results raises questions about the reliability of IDSDS protocol.

Other than that, our experiments do not contradict the claims from the paper, in particular, that simpler models tend to have better attribution quality and that the relative ordering of the attribution methods stays similar across different models.

The evaluations of all the attribution methods are shown on the figure 1a. Because of an extremely limited space, we can only include few visualisations. For more plots, including those looking at the results from different perspectives and more readable, please see our repository.

5.2 CONSISTENCY OF ATTRIBUTION METHOD RANKINGS: LOGITS VS. SOFTMAX PROBABILITIES

We performed logits vs softmax experiment for all of 13 models we are evaluating for ImageNet. For lack of space, we include two in figure 1c, the rest is available in our repository. Most of the time, two or three attribution methods change their place in the ranking, when switching from logits to softmax probabilities. That makes the decision of using logits even more debatable than we previously thought.

We also found that their claim "computing the attributions after the final softmax layer reduces the correctness as measured by our IDSDS for almost all methods, indicating that pre-softmax attributions are favorable" is not true. For ResNet-50, -101 and -152, the scores tend to increase, for other members of ResNet family, they stay the same and they only decrese for the VGG models. In addition to that, the results we got for ResNet-50 are very different from those in the paper (please see figure 4c in the original paper).

5.3 EVALUATION ON A DIFFERENT DATASET AND DIVERSE ARCHITECTURES

This experiment supported their claim that attribution methods quality decreases with increasing complexity of the model. We can see this on both VGG and ResNet families (figure 1b). We can see, however, a number of peculiar phenomena, which undermine other claims.

The most striking is that scores for the entire GradCAM family completely collapsed while going from ResNet to VGG, for VGG-19 BN being around zero. Notably, on ImageNet, the GradCAM family was the best performing family on VGG models, among the ones we tested. Even more surprisingly, GradCAM family performance on ResNet models improved in relation to ImageNet. This is a strong indication that IDSDS protocol is not as reliable as the authors claimed and shows the importance of not relying on just one dataset.

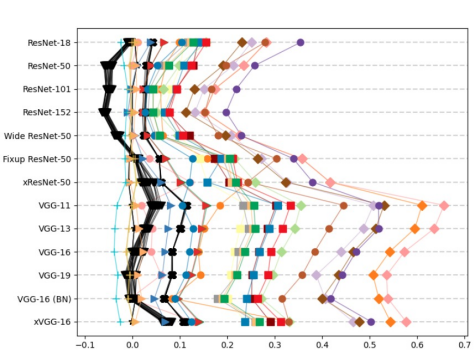
The relative ranking of the attribution methods stays similar when switching between datasets, however, only within one category (GradCAM-based, IG-based, and simpler gradient-based) – when we compare attribution methods from different families, we observe larger differences in the rankings. That also points to questioning the claim that IDSDS is reliable for comparing different attribution methods.

The score of *Integrated Gradients (zero baseline) + SmoothGrad* is around zero on all models when using CIFAR-100, even though it was performing much better on VGG models on ImageNet, peaking at 0.128 for VGG-11.

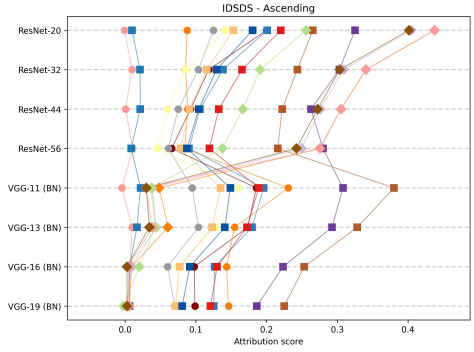
These observations are only the most standing out examples, there are more examples of differences and unexpected behaviour, for instance that scores for all attribution methods tend to increase for ResNet and decrease for VGG models while switching from ImageNet to CIFAR-100. We can not describe them all, but our repository contains more visualisations, in particular showing the attribution methods per family (IG, GradCAM, IxG) making the plots more readable. We expected that the scores should be more stable across datasets, not just models – it is hard to justify such differences given that both the datasets contains images of everyday objects – the differences would be more expected if we compared two different, niche datasets.

5.4 Testing the Claimed Network Similarity

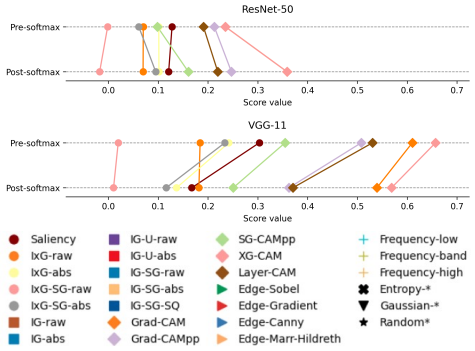
These experiments had to be implemented from scratch as explained above, which resulted in some confusion about the produced values. For the first sub-experiment testing this claim, the values produced by calculating the MAD of the target softmax outputs were not close to the limited examples given in the paper, being smaller by around a factor of 100. Despite this, the results do mostly support their claims with the ResNet family having a MAD between 0.00016-0.00020 and VGG being very consistent with 0.00011-0.00012. We further measured the MAD between logit outputs as we reasoned that it would give a better internal representation for raw values: ResNet varying from 0.41-0.44 and VGG from 0.25-0.28. Curiously these values align more with the paper's, being off by a factor of 10 and having similar most significant digits e.g. VGG 16 in the paper had a score of 0.028 and we found a logit score of 0.28. This may indicate that they authors were in fact using the logit MAD, however, this could be coincidental and, as they did not provide an implementation, we cannot know for certain. For the second sub-experiment, we measured the MAD of attribution maps between models on a few different methods. With GradCam, the method used in the paper, the values aligned much better, as with ResNet-50, we found 0.054 and the paper stated 0.047. For the dissimilar model comparison we found 0.201 and the paper stated 0.193. Both ResNet and VGG families did not vary much, staying around 0.04-0.06 for all methods tested: GradCam, SmoothGradCAM++ and ScoreCAM. All results can be found in the repository under `experiments/network-similarity` in their respective csv files.



(a) Evaluation results on ImageNet for experiment 2.1



(b) Evaluation results for CIFAR-100 for experiment 2.3



(c) Top: Pre- vs post-softmax for Resnet-50, Middle: for VGG-11, Bottom: Legend

Figure 1: Visualisations of the results for our experiments

6 Team Contribution

Alicja Zabicka Wrote the Reproducibility Challenge Plan. Wrote the code for experiment 2.2. Wrote sections 2.2 and 5.2. Wrote all the code for creating the visualisations.

Piotr Cieslik Wrote code for experiments 2.1 and 2.3, including the multiprocess management. Wrote the report, except sections 2.2, 2.4, 5.2 and 5.4.

Daniel Burgess Wrote three sub-experiments for experiment 2.4 and performed two of them. Wrote sections 2.4 and 5.4.

7 Conclusions

In this report we described what weaknesses (Hesse et al., 2024) has and our approach for reproducing it, including the experiments we designed to test the paper claims. In addition to reproducing the results they achieved, we designed novel experiments to verify the validity of their claims and the reliability of the IDSDS protocol. While verifying their claims, we found a number of arguments casting doubts at the validity of the evaluation protocol they introduce.

First, edge detection or entropy-based dummy methods can compete with Integrated Gradients on VGG models when evaluated on the entire validation split of ImageNet. Such result should not happen with a reliable evaluation protocol, especially when averaged on almost four thousand images.

Attribution methods ranking changes, although not significantly, when switching from logits to softmax probabilities. That casts further question on their choice of using logits, especially that they treat logits drop as "the ground truth for importance". In addition to that, we found that their claim of IDSDS scores decreasing when switching to softmax is not true.

While using IDSDS on CIFAR-100, we noticed differences in ranking of the attribution methods, compared to ImageNet, which shows that IDSDS is not robust for performing comparisons across different datasets. More notably, the scores for entire GradCAM family collapses when switching from ResNet to VGG architectures. This result significantly undermines the validity of the IDSDS protocol.

We found some evidence supporting their claims, for instance decreasing IDSDS scores when switching to more complex models, or low MAD difference between softmax of pretrained and finetuned models, as well as low MAD between attribution method outputs.

Overall, we conclude that even though there is merit in this method and we like that they made effort to overcome IDS limitations, the IDSDS method is not as reliable as the authors claim.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Benchmarking the attribution quality of vision models, 2024. URL `https://arxiv.org/abs/2407.11910`.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10:e0130140, 07 2015. doi: 10.1371/journal.pone.0130140.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820.