TOWARDS BEST PRACTICE IN EXPLAINING NEURAL NETWORK DECISIONS WITH LRP

Maximilian Kohlbrenner[1], *Alexander Bauer*[2], *Shinichi Nakajima*[2]
Alexander Binder[3], *Wojciech Samek*[1], *Sebastian Lapuschkin*[1]

[1]Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[2]Department of Computer Science, Technische Universität Berlin, Berlin, Germany
[3]ISTD Pillar, Singapore University of Technology and Design, Singapore, Singapore

ABSTRACT

Within the last decade, neural network based predictors have demonstrated impressive – and at times super-human – capabilities. This performance is often paid for with an intransparent prediction process and thus has sparked numerous contributions in the novel field of *explainable artificial intelligence (XAI)*. In this paper, we focus on a popular and widely used method of XAI, the *Layer-wise Relevance Propagation (LRP)*. Since its initial proposition LRP has evolved as a method, and a *best practice* for applying the method has tacitly emerged, based on humanly observed evidence. We investigate – and for the first time *quantify* – the effect of this current best practice on feedforward neural networks in a visual object detection setting. The results verify that the current, layer-dependent approach to LRP applied in recent literature better represents the model's reasoning, and at the same time increases the object localization and class discriminativity of LRP.

Index Terms— layer-wise relevance propagation, explainable artificial intelligence, neural networks, visual object recognition, quantitative evaluation

1. INTRODUCTION

In recent years, deep neural networks (*DNN*) have become the state of the art method in many different fields but are mainly applied as black-box predictors. Since understanding the decisions of artificial intelligence systems is crucial in numerous scenarios and partially demanded by law[1], neural network interpretability has been established as an important and active research area. Consequently, many approaches to explaining neural network decisions have been proposed in recent years, *e.g.* [1, 2, 3, 4]. The *Layer-wise Relevance Propagation* (*LRP*) [5] framework has proven successful at providing a meaningful intuition and measurable quantities describing a network's feature processing and decision making [6, 7, 8]. LRP attributes *relevance scores* R_i to the model inputs or intermediate neurons i by decomposing a model output of interest. The method follows the principles of *relevance conservation* and *proportional decomposition*. Therefore, attributions computed with LRP maintain a strong connection to the predictor output. While early applications of LRP use a single decomposition rule uniformly to all layers of a model [5, 9, 10] more recent work describes a trend towards assigning specific decomposition rules purposedly to layers wrt. function and position within the network [11, 8, 12, 13, 14]. This trend has tacitly emerged and formulates a *best practice* for applying LRP. Under qualitative evaluation, the attribution maps resulting from this current approach seem to be more robust against the effects of shattered gradients [10, 15, 11] and demonstrate an increased discriminativity between target classes [11, 12] compared to the uniform application of a single rule.

However, recent literature applying LRP-rules in a layer-dependent manner do not justify the beneficial effects of this novel variant quantitatively, but only based on human observation. In this paper, we design and conduct a series of experiments in order to verify whether a layer-specific application of different decomposition rules actually constitutes an improvement above earlier descriptions and applications of LRP [9, 16]. Our experiments are conducted on popular computer vision data sets with ground truth object localizations, the ImageNet [17] and PascalVOC [18] datasets, using different neural network models.

2. FEEDFORWARD NEURAL NETWORKS AND LRP

Feedforward neural networks constitute a popular architecture type, ranging from simple multi-layer perceptrons and shallower convolutional architectures to deeper and more complex Inception [19] and VGG-like architectures [20]. These types of neural network commonly use ReLU non-linearities and first pass information through a stack of convolution an pooling layers, followed by several fully connected layers. The good performance of feedforward architectures in numerous problem domains, and the availability as pre-trained models makes them a valuable standard architecture in neural network design.

2.1. Layer-wise Relevance Propagation

Consequently, feedforward networks have been subject to investigations in countless contributions towards neural network interpretability, including applications of LRP [5, 9, 16], which finds its mathematical foundation in *Deep Taylor Decomposition (DTD)* [21]. The most basic attribution rule of LRP (we here refer to as *LRP*$_z$) is defined as

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{z_j} R_j^{(l+1)} \tag{1}$$

and performs a proportional decomposition of an upper layer $(l+1)$ relevance value $R_j^{(l+1)}$ to obtain lower layer (l) relevance scores $R_i^{(l)}$ wrt. to the forward mappings z_{ij} (directed from layer inputs i to

This work was supported by the German Ministry for Education and Research as Berlin Big Data Centre (01IS14013A), Berlin Center for Machine Learning (01IS18037I) and TraMeExCo (01IS18056A). This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein.

[1]*e.g.* via the "right to explanation" proclaimed in the General Data Protection Regulation of the European Union

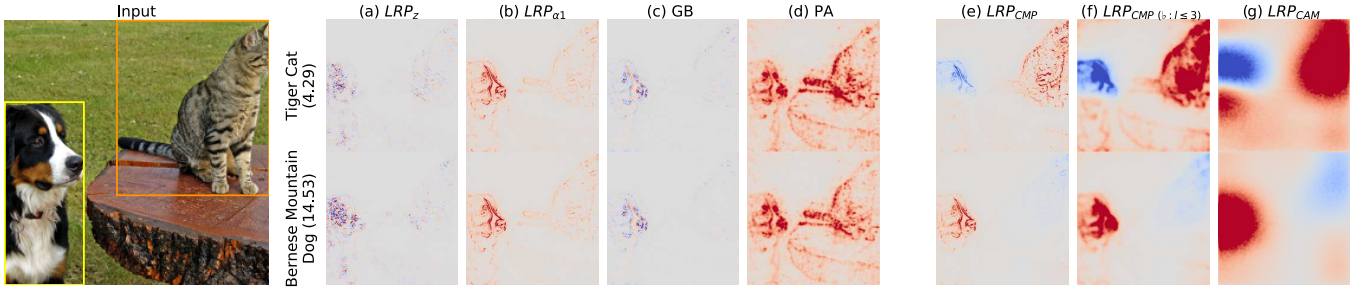

Fig. 1. Different attributions for the output classes "Tiger Cat" and "Bernese Mountain Dog" using the VGG-16 model. Network output strength (logit) is given in parenthesis. Uniformly applied rules *(a)* - *(d)* are not, or hardly class discriminative. LRP_z shows the effect of gradient shattering. Composite strategies *(e)* - *(g)* are sensitive to class-specific information and highlight features on different levels of scale.

outputs j during inference) and their respective aggregations z_j at the layer output. Note that Eq. (1) is conservative between layers and in general maintains an equality $\sum_i R_i^{(l)} = f(x)$ at any layer of the model.

Further purposed LRP-rules beyond Eq. (1) are introduced in [5] which can be understood as advancements thereof. So does the LRP_ε decomposition rule add a signed and small constant ε to the denominator in order to prevent divisions by zero and to diminish the effect of recessive mappings z_{ij} to the relevance decomposition. The $LRP_{\alpha\beta}$ rule performs and then merges separate decompositions for the activatory ($z_{ij} > 0$) and inhibitory ($z_{ij} < 0$) parts of the forward pass. Here, the α parameter permits a weighting of relevance distribution towards activations and inhibitions (β is given implicitely s.t. $\alpha + \beta = 1$ to uphold conservativity). The commonly used parameter $\alpha = 1$ (short: $\alpha 1$) can be derived from DTD and has been rediscovered in *ExcitationBackprop* [22]. Later work [23, 12] introduces $LRP_\flat$, a decomposition which spreads the relevance of a neuron uniformly across all its inputs. This rule (or alternatively the DTD_{z^B} rule [21]) has seen application in networks' input layers, and provides invariance to the decomposition process wrt. to translations in the input domain.

Earlier applications of LRP (*e.g.* [5, 9]) did use one single decomposition rule uniformly over the whole network, which often resulted in suboptimal "explanations" of model behavior [11]. So are LRP_z and LRP_ε respectively identical and highly similar to *Gradient*$\times$ *Input (G*$\times$*I)* in ReLU-activated DNNs [10] and – while working well for shallower models – demonstrate the effect of gradient shattering in overly complex attributions for deeper models [11]. The $LRP_{\alpha\beta}$ demonstrates robustness against gradient shattering and produces visually pleasing attribution maps, however is lacking in class- or object discriminativity [24, 11] by consistently attributing relevance to features unrelated to the object. Further, $LRP_{\alpha\beta}$ introduces the constraint of positive layer activations [21], which is in general not guaranteed, especially at the (logit) output of a model.

2.2. A Current Best Practice for LRP

A recent trend among XAI researchers and practitioners employing LRP is the use of a *composite strategy (CMP)* for decomposing the prediction of a neural network [12, 11, 8, 13, 14]. That is, different parts of the DNN are decomposed using different rules, which in combination are robust against gradient shattering while sustaining object discriminativity. Common among these works is the utilization of LRP_ε with $\varepsilon \ll 1$ (or just LRP_z) to decompose fully connected layers close to the model output, followed by an application of $LRP_{\alpha\beta}$ to the underlying convolutional layers (usually with $\alpha \in \{1, 2\}$). Here, the separate decomposition of the positive and negative forward mappings complements the localized feature activation of convolutional filters activated by, and feeding into ReLUs. A final decomposition step within the convolution layers near the input uses the $LRP_\flat$-rule[2]. Most commonly this rule (or alternatively the DTD_{z^B}-rule) is applied to the input layer only. In summary, we here describe this pattern of rule application as LRP_{CMP}. Fig. 1 provides a qualitative overview of the effect of LRP_{CMP} in contrast to other parameterizations and methods, which we will further discuss in Sec. 4. Note that the option to apply the $LRP_\flat$ decomposition to the first n layers near the input (instead of only the first) provides control over the local and semantic scale [23] of the computed attributions (see Fig. 1*(e)-(g)*). Previous works profit from this option for comparing DNNs of varying depth, and differently configured convolutional stacks [12], or by increasing readability of attributions maps aligned to the requirements of human investigators [13].

3. EXPERIMENTS

The declared purpose of LRP is to precisely and quantitatively inform about the features which contribute towards or against the decision of a model wrt. to a specific predictor output [5]. While the recent LRP_{CMP} exhibits improved properties above previous variants of LRP *by eyeballing*, an objective verification requires quantification. The visual object detection setting, as it is described by the Pascal VOC (PVOC) [18] or ImageNet [25] datasets – both of which include object bounding box annotations – delivers an optimal experimental setting for this purpose.

An assumed ideal model would, in such a setting, exhibit true object understanding by only predicting based on the object itself. A good and *representative* attribution method should therefore reflect that object understanding of the model closely *i.e.* by marking (parts of) the shown object as relevant and disregarding visual features not representing the object itself. Similar to [9], we therefore rely on a measure based on localization of attribution scores. In the following, we will evaluate LRP_{CMP} against other methods and variants of LRP on ImageNet using a pre-trained VGG-16 network, and on PVOC 2007 using a pre-trained (on PVOC 2012) CaffeNet model [9]. Both models perform well on their respective task and have been obtained from `https://modelzoo.co/`.

[2]read: $\flat$ ="flat", as in the musical $\flat$.

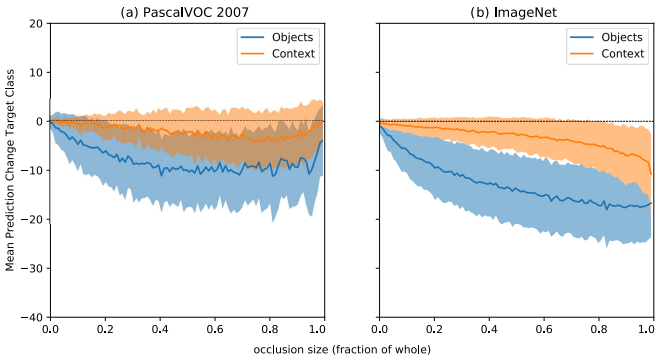

Fig. 2. Mean prediction changes measured in the logit outputs as a function of the occluded area, when occluding the pixels within (object) and without (context) the class-specific bounding boxes on PVOC 2007 (*left*) and ImageNet (*right*).

3.1. Verifying Object-centricity During Prediction

In practice, both datasets can not be assumed to be free from contextual biases (*c.f.* [8]), and in both settings models are trained to categorize images rather than localize objects. Still, we (neccessarily) assume that the models we use dominantly base their decision on the target object, as opposed to the image context. We verify our hypothesis in Fig. 2, showing for both models and datasets the $\Delta f(x)$ of the true class, when occluding the area within the object bounding box in contrast to occluding the remaining image area, relative to the unperturbed $f(x)$. We occlude by replacing the selected pixel coordinates with the respective mean values from the training set. Occluding the object area consistently leads to a sharper decrease in the output for the specific class. The trend is especially evident for smaller objects. This supports our claim that the networks base their decision mainly on the object as opposed to the image context.

3.2. Attribution Localization as a Quantitative Measure

This gives us a performance criterion for attribution methods in object detection and classification: In order to track the fraction of the total amount of relevance that is attributed to the object, we use the inside-total relevance ratio μ without, and a weighted variant μ_w within consideration of the object size:

$$\mu = \frac{R_{\text{in}}}{R_{\text{tot}}} \qquad \mu_w = \mu \cdot \frac{S_{\text{tot}}}{S_{\text{in}}} \tag{2}$$

While conceptually similar to the inside-outside ratio used in [9], μ and μ_w avoid numerical issues in edge cases wrt. bounding box size. Here, R_{in} is the sum of positive relevance in the bounding box, R_{tot} the total sum of positive relevance in the image and S_{in} and S_{tot} are the size of the bounding box and the image respectively, in pixels. The subscript w signals the addition of the normalization factor in μ_w considering the image and object size. Since correctly locating small objects is more difficult than locating image-sized objects, μ_w puts an emphasis on measuring the outcome for small bounding box sizes. In both cases, higher values indicate larger fractions of relevance attributed to the object area (and not background), and therefore are the desirable outcome.

The inside-total relevance ratio highly depends on the size of the bounding box, we thus report the average μ_w as an aggregate of results over differently sized objects and images in Tab. 1 and plot μ as a function of bounding box size in Fig. 3.

3.3. Experimental Setup

We perform our experiments on both the ImageNet and the PVOC 2007 datasets, since both collections provide large numbers of ground truth object bounding boxes.

For PVOC, we compute attribution maps for all samples (approx. 10.000) from PVOC 2007, using a model which has been pretrained on the multi label setting of PVOC 2012 [18, 9]. The respective model performs with a mean AP of 72.12 on PVOC 2007. Since PVOC describes a multi label setting, multiple classes can be present in the same image. We therefore evaluate for μ and μ_w once for each unique existing pair of $\{$ class $\times$ sample $\}$, yielding approximately 15.000 measurements. Images with a higher number of (smaller) bounding boxes thus effectively have a stronger impact on the results than images with larger, image-filling objects, while at the same time describing a *more difficult* setting. Many of the objects shown in PVOC images are not centered. In order to use all available object information in our evaluation, we rescale the input images to the network's desired input shape to avoid (partially) cropping objects.

On ImageNet [17] (2012 version), bounding box information does only exist for the 50.000 validation samples (one class per image) and can be downloaded from the official website[3]. We evaluate a pre-trained VGG-16 model from the keras model zoo, obtained via the iNNvestigate [26] toolbox. The model performs with a 90.1% top-5 accuracy on the ImageNet test set. For all images the shortest side is rescaled to fit the model input and the longest side is center-cropped to obtain a quadratic input shape. Bounding box information is adjusted correspondingly.

For computing attribution maps, we make use of existing XAI software packages, depending on the models' formats. That is, for the VGG-16 model we use the iNNvestigate [26] toolbox. For the PVOC data and the CaffeNet architecture, we compute attributions using the LRP Toolbox [27].

Both XAI packages support the same functionality regarding LRP, yet differ in the provided selection of other attribution methods. Our study, however, shall be focussed on the beneficial or detrimental effects between the variants of LRP used in literature. We compute attributions maps and compute values for μ and μ_w for four variants of LRP_{CMP} ($\alpha1, \alpha2$, each once with and without $LRP_\flat$ at the input), $LRP_{\alpha\beta}$ (both $\alpha1$ and $\alpha2$), and LRP_z, for both models. We complement the results with Guided Backprop [1] and for ImageNet with Pattern Attribution [2] only available in iNNvestigate, as well as LRP_{CAM} for demonstration purposes. LRP_{CAM} is equivalent to LRP_z (and CAM [28], absent the SoftMax) in the fully connected part of the model, but replaces the upsampling over the convolutional stack of CAM with applications of $LRP_\flat$ and distributes negative relevance. On both datasets, we evaluate attributions for the real class label, independent of network prediction.

4. RESULTS AND DISCUSSION

4.1. Qualitative Observations

Fig. 1 exemplarily shows attribution maps computed with different methods based on the VGG-16 model, for two object classes present in the ImageNet labels and the input image; "Bernese Mountain Dog" and "Tiger Cat". Attributions *(a)-(d)* result from uniform rule

[3] http://www.image-net.org/challenges/LSVRC/2012/index

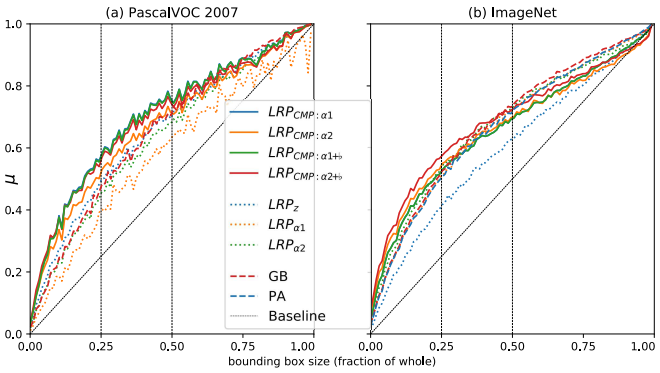

Fig. 3. Average in-total ratio μ as a function of bounding box size. Vertical lines mark thresholds of 25% and 50% covered image area. The baseline can be reached by uniformly attributing to all pixels of the image. Higher values are better.

application to the whole network. Next to applications of LRP_z and $LRP_{\alpha\beta}$, this includes Guided Backprop [1] and Pattern Attribution [2]. Neither of these maps demonstrate class-discriminativeness and prominently attribute scores to the same areas, regardless of the target class chosen for attribution. LRP_z additionally shows the effects of gradient shattering in a highly complex attribution structure due to its equivalence to G×I. Such attributions would be difficult to use and juxtapose in further algorithmic or manual analyses.

To the right, attribution maps *(e)*-*(g)* correspond to variants of LRP_{CMP}, which apply different decomposition rules depending on layer type and position. In Fig. 1 *(e)*, the $LRP_\flat$-rule is only applied to the input layer, while in *(g)* it is used for the first three convolutional layers and the whole convolutional stack in *(f)*. Both *(e)* and *(f)* use $\alpha1$. Here altogether, the visualized attribution maps correspond more to an "intuitive expectation". Fig. 1 *(e)*-*(g)* demonstrates the change in scale and semantic, from attributions to local features to a very coarse localization map, with changing placements of $LRP_\flat$. Further, it becomes clear that with an application of $LRP_{\alpha\beta}$ in upper layers, object localization is lost (see *(b)* vs *(g)*), while an application in lower layers avoids issues related to gradient shattering, as shown in *(e)*-*(f)* compared to *(a)*.

Note that the VGG-16 network used here never has been trained in a multi-label setting. Despite only receiving one object category per input sample, it has learned to distinguish between different object types, *e.g.* that a dog is not a cat. This in turn reflects well in the attribution maps computed after the LRP_{CMP} pattern.

4.2. Quantitative Results

Figs. 3 *(a)* and *(b)* show the average in-total ratio μ as a function of bounding box size, discretized over 100 equally spaced intervals, per size interval, for PVOC 2007 and ImageNet. Averages for μ and μ_w over the whole (and partial) datasets can be found in Tab. 1. Large values indicate more precise attribution to the relevant object. The assumed *Baseline* is the uniform attribution of relevance over the whole image, which is outperformed by all methods.

LRP_z performs noticably worse on ImageNet than on PVOC, which we trace back to the significant difference in model depth (13 vs 21 layers) affecting gradient shattering. We omit LRP_ε due to identity in results to LRP_z. $LRP_{\alpha\beta}$ has the tendency to attribute to all shown objects and suffers from the multiple classes per image in PVOC, where ImageNet shows only one class. Also, the similarity of attributions between *PA* and $LRP_{\alpha1}$ observed in Fig. 1 seem consistent on ImageNet and result in close measurements in Tab. 1.

Table 1. Average context attribution metrics for different analyzers and datasets. Row order is determined by μ_w. Higher μ_* are better.

Data	Analyzer	μ_w	$\mu_{\leq 0.25}$	$\mu_{\leq 0.5}$	μ
PVOC (*CaffeNet*)	$LRP_{CMP:\alpha2+\flat}$	**2.716**	**0.307**	0.421	0.532
	$LRP_{CMP:\alpha1}$	2.664	0.306	**0.426**	**0.539**
	$LRP_{CMP:\alpha1+\flat}$	2.598	0.301	0.421	0.535
	$LRP_{CMP:\alpha2}$	2.475	0.276	0.388	0.504
	LRP_z	2.128	0.236	0.353	0.480
	GB	1.943	0.212	0.335	0.470
	$LRP_{\alpha2}$	1.843	0.205	0.320	0.452
	$LRP_{\alpha1}$	1.486	0.163	0.273	0.403
	Baseline	1.000	0.100	0.186	0.322
ImageNet (*VGG-16*)	$LRP_{CMP:\alpha2+\flat}$	**1.902**	**0.397**	**0.534**	**0.714**
	$LRP_{CMP:\alpha2}$	1.797	0.368	0.505	0.693
	$LRP_{CMP:\alpha1}$	1.7044	0.3467	0.4887	0.6898
	$LRP_{CMP:\alpha1+\flat}$	1.7043	0.3466	0.4886	0.6898
	$LRP_{\alpha2}$	1.702	0.332	0.496	0.706
	GB	1.640	0.312	0.485	0.710
	$LRP_{\alpha1}$	1.609	0.306	0.475	0.699
	PA	1.591	0.303	0.471	0.698
	LRP_z	1.347	0.236	0.389	0.632
	Baseline	1.000	0.128	0.260	0.547

Tab. 1 shows that LRP_{CMP} clearly outperforms other methods consistently on large datasets. That is, the increased precision in attribution to relevant objects is especially evident in the presence of smaller bounding boxes in μ_w. This can also be seen in $\mu_{\leq 0.25}$ and $\mu_{\leq 0.5}$ in Tab. 1 and the left parts of Figs. 3 *(a)* and *(b)*, where a majority of the image shows contextual information or other classes. Once bounding boxes become (significantly) larger and cover over 50% of the image, all methods converge towards perfect performance. In both settings, $LRP_{CMP:\alpha2+\flat}$ ($\alpha2$ and using $\flat$) yields the best results, while overall the composite strategy is more effectful than a fine tuning of decomposition parameters.

4.3. Conclusion

In this study, we discuss a recent development in the application of Layer-wise Relevance Propagation. We summarize this emerging strategy of a composite application of multiple decomposition rules as LRP_{CMP} and juxtapose its effects to previous approaches which uniformly apply a single decomposition rule to all layers of the model. Our results show that LRP_{CMP} does not only yield more representative attribution maps, but also provides a solution against gradient shattering affecting previous approaches and improves properties related to object localization and class discrimination. The discussed beneficial effects are demonstrated qualitatively and verified quantitatively at hand of two large and popular computer vision datasets. While alternative approaches with the aim to achieve a similar effect to LRP_{CMP} may rely on multiple backward passes through the model or are unable to attribute negative relevance to class-contradicting features, LRP_{CMP} needs only one backward pass using established tools from the LRP framework.

5. REFERENCES

[1] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M.A. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. of International Conference on Learning Representations (ICLR)*, 2015.

[2] P.-J. Kindermans, K.T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," in *Proc. of International Conference on Learning Representations (ICLR)*, 2018.

[3] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. of International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

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

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, pp. e0130140, 2015.

[6] Y. Yang, V. Tresp, M. Wunderle, and P. A. Fasching, "Explaining therapy predictions with layer-wise relevance propagation in neural networks," in *Proc. of IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, pp. 152–162.

[7] A.W. Thomas, H.R. Heekeren, K.-R. Müller, and W. Samek, "Analyzing neuroimaging data through recurrent deep learning models," *CoRR*, vol. abs/1810.09945, 2018.

[8] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1096, 2019.

[9] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2912–2920.

[10] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191. Springer, 2019.

[11] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 193–209. Springer, 2019.

[12] S. Lapuschkin, A. Binder, K.-R. Müller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in *Proc. of IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1629–1638.

[13] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *CoRR*, vol. abs/1908.06943, 2019.

[14] L. Y. W. Hui and A. Binder, "Batchnorm decomposition for deep neural network interpretation," in *International Work-Conference on Artificial Neural Networks (IWANN)*, 2019, pp. 280–291.

[15] D. Balduzzi, M. Frean, L. Leary, J.P. Lewis, K. W.-D. Ma, and B. McWilliams, "The shattered gradients problem: If resnets are the answer, then what is the question?," in *Proc. of International Conference on Machine Learning (ICML)*, 2017, pp. 342–350.

[16] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Network Learning Systems (TNNLS)*, vol. 28, no. 11, pp. 2660–2673, 2017.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, Ma S., Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[21] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[22] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision (IJCV)*, vol. 126, no. 10, pp. 1084–1102, 2018.

[23] S. Bach, A. Binder, K.-R. Müller, and W. Samek, "Controlling explanatory heatmap resolution and semantics via decomposition depth," in *Proc. of IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2271–2275.

[24] J. Gu, Y. Yang, and V. Tresp, "Understanding individual decisions of cnns via contrastive backpropagation," in *Proc. of Asian Conference on Computer Vision (ACCV)*, 2018, pp. 119–134.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1106–1114.

[26] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks!," *Journal of Machine Learning Research (JMLR)*, vol. 20, pp. 93:1–93:8, 2019.

[27] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and S. Samek, "The LRP toolbox for artificial neural networks," *Journal of Machine Learning Research (JMLR)*, vol. 17, pp. 114:1–114:5, 2016.

[28] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.