ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

SmoothLRP: Smoothing LRP by Averaging over Stochastic Input Variations

Arne Peter Raulf[1,*], Sina Däubener[2,*,†], Ben Luis Hack[1,*], Axel Mosig[1], Asja Fischer[2,†]

1- Department of Bioinformatics, Ruhr University, Bochum, Germany

2- Department of Mathematics, Ruhr University, Bochum, Germany

Abstract. Explanations of neural networks predictions are a necessity for deploying neural networks in safety critical domains. Several methods were developed which identify most relevant input features, such as sensitivity analysis and layer-wise relevance propagation (LRP). It has been shown that the noise in the explanations from the sensitivity analysis can be reduced by averaging over noisy input images, a method referred to as SmoothGrad. We investigate the application of the same principle to LRP and find that it smooths the resulting relevance function leading to improved explanations. Moreover, it can be applied for restoring the correct label of adversarial examples.

1 Introduction

Due to the wide-spread use of deep neural networks (DNNs), the question of how decisions of these models can be interpreted gains rising importance. As a consequence a multitude of methods have been developed in the recently established field of explainable artificial intelligence. These methods range from gradient based approaches [14, 15, 16], where the gradient of the prediction in direction of the input is used to infer and highlight relevant areas - up to concept based approaches like for example TCAV [5], where human concepts like "stripes" for a zebra classification can be explicitly tested. One approach which has established itself as a prominent method for the interpretability of DNNs is layer-wise relevance propagation (LRP) [2, 8, 3]. LRP leverages the graph structure to extract meaningful explanations while at the same time fulfilling desired properties gradient based approaches do not necessarily. One property for example is the *conservation* property [9], which grants that each layer captures the same amount of information. However, Montavon et al. [10] noted that LRP can lead to unsatisfactory results when applied to classifiers that are not optimally trained or networks with specific structures (e.g., noisy first-layer filters or a large stride parameter in the first convolution layer). They suggest to mitigate these effects by replacing the explanation of a single input image by the explanations of multiple slightly translated versions of it. Another strategy for making DNNs more robust is by adding Gaussian noise to the training images [13, 17]. This approach was also shown to be effective for smoothing sensitivity maps, leading to a method referred to as SmoothGrad [15].

*Equal contribution.

†Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

We build on this idea by averaging over multiple stochastic variations of the input, a procedure we refer to as SmoothLRP. Our contributions are twofold:

- We analyze the effect of using SmoothLRP as a wrapper over existing approaches on the resulting heatmaps.
- We exploit the relevance variance to reverse adversarial examples back to benign examples.

Because of the high level similarity to SmoothLRP, we compare SmoothLRP also to LRP for Bayesian neural networks [4]. The idea for relevance quantification in this setting is to take the single predictions of the Monte Carlo approximation and calculate the LRP values for each of them. While the theoretical motivation is fundamental different, their approach like ours incorporates stochasticity into LRP. Bayesian-LRP (B-LRP) does this by taking model uncertainty into account, while SmoothLRP rather models input uncertainty, as we will explain in more detail.

2 SmoothLRP

First we give a brief description of the different LRP methods. In the following let f be a trained DNN with L hidden layers. We denote by $w_{i,j}^{(l)}$ the weight connecting the ith neuron of layer l with the jth neuron of layer $l+1$ and with $a_i^{(l)}$ the activation of the ith neuron in layer l after applying the non-linearity. Then, LRP-0 [2] defines the relevance of the ith neuron in layer l as

$$R_i^{(l)} = \sum_j \frac{a_i^{(l)} \cdot w_{i,j}^{(l)}}{\sum_{i'} a_{i'}^{(l)} \cdot w_{i',j}^{(l)}} R_j^{(l+1)} \ , \tag{1}$$

where $R_j^{(L+1)} = f_j(\boldsymbol{x})$ is the output of the jth output neuron and $R_i^{(0)}$ is the relevance at the input level. However, it has been shown that this 0-rule is equivalent to gradient $\times$ input [12] and gradients are known to be noisy [15]. Therefore, the 0-rule was extended by the ϵ-rule [2], where a small value is added to the denominator in eq. (1) to reduce the impact of low contributions. Another well known extension is the γ-rule where positive connections are enhanced [8] by up-weighting positive weights by a factor γ. Finally, LRP-composite (LRP-CMP) [6] applies different LRP rules at different layers.

Our approach merges vanilla LRP-rules and the idea of smoothing image explanations by averaging over stochastic input variations introduced by SmoothGrad. More formally, let $\boldsymbol{x}$ be an arbitrary input and $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\epsilon}$ the stochastic input variation, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \sigma^2 \cdot \boldsymbol{I})$ being a vector valued random variable following an uncorrelated multivariate normal distribution. We are interested in the expected relevance of $\tilde{\boldsymbol{x}}$ under $\mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \sigma^2 \cdot \boldsymbol{I})$, that is

$$\mathbb{E}_{p(\boldsymbol{\epsilon})}[R_i^{(0)}(\tilde{\boldsymbol{x}})] = \int p(\boldsymbol{\epsilon}) \cdot R_i^{(0)}(\boldsymbol{x} + \boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \ ,$$

and refer to the method estimating it as SmoothLRP (cf. figure 3, Algorithm 1).

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

3 Experiments

Throughout the experiments we used 50 noisy input versions with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.05^2)$ for estimating the relevance scores of SmoothLRP.

3.1 Investigating explainability properties

In this section we investigate whether SmoothLRP improves upon the vanilla LRP rules. For this, we start by comparing the relevance heatmaps generated by LRP-0, LRP-γ, LRP-ϵ, LRP-CMP, and BLRP to those provided by their smoothed counterparts on the often used "castle" image (cf. figure 1). To gain a direct comparison to the vanilla LRP, we extended the original code provided by the heatmapping tutorial (heatmapping.org) and Bykov et al. [4], which both use the pretrained VGG16 network provided by torchvision.

The resulting explanations are shown in figure 1. For all LRP methods we see minor improvements for their smoothed version, which we briefly point out in the following. For example, comparing the explanations provided by vanilla LRP to those of SmoothLRP an increase of the areas of positive relevance scores around the castle structure can be observed accompanied with an increase in the interrelation between the pixels. This can best be seen for LRP-0 but also for LRP-ϵ. For LRP-ϵ and LRP-CMP only the SmoothLRP versions classify the complete lampshade in the top as explicitly uninformative for the prediction. Also visible in all heatmaps is the reduction of noise, which can be seen by the area under the castle. Interestingly, while BLRP incorporates model uncertainty into their relevance scores, the 50% quantile of SmoothLRP shows slightly more details for the demarcation of relevant and non-relevant areas which stronger smoothing properties.

Since LRP-CMP is regarded as state of the art LRP method we investigated heatmaps produced by it to their smoothed versions for the first 300 images of the ImageNet validation data. In almost all cases we see the background noise reduction effect (cf. figure 2). On images with a lot of features, SmoothLRP helps to focus on the relevant parts and even sometimes shifts the whole focus as can be seen in figure 2 D). To gain a better evaluation we conducted a small survey in which 6 participants were asked to indicate which heatmap they found to provide better explanations for the corresponding label. Images with bold printed labels were the ones where the LRP heatmap was preferred by users. We found a strong correlation between preferences for LRP and the amount of details/edges shown by the heatmaps, which are often reduced by SmoothLRP.

3.2 Reversing adversarial examples

During our experiments, we found that adversarial examples (AEs) consistently showed a higher prediction variance over the noisy input variations as well as a higher variance in the relevance scores compared to their benign counterparts. This motivated us to investigate whether AEs can automatically be detected and reversed. The underlying hypothesis is that because AEs are created by adding a

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

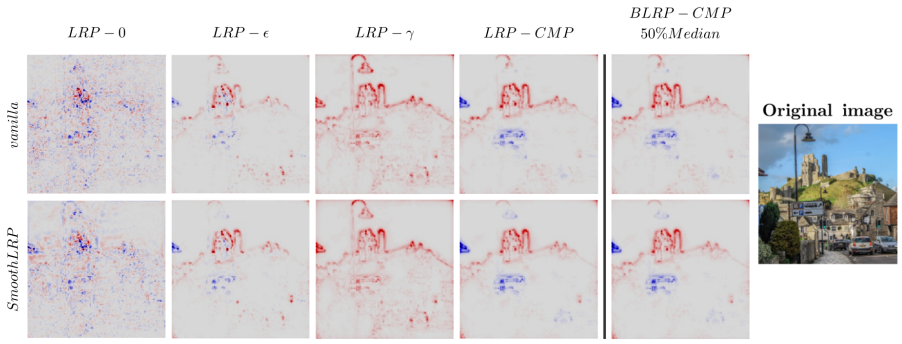

Fig. 1: **Comparison of heatmaps:** Red indicates positive and blue negative relevance. Note that this picture has different non-relevant pattern like the street lights and the signs. Original image of the castle (right).

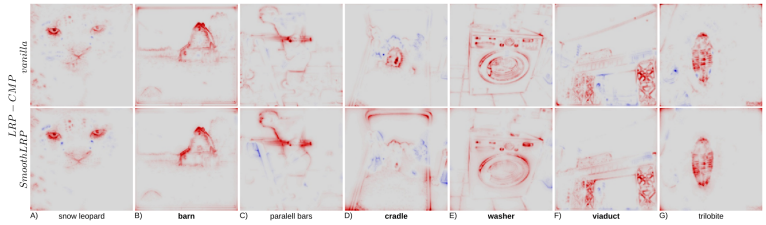

Fig. 2: **Comparison of LRP-CMP and smoothLRP-CMP heatmaps:** Red indicates positive and blue negative relevance. Bold labels indicate that users preferred heatmaps generated with LRP over SmoothLRP.

specific δ to the image, the stochastic variations added to the input image might reverse some of the well calibrated δ-changes. Therefore, pixels with heavily varying relevance scores might indicate the pixels relevant for the malicious class label. This experiment was conducted for the VGG16 network trained on the ImageNet data set provided by Tensorflow/Keras. For the relevance calculation we used uniform LRP-ϵ (with $\epsilon = 1e-9$) provided by the toolbox iNNvestigate [1].

We proceeded as described: (i) We applied the projected gradient decent attack [7] implemented in Foolbox [11] with perturbation strength 5 and 10 iterations to the first 300 images of the ImageNet validation set, out of which 253 turned to successful AEs, i.e. AEs leading to a misclassification. (ii) We did the same for the 20 Foolbox images which were then used to calculate the pixel-wise relevance variances over the 50 SmoothLRP scores for the benign images and their corresponding AEs. For each image/AE we then estimated the mean of these relevance variances and identified a mean based classification threshold for distinguishing between benign and adversarial images. (iii) We calculated the

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Algorithm 1 SmoothLRP

input $\boldsymbol{x}$ original image, σ^2 variance of the added noise, M number of stochastic input variations, D input dimension
  for $m = 1, \ldots, M$ **do**
    Sample $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \sigma^2 \cdot \boldsymbol{I})$
    Set $\tilde{\boldsymbol{x}}_m = \boldsymbol{x} + \boldsymbol{\epsilon}_m$
    for $i = 1, \ldots, D$ **do**
      Calculate $R_i^{(0)}(\tilde{\boldsymbol{x}}_m)$
    end for
    $R^{(0)}(\tilde{\boldsymbol{x}}_m) = \{R_i^{(0)}(\tilde{\boldsymbol{x}}_m)\}_{i=1}^{D}$
  end for
  Relevance set = $\{R^{(0)}(\tilde{\boldsymbol{x}}_m)\}_{m=1}^{M}$
  Relevance: $\frac{1}{M}\sum_{m=1}^{M} R^{(0)}(\tilde{\boldsymbol{x}}_m)$
output Relevance set, relevance

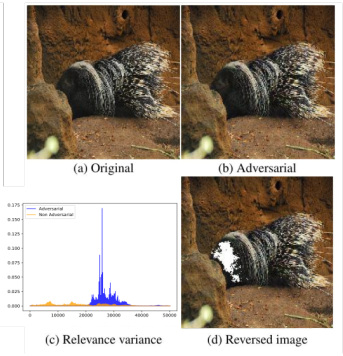

Fig. 3: Left: **Pseudo code** for the SmoothLRP algorithm. Right: **Reversing the label of an adversarial example.** (a) Benign image and its (b) adversarial version can be distinguished by their c) pixel-wise variances of the relevance values. (d) By changing 1000 pixels to white the correct label *porcupine* is restored.

mean of the pixel-wise relevance variance for the 253 successful ImageNet AEs and were able to correctly classify 138 of these based on the threshold classifier. (iv) For each of these 138 AEs we changed the 1000 pixels with the highest relevance variances simultaneously to white and repeated this procedure until the prediction label changed. Following this procedure, out of the 138 AEs 44 image labels have been successfully restored.

4 Conclusion and Discussion

In this paper we investigated the effect of smoothing existing LRP-rules by averaging over stochastic input variations, a technique we refer to as SmoothLRP. We showed that this method improves on current LRP-rules by increasing interrelations between pixels and by reducing the impact of non-relevant features. Further, we showed that while being more naive it performs comparable to Bayesian LRP, a method applying LRP to Bayesian neural networks, indicating that being able to reason about the model uncertainty does not improve the explainability compared to modeling simple input based uncertainty. Moreover, we showed how SmoothLRP can be employed for restoring the true label of adversarial examples. Even though these results were obtained by experimenting with a quite small dataset and are in this sense preliminary, they seem very promising for a deeper analysis in future.

References

[1] 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*, 20(93):1–8, 2019.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

[2] 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*, 10(7), 2015.

[3] A. Binder, S. Bach, G. Montavon, K.-R. Müller, and W. Samek. Layer-Wise Relevance Propagation for Deep Neural Network Architectures. In K. J. Kim and N. Joukov, editors, *Information Science and Applications (ICISA)*, pages 913–922. Springer Singapore, 2016.

[4] K. Bykov, M. M. C. Höhne, K.-R. Müller, S. Nakajima, and M. Kloft. How Much Can I Trust You? – Quantifying Uncertainties in Explaining Neural Networks. *arXiv preprint arXiv:2006.09000*, 2020.

[5] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viegas, and R. Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[6] S. Lapuschkin, A. Binder, W. Samek, and K. Müller. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. In *IEEE International Conference on Computer Vision Workshops*, 2017.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[8] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, 2019.

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

[10] G. Montavon, W. Samek, and K. R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15, 2018.

[11] J. Rauber, W. Brendel, and M. Bethge. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, International Conference on Machine Learning*, 2017.

[12] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[13] J. Sietsma and R. J. F. Dow. Creating artificial neural networks that generalize. *Neural Networks*, 4(1):67â79, 1991.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations, Workshop Track Proceedings*, 2014.

[15] D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[16] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[17] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4480–4488, 2016.