Neural Processing Letters
https://doi.org/10.1007/s11063-023-11166-8

Evaluation of the Explanatory Power Of Layer-wise Relevance Propagation using Adversarial Examples

Tamara R. Dieter[1,2] · Horst Zisgen[3]

Accepted: 1 February 2023

Abstract

Approaches for visualizing and explaining the decision process of convolutional neural networks (CNNs) have recently received increasing attention. Particularly popular approaches are so-called saliency methods, which aim to assign a valence to each input pixel based on its importance and influence on the classification via saliency maps. In our paper, we contribute by a novel analyzing approach build on adversarial examples to investigate the explanatory power of saliency methods exemplified by layer-wise relevance propagation (LRP). Based on the hypothesis that distinct decisions, such as an image's classification and the classification of its corresponding adversarial examples, should yield to dissimilar saliency maps to provide transparent rationales, we break down relevance scores of images and corresponding adversarial examples and analyze them using a comprehensive statistical evaluation. It turns out that different relevance decomposition rules of LRP do not lead to clearly distinguishable saliency maps for images and corresponding adversarial examples, neither in terms of their contour lines, nor in terms of the statistical analysis.

Keywords Deep learning · Layer-wise relevance propagation · Adversarial examples · Explainable artificial intelligence · Saliency maps

1 Introduction

Deep learning (DL) has achieved impressive performance in a lot of application areas, especially in object recognition [34]. However, their nested non-linear structure combined with millions of parameters makes deep learning a black-box method that arrives at decisions in a non-transparent manner. Therefore, the explainability of DL models [9, 13, 21, 33, 35] and

✉ Tamara R. Dieter
tamara.dieter@h-da.de

Horst Zisgen
horst.zisgen@h-da.de

[1] Department of Computer Science, University of Applied Sciences, Darmstadt, Germany

[2] Institute for the Protection of Terrestrial Infrastructures, German Aerospace Center (DLR), Sankt Augustin, Germany

[3] Department of Mathematics and Natural Science, University of Applied Sciences, Darmstadt, Germany

Published online: 10 March 2023

the development of associated techniques to visualize, explain, and interpret the behavior of DL algorithms have recently received increasing attention [3, 8, 16, 19, 22, 24, 25, 28, 30, 31, 36, 38, 40]. Apart from methods optimizing activations [10, 25, 36] and inverting input representations [8, 22], visualizing and explaining the decision process of *convolutional neural networks (CNNs)* by saliency methods is an intensively discussed approach [1, 3, 16, 19, 20, 24, 28, 30, 31, 38–40].

Saliency methods attempt to explain an algorithm's decision by assigning pixel-level values that reflect the importance of input components in terms of their contribution to the classification result. Therefore, saliency methods generally lead to so-called *saliency maps* [30] (also known as *input contribution heatmaps* or *feature importance maps*), which try to explain the decision process of CNNs through

i. *Input Modification*, i.e., assigning a relevance to a pixel based on the drop in prediction probability caused by the pixel's perturbation [20, 38, 39],
ii. *Class Activation*, i.e., combining the activation pattern of a higher-level layer with further information, such as the network's output [40],
iii. *Backpropagation*, i.e., tracing the contribution of the output nodes backwards through the network to the input nodes [3, 16, 38].

A particularly common backpropagation approach is *layer-wise relevance propagation (LRP)* introduced by Lapuschkin [3, 19]. Layer-wise relevance propagation relies on the assumption that the total amount of relevance is preserved when decomposing the classification decision backwards to the pixel-wise relevance scores. This so-called layer-wise conservation principle postulates that the sum of relevance assigned to neurons in a CNN layer remains the same for two adjacent layers. Despite the popularity of using saliency methods to explain DL models, a significant number of papers have been published addressing more intensively the stability and robustness of saliency methods [1, 3, 12, 15, 18, 19, 23, 28]. Therefore, our work aims to explore the use of adversarial examples as a further tool to help evaluating the robustness and explanatory power of techniques devoted to the explainability of DL models, exemplified by LRP.

1.1 Related Work on the Explanatory Power of Saliency Methods

In the work of Samek et al. [28], the amount of changed classification probability is presented as possible measure to evaluate the explanation of the decision process provided by LRP. For the investigation of the explanatory power of LRP, they suggest replacing the input variables considered most relevant with samples from a probability distribution, such as the Uniform or Dirichlet distribution. In this case, a large decrease in classification probability caused by a perturbation of the input variables with the highest relevance scores is considered to be an indicator of a suitable explanation. A similar idea is followed by Bach et al. [3] who evaluated the impact of single value perturbations on the detection result by flipping pixels with highly positive and highly negative relevance scores, as well as pixels with relevance scores close to zero. Lapuschkin [19] presents a more generalized approach by employing an iterative greedy procedure to evaluate the expected behavior of LRP.

The work of Ghorbani et al. [12], on the other hand, shows for various gradient-based methods and DeepLIFT that the same object classification for two extremely similar images can be explained by different saliency maps. They perform slight modifications to the input images to ensure a similar classification of the modified images and the originals, however, leading to a substantial difference in saliency maps. Similarly, Kindermans et. al. [18] analyze the invariance of saliency map generating methods to transformations of the input data which have no impact on the prediction outcome. Instead of modifying input images, Heo et al. [15] adversarially manipulate the classification model, leaving the model accuracy unchanged while achieving a dramatic change in explanation. Therefore, they are able to perform model manipulations that result in modified models classifying an input object with nearly the same classification probability as the original model. However, the saliency maps differ significantly in dependence of the underlying model. Further, Adebayo et al. [1] present a sanity check for saliency methods comprising model randomization and data randomization tests. Based on their observations, they find that some saliency methods (e.g., gradient$\odot$input) can be interpreted as implicitly implemented techniques analogous to edge detection tending to detect edges rather than explain decisions.

Similarities in saliency maps of marginally perturbed images intentionally designed to cause a major shift in classification (also known as *adversarial examples* [29]) and the corresponding originals are also observed by other authors, such as Gu and Tresp [14] and Brama et al. [5]. Aiming to use saliency maps of adversarial examples to develop defense strategies against adversarial attacks [29, 37], Brama et al. [5] use binarized saliency maps based on the 5% of the highest scoring pixels (i.e., pixels most relevant for the classification result) to reveal class-discriminating information and illustrate similarities. Neither the work by Brama et al. [5] nor the work by Gu and Tresp [14] exceeds a visual comparison of contour lines within the saliency maps. This applies to the work of Montavon et al. [24] as well. However, visual inspection is insufficient to assess whether an explanation is model-sensitive, as Adebayo et. al. demonstrate in [1].

1.2 Contribution

In this paper, we investigate the use of minimally invasive classification shift perturbations, more precisely adversarial examples [29, 37], to evaluate the robustness and stability of the explanations obtained from saliency maps exemplified by LRP. Our approach focuses on a relevance score-independent input modification causing a completely different classification, and therefore clearly separates from [3, 19, 28], which measures the variation in classification probabilities following relevance score-dependent perturbation of input variables. This also distinguishes our approach from the approach taken by Ghorbani et al. [12], where the input image is perturbed as well, yet without changing the classification decision. Moreover, the approach of Ghorbani et al. [12] is motivated by the expectation that, given a reproducible and consistent explanatory pattern, a minor input change should not affect the classification decision, and hence the saliency map.

In contrast, we present a novel method to analyze the explanatory power of LRP by breaking down LRP-relevance scores of images and corresponding adversarial examples based on the hypothesis that distinct decisions should yield to dissimilar saliency maps even for small changes of the input variables to provide comprehensible rationales. In other words: To explain a turn in a classifier's original decision, the decision turn should result in a variation in the saliency map for a significant proportion of the input variables considered most relevant for achieving the original decision.

Furthermore, we present a novel approach to statistically compare LRP-based saliency maps using relevance score distributions and relevance score rankings. We provide a comprehensive statistical analysis of LRP-based saliency maps of images and corresponding adversarial examples in terms of changes in relevance score distribution and variations in

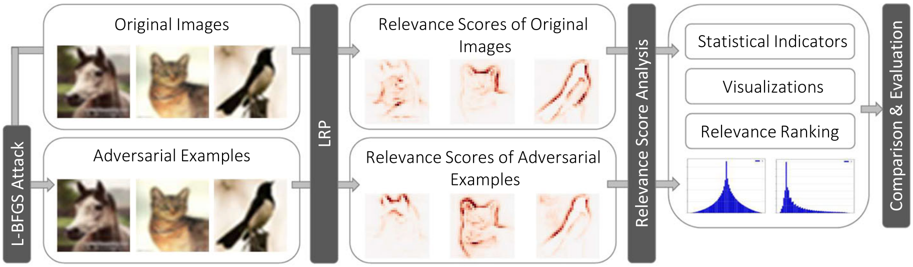


Fig. 1 Procedure of analyzing the explanatory power of LRP using adversarial examples. First, the adversarial examples are created based on the original data (cf. Sect. 3.2). Then, the relevance scores are generated for original images and the adversarial examples using LRP (cf. Sect. 3.3). Finally, the relevance scores are comprehensively analyzed, compared, and evaluated using established statistical methods (cf. Sect. 3.4)

the scores of components marked highly relevant by LRP before and after image perturbation. Consequently, our approach extends current analysis of the explanatory power of saliency maps [3, 5, 14] by going beyond a rather subjective visual comparison of contour lines. Using multiple relevance decomposition rules of LRP, we demonstrate that different decomposition rules do not produce clearly distinguishable saliency maps for images and corresponding adversarial examples, neither regarding their contour lines, nor regarding the statistical measures mentioned above. Finally, we assess the suitability of our presented approach as potential evaluation tool for the adequacy of saliency methods and thus as extension of existing methods [1, 12, 15, 18] by analyzing whether the differences in saliency maps for pairs of images and adversarial examples are significant to explain the difference in classification.

The saliency maps and adversarial examples underlying our analyses are generated based on a simple CNN architecture using CIFAR-10 [4] data and the L-BFGS attack [6, 29, 37]. A schematic representation of our approach is shown in Fig. 1. The related source code, as well as the generated adversarial examples can be found in [7].

1.3 Outline

The remaining paper is organized as follows. Section 2 gives a brief introduction to the fundamental principles and applied methodologies underlying the analysis of this work. In Sect. 3, the experimental setup for generating the adversarial examples, as well as the relevance scores and the procedure of their analysis are described, including the underlying dataset and the applied CNN architecture. Finally, the most significant results of the analysis are summarized in Sect. 4, followed by a brief discussion and a conclusion (Sect. 5).

2 Fundamental Principles

This section briefly outlines the basic principles of layer-wise relevance propagation and adversarial examples including general definitions and generation techniques.

2.1 Layer-wise Relevance Propagation

Layer-wise relevance propagation represents a methodical approach aiming to increase the transparency and interpretability of individual classification decisions. LRP strives to identify important components by decomposing the classifier's output into individual contributions of

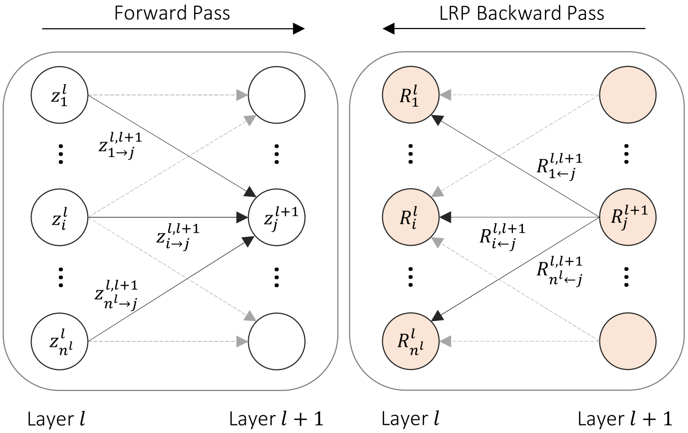


Fig. 2 Principle procedure of layer-wise relevance propagation, showing the idea of redistributing the relevance score R_j^{l+1}, $j \in \{1, \dots, n^{l+1}\}$, of the j-th component of layer $l+1$ (right) in dependence of the corresponding input components' forward contributions (left)

the input components. The application of LRP is based on the assumption that the classifier is decomposable into $n_L \in \mathbb{N}$ individual size $n^l \in \mathbb{N}$ layers and therefore, the classification function $f : \mathbb{R}^n \longrightarrow \mathcal{C}$ can be represented as composition of the functions f^l with $l \in \{1, \dots, n_L\}$, i.e., $f = f^{n_L} \circ ... \circ f^1$, whereby $\mathcal{C}$ denotes the set of available classes. The classification function is required to create mappings between intermediate representations of the input data $X \in \mathbb{R}^n$, which are generally denoted by $z^l \in \mathbb{R}^{n^l}$. The mapping between the i-th component of layer l and the j-th component of layer $l+1$ is defined by $z_{i\to j}^{l,l+1}$, so that

$$z_j^{l+1} = \sum_{i=1}^{n^l} z_{i\to j}^{l,l+1} \tag{1}$$

holds for all $j \in \left\{1, \dots, n^{l+1}\right\}$ (see Fig. 2).

Using layer-wise relevance propagation, the degree of a component's influence on the final decision score is measured by the relevance score R which constitutes a relative measure of a component's contribution to the network's outcome. The sign of R indicates the direction of the contribution, whereby $R > 0$ implicates a positive and $R < 0$ a contradictory contribution to the classifier's outcome, i.e., a contribution that contradicts the final decision score. A component with a relevance score close to zero is expected to be irrelevant with regard to the decision made by the classifier.

Proceeding from the model's final layer n_L with an initial relevance score of $R^{n_L} = f(X)$, LRP successively propagates the relevances backwards through the network until the input layer is reached. Considering a multi-class classification problem solved via deep neural networks, the classifier's prediction usually results in a vector containing probabilities for each existing class. In this case, the relevance score R^{n_L} is initialized with the value of the class that is supposed to be explained. Under the assumption that the relevance R_j^{l+1} of each component z_j^{l+1} of layer $l+1$ (cf. Eq. 1) has already been identified, the relevances of the previous layer's components z_i^l is given by

Table 1 Relevance message based on commonly used decomposition rules

Relevance decomposition rule	Relevance message
Basic decomposition rule (LRP-0)	$R_{i\leftarrow j}^{l,l+1} = \frac{z_{i\to j}^{l,l+1}}{z_j^{l+1}} R_j^{l+1}$
ε-Rule (LRP-ε)	$R_{i\leftarrow j}^{l,l+1} = \frac{z_{i\to j}^{l,l+1}}{z_j^{l+1}+\varepsilon\cdot\mathrm{sgn}\left(z_j^{l+1}\right)} R_j^{l+1}$ with $\varepsilon > 0$
$\alpha\beta$-Rule (LRP-$\alpha\beta$)	$R_{i\leftarrow j}^{l,l+1} = \left(\alpha \frac{\left(z_{i\to j}^{l,l+1}\right)^+}{\left(z_j^{l+1}\right)^+} + \beta \frac{\left(z_{i\to j}^{l,l+1}\right)^-}{\left(z_j^{l+1}\right)^-}\right) R_j^{l+1}$ with $\alpha \geq 1$, $\beta \in \mathbb{R}, \alpha + \beta = 1$

$$R_i^l = \sum_{j=1}^{n^{l+1}} R_{i\leftarrow j}^{l,l+1} \tag{2}$$

(cf. Fig. 2). The relevance message $R_{i\leftarrow j}^{l,l+1}$, directed from component j to component i, describes the ratio of the relevance score R_j^{l+1} that can be traced back to the i-th component of layer l and can be determined, for example, according to the decomposition rules listed in Table 1. The signum function occurring in Table 1 is defined as follows:

$$\mathrm{sgn}\left(z_j^{l+1}\right) = \begin{cases} 1 & \text{if } z_j^{l+1} \geq 0 \\ -1 & \text{otherwise} \end{cases} .$$

Positive forward contributions $\left(z_{i\to j}^{l,l+1}\right)^+$ and and negative forward contributions $\left(z_{i\to j}^{l,l+1}\right)^-$ are defined by

$$\left(z_{i\to j}^{l,l+1}\right)^+ = \begin{cases} z_{i\to j}^{l,l+1} & \text{if } z_{i\to j}^{l,l+1} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and

$$\left(z_{i\to j}^{l,l+1}\right)^- = \begin{cases} 0 & \text{if } z_{i\to j}^{l,l+1} \geq 0 \\ z_{i\to j}^{l,l+1} & \text{otherwise} \end{cases} . \tag{4}$$

Further decomposition rules, as well as a detailed description of the aforementioned decomposition rules can be found in [3], [19] and [24].

2.2 Adversarial Examples

Adversarial examples are especially common in the area of image classification and object recognition, intentionally designed to deceive machine learning models and provoke misclassifications with high probabilities. They are characterized by a close resemblance to the training data and cannot be differentiated from regular input images by human observers (see Fig. 3).

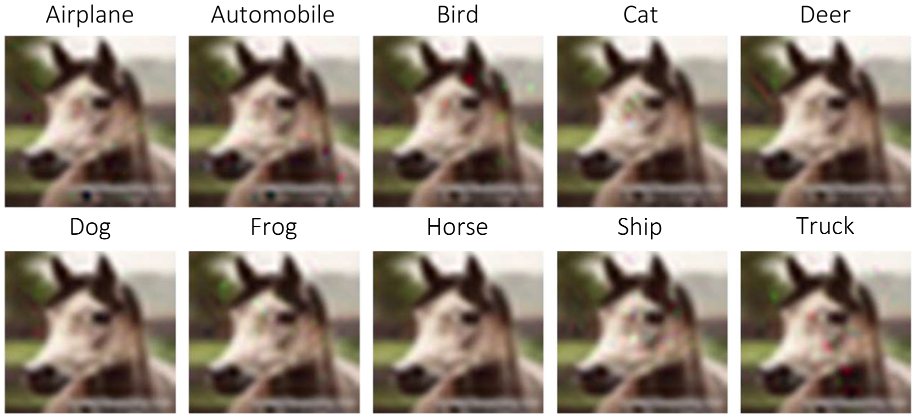


Fig. 3 Results of the L-BFGS attack targeting the classes *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *ship* and *truck* based on an image $X \in \mathbb{R}^n$ (bottom row, third from the left) originally assigned to the class *horse*. All adversarial examples show a classification probability $p_{c'}$ over 99 % towards the attack's corresponding target class $c' \in \mathcal{C}$

Based on a pre-trained classifier $f : \mathbb{R}^n \longrightarrow \mathcal{C}$, the generation of adversarial examples can be mathematically defined as constrained optimization problem. Given a set of potential classes $\mathcal{C}$ as well as a reference image $X \in \mathbb{R}^n$ belonging to the class $c \in \mathcal{C}$, the optimization problem reads as follows:

$$\begin{aligned} \min_{\eta} \quad & \|\eta\|_p \\ \text{s.t.} \quad & f(X') = c' \\ & f(X) = c \\ & c' \neq c \\ & L \leq X' \leq U \end{aligned} \tag{5}$$

with $\eta = X - X'$ meaning the discrepancy between the reference image and the adversarial example $X' \in \mathbb{R}^n$. The parameters $L, U \in \mathbb{R}^n$ represent the component-wise lower and upper bounds on the pixel values of the adversarial example belonging to the target class $c' \in \mathcal{C}$. The determination of the minimal perturbation η^*, that is needed to provoke a misclassification of the reference image, is a complex, nontrivial problem. There is a wide range of algorithms, so called *adversarial attacks*, which enable the approximate solution of this problem. Adversarial attacks, such as the *Fast Gradient Sign method* or the *L-BFGS attack* (see Sect. 3.2), are usually based on different algorithmic approaches and assumptions. An extensive survey of existing adversarial attacks is given by [29] and [37].

3 Experimental Settings

3.1 Dataset and Network Architecture

The generation of the adversarial examples forming the foundation of the analysis covered within this paper and the training of the underlying classifier is based on the dataset CIFAR-10. In the research area of machine learning, CIFAR-10 is a commonly used benchmark dataset of RGB images characterized by a comparatively low image resolution (32×32). The dataset comprises 60,000 images (50,000 training and 10,000 test samples) belonging to the classes *airplane*, *automobile*, *bird*, *cat*, *deer*, *dog*, *frog*, *horse*, *ship* and *truck*, which consist of 6,000 samples each. The available classes are denoted by $c_i \in \mathcal{C}$ for $i \in \{1, \ldots, 10\}$ and thus, $\mathcal{C} = \{c_i \mid i \in \{1, \ldots, 10\}\}$ holds.

In the field of image classification, there are already a considerable amount of CNN architectures that achieve excellent results on CIFAR-10 [4]. These CNNs are generally very deep and equipped with innovative architectural elements (e.g., skip connections [17]). To circumvent potential dependencies on specific architectural design and ensure a straightforward and transparent analysis of relevant features, we deliberately selected a CNN characterized by a more simple and less deep network architecture. The network's architectural design is based solely on fundamental structural elements, such as convolutional, max-pooling, and fully-connected layers, whose configuration is inspired by the state-of-the-art classifier VGG [17] (cf. Table 2). Unlike VGG networks, our CNN architecture features a significantly smaller number of trainable parameters. VGG16 [31], for instance, includes approximately 138 million trainable parameters, while our CNN only consists of 307,936 trainable parameters (cf. Table 2). The CIFAR-10-based training and evaluation of the selected network leads to a training accuracy of 90.87 % and a validation accuracy of 89.02 %. Therefore, 8,902 test images and 47,552 training images are correctly classified.

3.2 Generation of Adversarial Examples

The generation of the adversarial examples is based on 8,902 correctly classified images of the CIFAR-10's test dataset using the *L-BFGS attack* (cf. Sect. 2.2). The L-BFGS attack is an iterative white-box attack based on the *limited memory BFGS method for bound constrained optimization* (short L-BFGS-B), a numerical optimization algorithm described in detail by Byrd et al. [6]. Since the L-BFGS attack is a targeted adversarial attack, the attack's desired target class $c' \in \mathcal{C}$ needs to be specified in advance. To reach a wide variety of adversarial examples for later analysis, every class of CIFAR-10 is chosen once to be the attack's target class. Therefore, the attack is executed nine times for each correctly classified image $X \in \mathbb{R}^n$, i.e., for all targets $c' \in \mathcal{C}$ with $c' \neq y$. In this case, the parameter $y = c$ for $c \in \mathcal{C}$ denotes the image's true label. The implementation of the L-BFGS attack provided by the Python library Foolbox 2.4.0 [11, 26] was used to create the adversarial examples. Further information regarding the algorithmic specification of the attack's implementation can be found in [32].

3.3 Generation of Relevance Scores

The relevance scores for both, adversarial examples and original images (cf. Fig. 1), are determined according to the basic relevance decomposition rule (LRP-0), the ε-rule (LRP-ε), as well as the $\alpha\beta$-rule (LRP-$\alpha\beta$) (cf. Table 1) implemented in Python 3.6.10 using Tensorflow 2.1.0. The advanced decomposition rules LRP-ε and LRP-$\alpha\beta$ are executed for different parameter values, i.e., for $\varepsilon \in \{0.0001, 0.01, 0.1, 1\}$ and $\alpha \in \{1, 2\}$, to evaluate the parameters' effects on the final relevance scores as well. Given the tensorial representation of the input images, the application of the relevance decomposition rules according to Sect. 3.3 results in 3072 relevance scores per adversarial example or original image, respectively.

To allow a clear distinction between relevant and irrelevant input components, especially when passing on off-manifold data like adversarial examples, the softmax pre-activation values are used as initial relevance scores, instead of utilizing the classifier's final probabilistic

Table 2 Detailed specification of the CNN architecture

No.	Layer type	Output dim.	Trainable parameters	
			No. Biases	No. Weights
1	Convolution2D	(32,32,32)	32	864
2	BatchNormalization	(32,32,32)	0	64
3	Activation (ReLU)	(32,32,32)	0	0
4	Convolution2D	(32,32,32)	32	9,216
5	BatchNormalization	(32,32,32)	0	64
6	Activation (ReLU)	(32,32,32)	0	0
7	MaxPooling2D	(16,16,32)	0	0
8	Dropout (0.2)	(16,16,32)	0	0
9	Convolution2D	(16,16,64)	64	18,432
10	BatchNormalization	(16,16,64)	0	128
11	Activation (ReLU)	(16,16,64)	0	0
12	Convolution2D	(16,16,64)	64	36,864
13	BatchNormalization	(16,16,64)	0	128
14	Activation (ReLU)	(16,16,64)	0	0
15	MaxPooling2D	(8,8,64)	0	0
16	Dropout (0.3)	(8,8,64)	0	0
17	Convolution2D	(8,8,128)	128	73,728
18	BatchNormalization	(8,8,128)	0	256
19	Activation (ReLU)	(8,8,128)	0	0
20	Convolution2D	(8,8,128)	128	147,456
21	BatchNormalization	(8,8,128)	0	256
22	Activation (ReLU)	(8,8,128)	0	0
23	MaxPooling2D	(4,4,128)	0	0
24	Dropout (0.4)	(4,4,128)	0	0
25	Flatten	(,2048)	0	0
26	Dense	(,10)	10	20,480
27	Activation (Softmax)	(,10)	0	0

outcome $p_c \in [0, 1]$, $c \in \mathcal{C}$. To avoid incredible large relevance scores and enable comparability while maintaining the relevance scores' ratio and signs within each image, the final relevance scores are normalized separately in dependence of the underlying image using the maximum norm. As opposed to [24], there is no composition of different decomposition rules in dependence of a layer's position within the architecture of the CNN.

3.4 Analysis of the Relevance Scores

The analysis is conducted based on the relevance scores of 77,402 generated adversarial examples, as well as the relevance scores of the correctly classified test images of CIFAR-10 (cf. Sect. 3.3). For our investigations, we applied established methods of descriptive statistics and exploratory data analysis, such as expected values, standard deviations, quantile values and ranges, as well as visual evaluation and verification via histograms and saliency maps.

Table 3 Quantiles of the adversarial examples' relevance scores obtained by applying LRP-0, LRP-ε and LRP-$\alpha\beta$

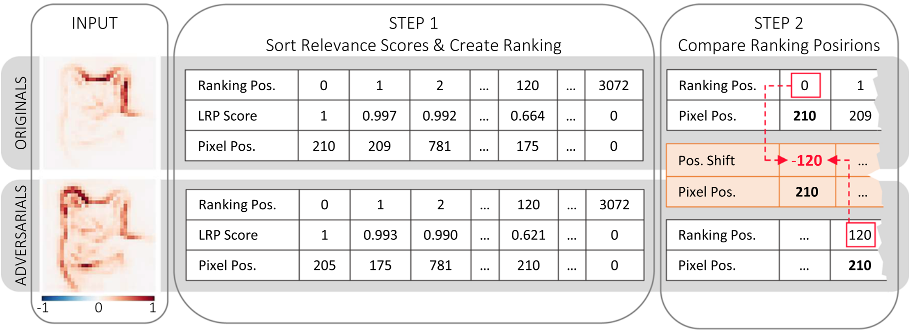
Quantile	Relevance decomposition rule						
	LRP-0	LRP-ε				LRP-$\alpha\beta$	
		$\varepsilon = 0.0001$	$\varepsilon = 0.01$	$\varepsilon = 0.1$	$\varepsilon = 1$	$\alpha = 1$	$\alpha = 2$
0.001	−0.6375	−0.6418	−0.7014	−0.5862	−0.4517	0.0000	−0.3164
0.005	−0.3570	−0.3677	−0.4848	−0.3413	−0.2265	0.0000	−0.1628
0.010	−0.2363	−0.2739	−0.3935	−0.2541	−0.1529	0.0001	−0.1092
0.100	−0.0553	−0.0593	−0.1200	−0.0514	−0.0162	0.0017	−0.0103
0.500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0181	0.0000
0.900	0.0553	0.0593	0.1197	0.0516	0.0197	0.1353	0.0183
0.990	0.2665	0.2775	0.4012	0.2648	0.2604	0.4758	0.2634
0.995	0.3635	0.3746	0.4984	0.3605	0.4005	0.5960	0.4058
0.999	0.6508	0.6569	0.7356	0.6364	0.7581	0.8526	0.7696

Table 4 Quantiles of the original images' relevance scores obtained by applying LRP-0, LRP-ε and LRP-$\alpha\beta$

Quantile	Relevance decomposition rule						
	LRP-0	LRP-ε				LRP-$\alpha\beta$	
		$\varepsilon = 0.0001$	$\varepsilon = 0.01$	$\varepsilon = 0.1$	$\varepsilon = 1$	$\alpha = 1$	$\alpha = 2$
0.001	−0.6565	−0.6555	−0.7047	−0.5982	−0.4414	−0.2043	−0.3217
0.005	−0.3749	−0.3798	−0.4669	−0.3482	−0.2276	−0.0385	−0.1703
0.010	−0.2769	−0.2828	−0.3709	−0.2595	−0.1562	−0.0072	−0.1161
0.100	−0.0581	−0.0611	−0.1055	−0.0520	−0.0174	0.0016	−0.0115
0.500	0.0000	0.0000	0.0000	0.0000	0.0000	0.0189	0.0000
0.900	0.0583	0.0613	0.1001	0.0537	0.0231	0.1486	0.0223
0.990	0.2770	0.2830	0.3743	0.2817	0.3059	0.5140	0.3034
0.995	0.3744	0.3795	0.4714	0.3807	0.4556	0.6335	0.4518
0.999	0.6530	0.6549	0.7089	0.6521	0.7976	0.8729	0.7997

The statistical analysis is performed separately for relevance scores above and below zero, due to the different interpretation of positive and negative relevance scores (cf. Sect. 2.1).

Input components with high positive or particularly contradictory contributions to the classifier's outcome, i.e., components with high absolute relevance scores, are of particular interest in the context of our analyses. If the hypothesis that different classification decisions must lead to distinct saliency maps to provide reasonable explanations holds, there should be a significant discrepancy between adversarial examples and original images, especially in extreme value ranges. Therefore, the focus is on the components with the most extreme relevance scores in each sample, as well as the largest 1 % of the positive relevance scores and the smallest 1 % of the negative relevance scores. This selection is due to the determined quantile values shown in Table 3 and Table 4 . In addition to the analysis of significantly influential components of adversarial examples and original images, non-influential components, i.e., components with relevance scores close to zero, are examined as well. Due to


INPUT
ORIGINALS
ADVERSARIALS
-1
0
1
STEP 1
Sort Relevance Scores & Create Ranking
STEP 2
Compare Ranking Posirions

Ranking Pos.	0	1	2	...	120	...	3072
LRP Score	1	0.997	0.992	...	0.664	...	0
Pixel Pos.	210	209	781	...	175	...	0

Ranking Pos.	0	1	2	...	120	...	3072
LRP Score	1	0.993	0.990	...	0.621	...	0
Pixel Pos.	205	175	781	...	210	...	0

Ranking Pos.	0	1
Pixel Pos.	**210**	209
Pos. Shift	**-120**	...
Pixel Pos.	**210**	...
Ranking Pos.	...	120
Pixel Pos.	...	**210**

Fig. 4 Description of the procedure for generating the relevance ranking based on the relevance scores of original images and corresponding adversarial examples (see input). Step 1: The relevance scores (here LRP Score) are sorted for each image and associated adversarial examples individually in descending order. The pixel with the highest score is ranked first, while the pixel with the lowest score is ranked last. A pixel can be identified by its position in the image frame (here Pixel Pos.). Step 2: The position of each pixel in the original relevance ranking and the adversarial relevance ranking is compared, resulting in a positional relevance shift for each pixel (here Pos. Shift)

their comparatively low relevance scores, components with a positive score below 0.001 and a negative score above −0.001 are assumed to be irrelevant for the classification.

Furthermore, we establish a ranking describing the relevance shift between components of original images and components of adversarial examples triggered by the application of the adversarial attack (cf. Fig. 4). Therefore, the input components of each image are sorted separately and in descending order according to their relevance score, without distinguishing between positive and negative values. The relevance shift of each component is defined by the difference between the position of a component in the relevance ranking based on the original image and the position of a component in the relevance ranking based on the corresponding adversarial example. Hence, a positive shift indicates a positional degradation and a negative shift a positional enhancement of a component when looking at adversarial examples. A shift of zero implies that the position of a component remains unchanged. When analyzing the positional shift of individual components, we focus primarily on the components with the largest or the largest 1 % of the positive relevance scores, similar to the statistical evaluation. Additionally, we investigate the change in position for the most relevant 10 % of the components. In the following, the relevance ranking for components of adversarial examples is referred to as *adversarial relevance ranking* and the ranking based on components of original images is referred to as *original relevance ranking*.

4 Results

4.1 Classification Accuracy

The application of the L-BFGS attack according to the experimental setup sketched above (cf. Sect. 3.2) results in a total of 77,402 adversarial examples which corresponds to a success rate of 96.61 %. The vast majority of adversarial examples, more precisely 99 %, show a classification probability $p_{c'}$ above 93.37 % towards their respective target class $c' \in \mathcal{C}$. Only in 0.5 % of the cases, $p_{c'} \leq 54.82$ % holds. Hence, in these cases, the adversarial attack

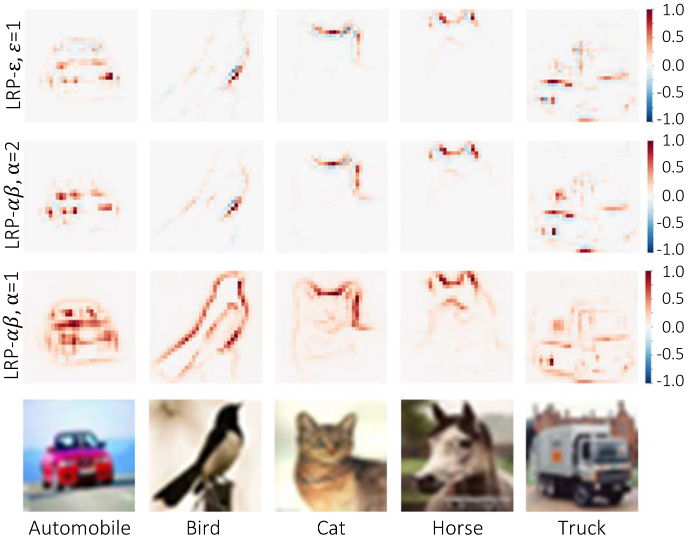


Fig. 5 Results of the application of LRP-ε and LRP-$\alpha\beta$ to correctly classified images of CIFAR-10 belonging to the classes *automobile*, *bird*, *cat*, *horse* and *truck*

is not able to create powerful adversarial examples which convincingly fool the underlying CNN. In 0.1 % of the cases, $p_{c'}$ is even lower than 26.62 %.

A closer look at these low-probability adversarial examples (i.e., $p_{c'} \leq 54.82$ %) revealed, that most of them result from images originally assigned to the classes *dog* and *cat*. It can also be observed that 39 out of 84 low-probability adversarial examples, that result from the L-BFGS attack targeting the class *cat*, are originally assigned to the class *dog*. A similar picture was found while examining the adversarial examples of the target class *dog*. This observation is not surprising since these classes seem to be generally mistaken by the underlying CNN, due to their close visual appearance. Furthermore, it was striking to see that regardless of the image's true label $y = c$ with $c \in \mathcal{C}$ and the attack's target class $c' \in \mathcal{C}$, the remaining probability $1 - p_{c'}$ was typically assigned to one or two other classes at a lower but similar level. In this context, the second highest probability was almost always assigned to the image's true class c. Despite the images' low resolution of 32×32, the adversarial examples cannot be distinguished from their original image by human observers, as Fig. 3 compellingly illustrates.

4.2 Relevance Scores

4.2.1 Visual Evaluation

The visual verification and direct comparison between the input contribution heatmaps of an original image and its corresponding adversarial examples reveal no significant differences (cf. Figs. 5and 6). Despite strongly divergent classification decisions and high classification accuracy, there is almost no difference between components of adversarial examples marked relevant and relevant components of original images. Even though individual pixels undergo minor changes in the absolute magnitude of their relevance scores and some previously insignificant pixels seems to become relevant as a result of the adversarial attack, the majority

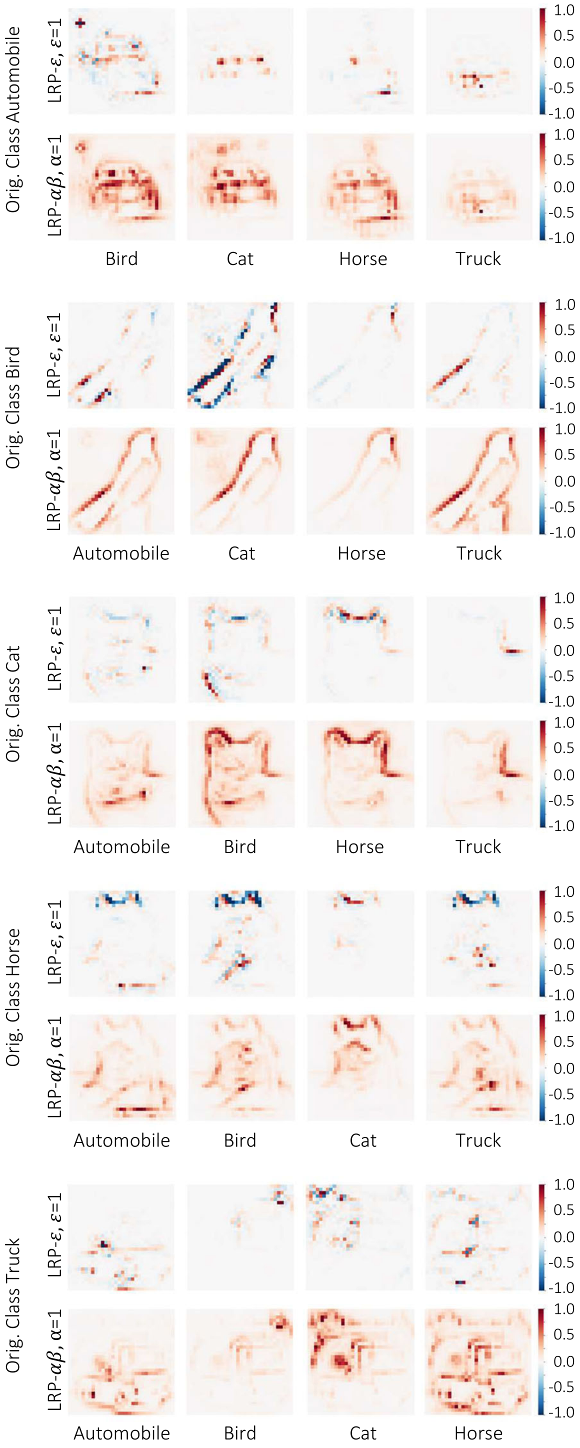


Fig. 6 Results of the application of LRP-ε and LRP-$\alpha\beta$ to adversarial examples belonging to the target classes $c' \in \{automobile, bird, cat, horse, truck\}$

of pixels appear to have a strong impact on both, the classification of the adversarial example and the classification of the original image.

This can be observed especially for input contribution heatmaps derived from relevance scores obtained by applying LRP-$\alpha\beta$ with $\alpha = 1$ (cf. Figs. 5 and 6). Even tough some background components seem to become relevant for the classifier's outcome through the changes induced by the adversarial attack, the contours of the original objects are clearly visible in the input contribution heatmaps of the adversarial examples. This implies that components marked relevant for original images seem to be relevant for adversarial examples as well, albeit leading to a significantly different classification decision with a high accuracy towards the pre-defined target class (here *automobile*, *bird*, *cat*, *horse* and *truck*). In some cases the negative relevance scores even overpower the positive ones (e.g. Fig. 6 original class *bird*, target class *cat*) replicating the original object's contour lines. Thus, the heatmaps seem to clearly contradict the result of the classifier.

Accordingly, a visual verification seems to be ambiguous and not sufficient to explain the entirely different classification results of original images and adversarial examples. Furthermore, the question arises whether a visual verification of relevance scores based on human interpretation of contour lines can actually explain the influence of a component within the complex structure of a deep neural network. However, to allow the evaluation to be based on more than a visual inspection a statistical evaluation of the differences in saliency maps is presented in the following section.

4.2.2 Statistical Evaluation

Regardless of the applied decomposition rule, the statistical evaluation shows that on average 0.36 % of the adversarial components have a relevance score of zero, and therefore are considered non-influential to the final decision score $f(X')$, $X' \in \mathbb{R}^n$. For correctly classified images, on average only 0.24 % of the components have a relevance score of zero. Looking at the quantile values in Tables 3 and 4, only 1 % of the positive and 1 % of the negative relevance scores appear to be significant for the final classification decision. However, the majority of the components seem to have no significant impact according to LRP, as their relevance scores are an order of magnitude lower than those below and above the 1 % and 99 % quantiles, respectively. This is also reflected by the relevance scores' expected value, which ranges from zero (LRP-0) to 0.0516 (LRP-$\alpha\beta$, $\alpha = 1$) for adversarial examples and from zero to 0.055 for original images.

Considering the relevance scores' quantile values, summarized in Tables 3 and 4, there is no discernible difference between the relevance scores of adversarial examples, and the relevance scores of original images. In both cases, the relevance scores obtained by LRP-0 and LRP-ε with $\varepsilon \in \{0.0001, 0.01, 0.1\}$ are symmetrically distributed around zero. The distributions of the relevance scores obtained by LRP-ε with $\varepsilon = 1$ and LRP-$\alpha\beta$ with $\alpha \in \{1, 2\}$, on the other hand, are slightly skewed to the right, which is due to the nature of the applied decomposition rules.

In the case of LRP-ε, the parameter ε absorbs a certain amount of relevance and thus eliminates weak or contradictory contributions as ε grows. Accordingly, with an increasing parameter value the number of irrelevant components increases and only the most salient components survive, which is also reflected by the quantile values in Tables 3 and 4. Furthermore, it can be observed that the relevance scores' standard deviation also declines with growing ε, showing values below 0.1184. Additionally, the gaps between the quantile values of the relevance scores change for $\varepsilon = 1$ and the relevance scores of allegedly influential components tend to become even larger. This seems to allow a more precise distinction of relevant and irrelevant features. In contrast to LRP-ε, the observed distribution shift for the relevance scores obtained by LRP-$\alpha\beta$ is due to the different weighting of positive and negative forward contributions. Especially interesting is the significant difference in the lower quantile values for relevance scores of adversarial components and components of original images for $\alpha = 1$. In the case of original images, 2% of the components are assigned a relevance score less than zero, whereas only 0.1 % of the relevance scores associated with adversarial examples are in the negative value range. Nevertheless, in both cases, the relevance scores have similar expected values (0.0516 for adversarial examples, 0.055 for original images) and standard deviations (0.0930 for adversarial examples, 0.1013 for original images). Similar to LRP-ε the variation of the gap between the relevance scores' quantile values can be observed for LRP-$\alpha\beta$ as well.

Regardless of the applied decomposition rule (i.e., LRP-0, LRP-ε or LRP-$\alpha\beta$), the examination and direct comparison of the relevance scores for both adversarial examples and original images using quantiles, expected values and standard deviations revealed no major differences between their relevance scores. Even the analysis of highly influential or non-influential components showed neither significant differences between the relevance scores of adversarial examples and original images, nor general differences between positive and negative relevance scores. Hence, the statistical analysis indicates a rather ambiguous behavior of LRP as well (cf. Sect. 4.2.1), supporting the conjecture of insufficient explanatory power, especially when considering defective data such as adversarial examples.

4.2.3 Relevance Ranking

The results above are also supported by the established relevance ranking for components of original images and components of adversarial examples, as well as by the direct comparison of their ranking position according to Sect. 3.4. Particularly striking are the results of the relevance ranking for adversarial examples and original images based on the relevance scores obtained by applying LRP-$\alpha\beta$ with $\alpha = 1$.

The examination of the components with the highest relevance score in each original image pursuant to LRP-$\alpha\beta$ with $\alpha = 1$ shows, that 96.72% of these components are also among the 10% of the components most relevant for the classification of the corresponding adversarial example. In 75.21% of the cases, the most relevant component of the original image even belongs to the 1% top-scored components of the associated adversarial example. When looking at the positional shift of an original image's most relevant component, which still belongs to the top 1% or top 10% of the most relevant components in the adversarial ranking, a comparatively small change in position can be observed. For the majority of the components (more precisely 70% of them), the change in position is below 7 when considering the top 1% of the components within the adversarial ranking, and less than 54 when taking the top 10% into account. This observation is also illustrated by Fig. 7. In 3.75% of the cases, the most relevant component of the original image and the most relevant component of the corresponding adversarial example are identical.

Considering the 1% of the original images' most relevant components according to LRP-$\alpha\beta$ with $\alpha = 1$, it can be observed that 43.37% of them also belong to the 1% of the top-ranked components of the corresponding adversarial examples. Approximately 93% of original images' top 1% even belong to the top 10% of the adversarial components mainly responsible for the classifier's outcome. As illustrated by Fig. 7, the absolute positional change of an original component within the adver-sarial relevance ranking is less than 10 in 70% of the cases when looking at the 1% of the top-ranked adversarial components, and less than 54 when considering the top 10%. Of particular interest is the change in position of the most

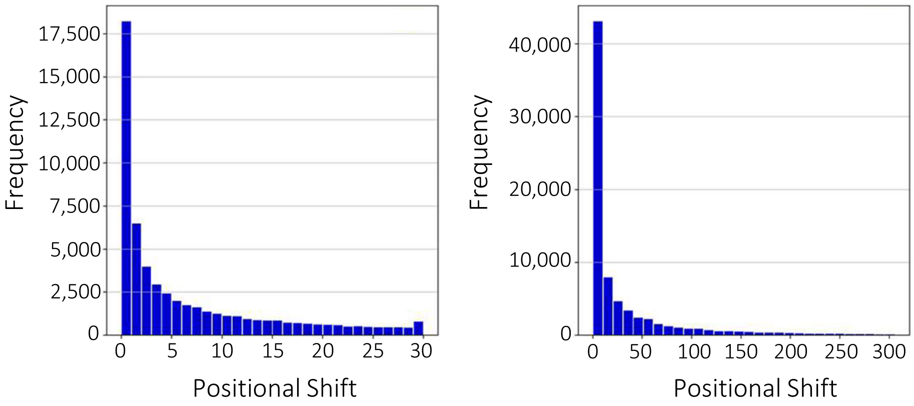


Fig. 7 Frequency distribution of the positional change of the most relevant component of each original image according to LRP-$\alpha\beta$, $\alpha = 1$, which also belongs to the top 1% (left) or the top 10% (right) of the adversarial input components

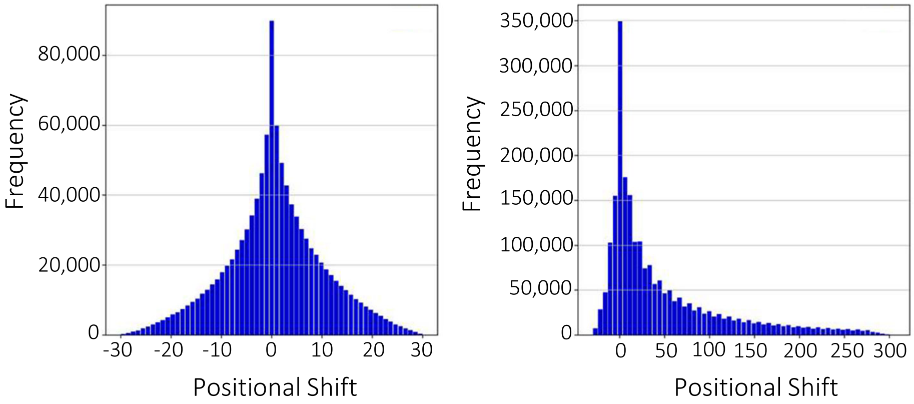


Fig. 8 Frequency distribution of the positional change of the 1% most relevant components of each original image according to LRP-$\alpha\beta$, $\alpha = 1$, which also belong to the top 1% (left) or the top 10% (right) of the adversarial input components

relevant 1% of the original components, which also belong to the 1% of the most relevant components of the respective adversarial examples. Here, an average positional change of zero can be observed. In the case of the components that are given 10% of the highest relevance scores in the original image, 64.12% are also among the most relevant 10% in the adversarial relevance ranking, even though the adversarial examples are classified with an equally high accuracy. In fact, this even applies to strong attacks where the target is very different from the origin, e.g. images originally belonging to the class *bird* vs. their corresponding adversarial examples belonging to the target class *truck*.

Furthermore, the positional change of the components that belong to both, the top 10% of the original images and to the top 10% of the corresponding adversarial examples, averages zero as well (cf. Fig. 9and Fig. 8, left). For 50% of these components, the absolute change in position is below 51. The relevance ranking and positional shift analysis were performed analogously based on relevance scores obtained by LRP-$\alpha\beta$ with $\alpha = 2$ and LRP-ε with $\varepsilon = 1$. The overall tendencies of these results are similar to the results based on relevance

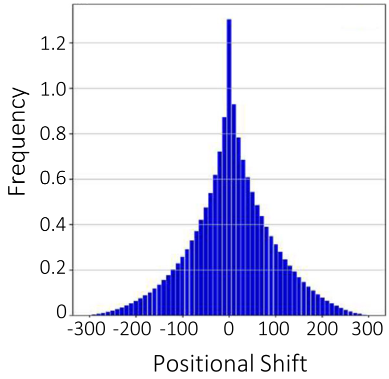


Fig. 9 Frequency distribution of the positional change of the 10% most relevant components of each original image according to LRP-$\alpha\beta$, $\alpha = 1$, which also belong to the top 10% of the adversarial input components

Table 5 Share of the original images' top-ranked components that also belong to the corresponding share of the adversarial examples' top-ranked components based on the relevance scores obtained by LRP-$\alpha\beta$ with $\alpha = 2$ and LRP-ε with $\varepsilon = 1$

Rule	Share orig. images	Share adversarial examples	
		1%	10%
	Most relevant	54.97%	79.93%
LRP-$\alpha\beta$	1%	34.00%	76.03%
	10%	–	52.40%
	Most relevant	44.14%	68.85%
LRP-ε	1%	27.12%	64.18%
	10%	–	42.15%

scores obtained by LRP-$\alpha\beta$ with $\alpha = 1$. Therefore, these results will not be discussed further. However, the main results can be found in Table 5.

Since the majority of top-ranked 1% experience only a marginal change in position of 2.3 per thousand and the majority of the top-ranked 10% merely undergo a relative position change of 1.76%, these top-score shifts between original images and adversarial examples cannot be considered a reliable foundation for explaining the change in classification w.r.t. adversarial examples and original images. Given the fact that the analysis did not discriminate between class affiliations or target class dependencies, these results indicate a general characteristic problem of layer-wise relevance propagation.

4.3 Discussion

Adversarial examples are generally characterized by high similarity to the original data. Therefore, edges in images rarely undergo significant changes in adversarial attacks (cf. Fig. 3). This feature is clearly highlighted by LRP by carving out almost identical contour lines for both the original image and the adversarial examples (cf. Figs. 5 and 6) while they are classified differently. Consequently, LRP emphasizes the image contour lines rather than actually explaining the network's decision. This finding is also supported by the observations of Adebayo et al. [1], who show that some saliency methods (e.g., gradient$\odot$input) work like an edge detector, in combination with the work of Ancona et al. [2], who show that gradient$\odot$input is strongly related to LRP and even equivalent in some configurations. Taking further into account that visual comparison between object contours and saliency maps is a "poor guide in judging whether the saliency map is sensitive to the underlying model"

[1], our results lead to the conclusion that assigning a relevance score to individual input components based on a layer-wise conservation principle to measure their importance in the decision process does not properly explain the behavior of a deep neural network.

However, this rationale is not inconsistent with other evaluations using relevance score-dependent perturbations of input components to analyze the explanatory power of LRP [3, 19, 28]. Since these approaches change components of objects marked relevant, they technically change the edges of these objects which should reduce the classification probability. But this does not explain how the network arrives at its decisions, because in the case of adversarial examples—where components allegedly relevant according to LRP mainly remain unchanged (cf. Figs. 7 and 8)—the classification decision is completely overturned.

5 Conclusion

In this paper, we presented a comprehensive statistical analysis and a novel approach to evaluate the explanatory power of LRP using adversarial examples as relevance score-independent perturbation. The performed analyses demonstrate that there is no significant difference between the saliency maps of adversarial images and the corresponding original ones. This leads to the conclusion that there is no evidence that LRP in its current version explains the CNN’s decision process for original images or adversarial examples in a comprehensible way. Nevertheless, our analyses show that adversarial examples offer the potential to uncover inconsistencies in the robustness and stability of explanations obtained by saliency methods. We believe that adversarial examples are a useful addition to the means of evaluating the explanatory power of such methods.

While our work was a first step in this direction, the presented approach can be used for consistency evaluations of other explainability methods (e.g., LIME [27, 35]) as well.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Adebayo J, Gilmer J, Muelly M, Goodfellow IJ, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: NeurIPS
2. Ancona M, Ceolini E, Öztireli C, Gross M (2018) Towards better understanding of gradient-based attribution methods for deep neural networks. In: ICLR
3. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE. https://doi.org/10.1371/journal.pone.0130140
4. Benchmarks.AI: CIFAR-10 . https://benchmarks.ai/cifar-10 (Last visited: 19.07.2022)
5. Brama H, Grinshpoun T (2020) Heat and blur: an effective and fast defense against adversarial examples. CoRR arXiv:2003.07573
6. Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput. https://doi.org/10.1137/0916069

7. Dieter T, Zisgen H Tools for evaluating the explanatory power of LRP. https://github.com/tamaradi/Evaluation-of-the-Explainatory-Power-of-LRP (Last visited: 02.01.2023). https://doi.org/10.5281/zenodo.7498422
8. Dosovitskiy A, Brox T (2016) Inverting visual representations with convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4829–4837. https://doi.org/10.1109/CVPR.2016.522
9. Došilović FK, Brčić M, Hlupić N (2018) Explainable artificial intelligence: a survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040
10. Erhan D, Bengio Y, Courville AC, Vincent P (2009) Visualizing higher-layer features of a deep network (Technical report)
11. Foolbox: Welcome to Foolbox Native. https://foolbox.readthedocs.io (Last visited: 17.07.2022)
12. Ghorbani A, Abid A, Zou J (2018) Interpretation of neural networks is fragile. In: Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.33013681
13. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89. https://doi.org/10.1109/DSAA.2018.00018
14. Gu J, Tresp V (2019) Saliency methods for explaining adversarial attacks. CoRR arXiv:1908.08413
15. Heo J, Joo S, Moon T (2019) Fooling neural network interpretations via adversarial model manipulation. In: NeurIPS
16. Iwana BK, Kuroki R, Uchida S (2019) Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). https://doi.org/10.1109/ICCVW.2019.00513
17. Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev. https://doi.org/10.1007/s10462-020-09825-6
18. Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, Erhan D, Kim B (2019) The (un)reliability of saliency methods. Springer International Publishing, pp 267–280
19. Lapuschkin S (2018) Opening the machine learning black box with layer-wise relevance propagation. Ph.D. thesis, Technische Universität Berlin. https://doi.org/10.14279/depositonce-7942
20. Li J, Monroe W, Jurafsky D (2016) Understanding neural networks through representation erasure. CoRR arXiv:1612.08220
21. Lipton ZC (2018) The mythos of model interpretability. ACMQueue 16:31–57
22. Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5188–5196. https://doi.org/10.1109/CVPR.2015.7299155
23. Molnar C (2022) Interpretable machine learning, 2 edn
24. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR (2019) Layerwise relevance propagation: an overview. Springer, pp 193–209
25. Nguyen AM, Yosinski J, Clune J (2016) Multifaceted feature visualization: uncovering the different types of features learned by each neuron in deep neural networks. CoRR arXiv:1602.03616
26. Rauber J, Brendel W, Bethge M (2017) Foolbox: a Python toolbox to benchmark the robustness of machine learning models. In: Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning
27. Ribeiro MT, Singh S, Guestrin C (2016) 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. https://doi.org/10.1145/2939672.2939778
28. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR (2017) Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems 28(11):2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820
29. Serban AC, Poll E, Visser J (2018) Adversarial examples - a complete characterisation of the phenomenon. CoRR. arXiv:1810.01185
30. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations (ICLR), Workshop Track Proceedings
31. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Y. Bengio, Y. LeCun (eds.) ICLR
32. Tabacof P, Valle E (2016) Exploring the space of adversarial images. In: 2016 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/IJCNN.2016.7727230
33. Tjoa E, Guan C (2021) A survey on explainable artificial intelligence (XAI): towards medical XAI. IEEE Transactions on Neural Networks and Learning Systems 32:4793–4813

34. Wang X (2016) Deep learning in object recognition detection and segmentation. Found Trends Signal Process 8(4):217–382
35. Xie N, Ras G, van Gerven M, Doran D (2022) Explainable deep learning: a field guide for the uninitiated. J. Artif Intell Res 73:329–396. https://doi.org/10.1561/2000000071
36. Yosinski J, Clune J, Nguyen AM, Fuchs TJ, Lipson H (2015) Understanding neural networks through deep visualization. In: Deep Learning Workshop, 31 st International Conference on Machine Learning
37. Yuan X, He P, Zhu Q, Bhat RR, Li X (2019) Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2018.2886017
38. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: ECCV
39. Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A (2015) Object detectors emerge in deep scene CNNs. In: International Conference on Learning Representations
40. Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.319