Analyzing and Validating Neural Networks Predictions

Alexander Binder[1] ALEXANDER_BINDER@SUTD.EDU.SG
Wojciech Samek[2,5] WOJCIECH.SAMEK@HHI.FRAUNHOFER.DE
Grégoire Montavon[3] GREGOIRE.MONTAVON@TU-BERLIN.DE
Sebastian Bach[2] SEBASTIAN.BACH@HHI.FRAUNHOFER.DE
Klaus-Robert Müller[3,4,5] KLAUS-ROBERT.MUELLER@TU-BERLIN.DE

[1] Information Systems Technology and Design, Singapore University of Technology and Design, 487372, Singapore
[2] Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[3] Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany
[4] Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea
[5] Berlin Big Data Center (BBDC), Berlin, Germany

Abstract

We state some key properties of the recently proposed Layer-wise Relevance Propagation (LRP) method, that make it particularly suitable for model analysis and validation. We also review the capabilities and advantages of the LRP method on empirical data, that we have observed in several previous works.

1. Introduction

Neural networks are known to excel in many fields, such as image recognition (Krizhevsky et al., 2012), object detection (Girshick et al., 2014), video classification (Karpathy et al., 2014), machine translation (Sutskever et al., 2014), reinforcement learning (Koutník et al., 2013; Mnih et al., 2015) among many others. Many published works tackle questions such as how to design neural network architectures capable of solving a particular machine learning problem. Other papers present results on improving training procedures. For all those, performance is usually measured in terms of scores averaged over a large dataset.

Alternatively one can ask what makes neural networks, and predictors in general, arrive at a certain prediction. Given their nonlinearity and deeply nested structure, such an understanding yields a non-trivial problem which has been recently approached, e.g., by computing local gradients (Simonyan et al., 2013), performing a deconvolution (Zeiler and Fergus, 2014), or decomposing the prediction using layer-wise relevance propagation (LRP, Bach et al., 2015). In the next section we briefly introduce LRP and discuss some of it's key properties, that make it particularly suitable for model analysis and validation. In Section 3 we empirically compare the explanations provided by the LRP method to the ones provided by gradient-based sensitivity analysis and the deconvolution approach. Finally, in Section 4 we describe how the LRP method can be used for analyzing and validating machine learning models.

2. Layer-Wise Relevance Propagation

We briefly introduce the layer-wise relevance propagation (LRP) method of Bach et al. (2015) for explaining neural networks predictions. LRP explains a prediction $f(\boldsymbol{x})$ associated to an input $\boldsymbol{x}$ by decomposing it into relevance scores $[f(\boldsymbol{x})]_p$ for each pixel p. These scores indicate how relevant each pixel is for the neural network prediction. The decomposition obeys a conservation principle

$$f(\boldsymbol{x}) = \textstyle\sum_p [f(\boldsymbol{x})]_p \tag{1}$$

that requires that the sum of relevance scores must match the model output. Obviously this simple specification allows for trivial decompositions, e.g. choosing any random distribution over pixels and multiplying it with $f(\boldsymbol{x})$.

LRP forces meaningful decompositions by making use of the structure of the neural network and its parameters: It defines for each neuron in the network a redistribution rule that is conservative and that redistributes the relevance assigned to the neuron onto its input neurons based on to their weighted activations. See (Bach et al., 2015; Montavon et al., 2015) for a definition of these redistribution rules.

Decomposition approaches such as LRP differ from a simple visualization where one finds for a vector of pixels $(x_p)_p$ an associated vector of same dimensions, that visu-

Submitted to the ICML 2016 Workshop on Visualization for Deep Learning.

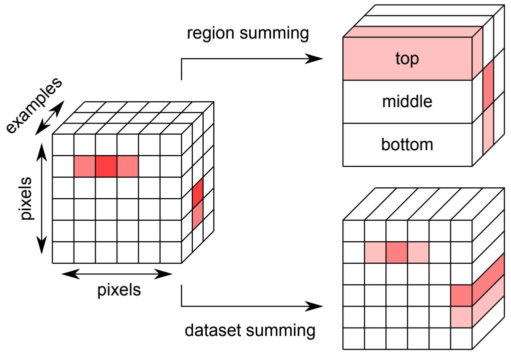

Figure 1. Two possible ways of aggregating the results of an LRP analysis performed on a dataset of images. Red color indicates relevant pixels or regions.

alizes the relevant structure in the image that caused a certain classification. Instead, decompositions favor aggregate analysis, as relevance scores can be meaningfully pooled spatially, or averaged over a dataset. These aggregations produce a coarser decomposition, however they still satisfy a conservation property. Example of possible aggregations are given in Figure 1. For example, we can perform region-wide pooling by considering group of pixels

$$\forall \boldsymbol{x} : [f(\boldsymbol{x})]_{\mathcal{R}} = \textstyle\sum_{p \in \mathcal{R}} [f(\boldsymbol{x})]_p$$

for analysis, with $\mathcal{R} \in \{\text{top}, \text{middle}, \text{bottom}\}$ the different regions of the image. At this coarser level of granularity, the conservation property still holds: $f(\boldsymbol{x}) = [f(\boldsymbol{x})]_{\text{top}} + [f(\boldsymbol{x})]_{\text{middle}} + [f(\boldsymbol{x})]_{\text{bottom}}$. We can also perform dataset aggregation, where we compute expected pixel-wise scores over some data distribution

$$[f]_p = \mathrm{E}[[f(\boldsymbol{x})]_p].$$

Again, a conservation property $\sum_p [f]_p = \mathrm{E}[f(\boldsymbol{x})]$ still holds. These aggregate explanations are used in Section 4.1 for analyzing the use of context by different classifiers.

3. Empirical Evaluation of LRP

In this section we show empirically how the LRP method compares to other methods for explanation, in particular, gradient-based sensitivity analysis, and deconvolution.

In the following we like to show a prototypical difference between methods using gradients for the visualization of regions in an image and decomposition-based methods. A gradient measures an infinitesimal local variation at a certain point rather than a decomposition of a total prediction score. This has implications for the content of a visualization. To see this, we consider explanations provided by three popular techniques, the gradient-based method of (Simonyan et al., 2013), the explanations provided by the Deconvolution approach of (Zeiler and Fergus, 2014) and the recently proposed LRP method for decomposition (Bach et al., 2015).

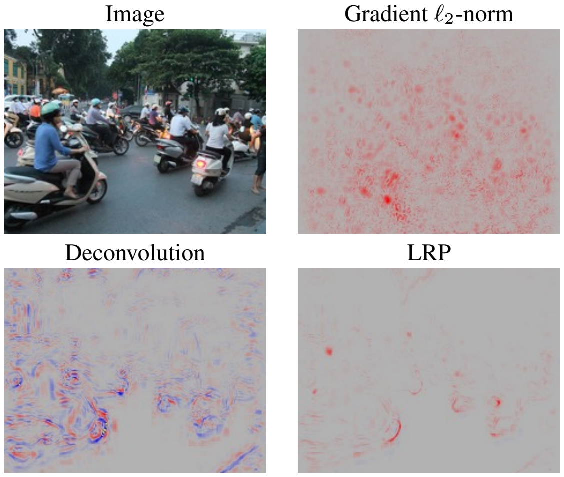

Figure 2. An image of a complex scene and its explanations for the class scooter by different methods. For LRP, we use the parameter $\beta = 1$ described in (Bach et al., 2015).

Figure 2 shows the qualitative difference between norms of backpropagated gradients on the one side and Deconvolution and LRP-type decomposition on the other side. In terms of gradients it is a valid explanation to put high norms on the empty street parts - there exist a direction in parameter space in which the classifier prediction can be increased by putting motor-bike like structures in there. By linearity of the gradient, it implies that the negative direction would decrease the classifier prediction equally strong. As a consequence, regions consisting of pure background may have a notable sensitivity. Deconvolution and LRP on the other side point to real scooter-like structures present in the image. Another aspect can be demonstrated using the noise images from (Nguyen et al., 2014). All methods are able to identify the spurious structures that lead to a prediction of a certain class, as can be seen in Figure 3.

An important question is whether one can measure quantitatively the meaningfulness of visualizations in terms of performance measures over data sets. The work in (Samek et al., 2015) employed perturbation analysis in order to compare visualizations. This is based on three key ideas. Firstly, if we modify a region, that is highly important for the prediction by a classifier, then we expect a steeper decline of the prediction score, compared to the modification of a lesser important, e.g. background region. Secondly, any pixel-wise or region-wise score computed by a visualization method implies an ordering of regions in an image. Thirdly, one can modify or perturb regions in the order implied by a score, measure the decline of the prediction score, and repeat this process with many random perturbations. The average decline over many repetitions can be

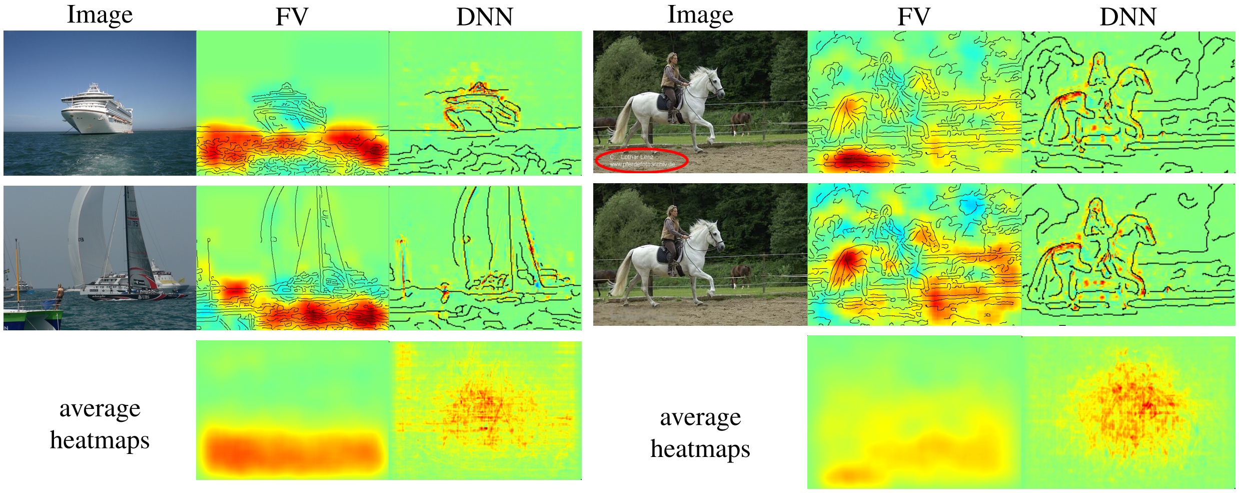

Figure 4. Top: Images of the classes "boat" and "horse", processed by the FV and DNN models and heatmapped using LRP. Bottom: Average heatmap scores over a random sample (of size between 47 and 177) of the distribution for each class and model. On the second image of class "horse", the copyright tag (marked by the red ellipse) has been removed. See (Lapuschkin et al., 2016) for more information.

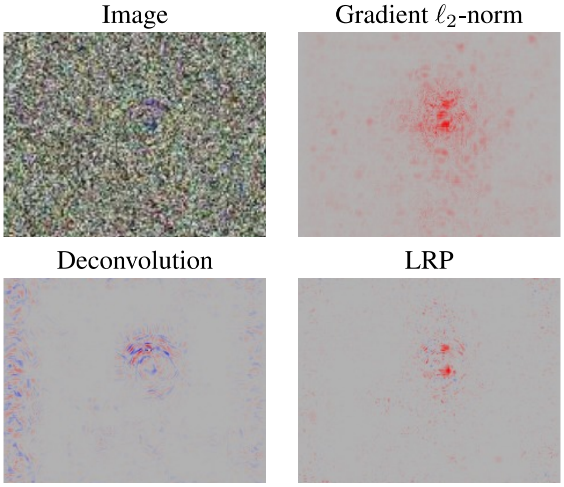

Figure 3. Image from Nguyen et al. (2014) explained by different methods.

reported as a measure of how well a pixel- or region-wise score identifies regions important for a prediction. (Samek et al., 2015) employed simple random draws from a uniform distribution found that deconvolution and LRP is always better than random orderings of regions, which however does not always holds for norms of gradients. This approach can be combined with more advanced region perturbations, e.g. using the seamless image fusion as used in (Zhou et al., 2014) for the image and a proposed perturbation, or (Zintgraf et al., 2016).

4. Model Analysis and Validation

Having introduced the LRP method and demonstrated empirically its advantageous properties, we finally describe how the method can be used for analyzing and validating machine learning models.

4.1. Usage of Context and Artefacts

The usage of explanation methods is not limited to deep neural networks. The work of Lapuschkin et al. (2016) has used LRP to highlight differences between Fisher Vector (FV) based classifiers and deep neural nets. A general observation made in (Lapuschkin et al., 2016) was that the pixel-wise scores can be used together with bounding box ground truth, and measuring the relevance inside the bounding box and the total relevance. The ratio of these pooled relevance scores can be considered as a measure of context usage in classifiers. Deploying this measure of context usage helped to identify a bias in the prediction of boats and horses made by the FV classifier.

Figure 4 is taken from Lapuschkin et al. (2016) and supports two observations. The first observation is the insight that Fisher vectors identify boats primarily by the water. A similarly anomalous second observation was made for the class horses, as shown in Figure 4. The FV classifier performed almost on par with the neural net, however it used for many images a copyright tag as cue to identify horse images. While these two correlations yield a good accuracy on the test set, they are not plausible. This analysis opens up opportunities for building better classifiers by augmenting training data in a more informed manner. Data augmentation schemes, either through artificial transformations, or active sampling, are especially useful when training data is not an abundant resource.

4.2. Comparing Different Neural Networks

The methods above can also be employed to highlight differences between different architectures. As an example, we noticed was that when comparing BVLC CaffeNet (Jia et al., 2014) to GoogleNet (Szegedy et al., 2014), the latter tends to focus more on faces with animals than the former. Figure 5 gives examples for this phenomenon.

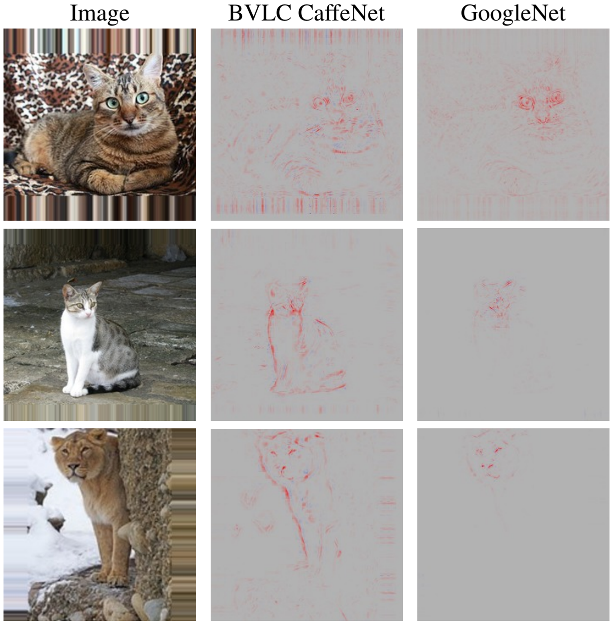

Figure 5. Images of animals with explained predictions of different neural networks (see also Lapuschkin et al. (2016)).

While this seems to be a trivial observation at the first sight, this usage of human understanding can be employed to either augment architectures, or, usually easier, to augment training data for the aim of improving predictions. Note that for animals, their faces are usually more discriminative than their body shapes or furs.

References

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):e0130140, 2015.

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, pages 580–587, 2014.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Int. Conf. on Multimedia*, pages 675–678, 2014.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE CVPR*, pages 1725–1732, 2014.

J. Koutník, G. Cuccu, J. Schmidhuber, and F. J. Gomez. Evolving large-scale neural networks for vision-based reinforcement learning. In *GECCO*, pages 1061–1068, 2013.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. in NIPS*, pages 1106–1114, 2012.

S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *IEEE CVPR*, 2016.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015.

G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *CoRR*, abs/1512.02479, 2015.

A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014.

W. Samek, A. Binder, G. Montavon, S. Bach, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *CoRR*, abs/1509.06321, 2015.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Av. in NIPS*, pages 3104–3112, 2014.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.

B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *CoRR*, abs/1412.6856, 2014.

L. M. Zintgraf, T. S. Cohen, and M. Welling. A new method to visualize deep neural networks. *CoRR*, abs/1603.02518, 2016.