Abstracts

Explaining the decisions of deep neural networks and beyond

Grégoire Montavon and Wojciech Samek
(joint work with Klaus-Robert Müller, Sebastian Lapuschkin, Alexander Binder, Jacob Kauffmann)

Machine learning models have become increasingly complex and this complexity has allowed them to reach high prediction accuracy on challenging datasets. In some cases, improved predictivity has come at the expense of interpretability, in particular, complex models tend to be perceived as black-boxes.

A lack of interpretability is problematic, not only because interpretability is desirable in itself (e.g. to extract useful insights from a model or from the modeled data), but also because common measurements of prediction accuracy can become strongly unreliable when certain assumptions about the training data are not met. Real-world datasets are typically not representative of all possible cases and the truly relevant variables may correlate with other irrelevant variables. In such circumstances, one would need to ensure that the machine learning model does not rely on these irrelevant variables. An assessment based purely on test set accuracy would be oblivious to the exact decision strategy and could overestimate the true prediction performance. This phenomenon has been referred to as the 'Clever Hans' effect [9]. Only an extension of the dataset with specific test cases, or an inspection of the model, e.g. via interpretability techniques [3, 16, 12], is capable of highlighting the improper decision structure.

In this talk, we look at the question of explaining the predictions of *deep neural networks*, a successful machine learning approach that has been used increasingly in real-world applications. A challenge for getting these explanations is the complexity of the decision function, which makes it hard to apply simple explanation methods developed in the context of linear models, e.g. based on first-order Taylor expansions. In particular, DNN decision functions are highly nonlinear and multi-scale, with a gradient that is highly varying or 'shattered' [4]. Also, local searches in the input space easily result in 'adversarial examples' [13] where the prediction no longer corresponds to the observed pattern in the input.

Layer-wise relevance propagation (LRP) [3] is a technique that was proposed to robustly explain the neural network decision in terms of input features. It was shown to work on numerous models in a wide range of applications [14, 5, 15]. LRP departs from the neural network's function representation to consider instead its *graph* structure. Specifically, the LRP algorithm performs an iterative redistribution of the neural network output to the lower layers. Redistribution from each layer to the layer below is achieved by means of propagation rules that satisfy a conservation property analogous to Kirchoff's conservation laws in electrical circuits. The LRP algorithm terminates once the input layer has been reached. The LRP algorithm can be motivated as decomposing a complex problem

(analyzing a highly nonlinear function) into a collection of simpler subproblems (treating each neuron individually).

Furthermore, it was shown that the LRP algorithm can be interpreted as a collection of Taylor expansions performed at each layer and neuron of the neural network [11]. Specifically, the 'relevance' received by a given neuron is approximately the product of the neuron activation and a locally constant term. In turn, the LRP redistribution step can be interpreted as (1) identifying the linear terms of a Taylor expansion of the relevance expressed as a function of activations in the lower layer, and (2) propagating to the lower layer accordingly. A connection can be made between different proposed LRP propagation rules and the choice of reference point at which the Taylor expansion is performed [11, 10]. This Taylor-based view on the LRP algorithm allows in particular to verify that the corresponding reference points are meaningful, for example, that they satisfy domain membership constraints. This interpretation of LRP as a collection of Taylor expansions is referred to as "deep Taylor decomposition" [11].

The LRP algorithm has been successfully applied to various data types and problems, ranging from computer vision and natural language processing tasks such as classification of concepts in images [3], age prediction [8] or categorization of text documents [2], over reinforcement learning tasks such as playing computer games [9], to various medical data analysis tasks, e.g., decoding of fMRI signals [14] or therapy outcome prediction [15]. In these diverse applications, LRP explanations provide additional insights into the decision strategies used by the model, which not only help to better understand the data, including its biases and artifacts [8, 9], but also help to analyze the learning processes and model's decision strategies [9].

In the second part of the talk, two recent advances that broaden the usefulness of explanation methods are discussed. First, Spectral Relevance Analysis (SpRAy) [9], a dataset-wide analysis of individual explanations that summarizes the overall decision structure of the model into a finite and easily interpretable set of prototypical decision strategies. This analysis allows to systematically investigate complex models on large datasets. It has unveiled in commonly used datasets, artifacts, that tend to systematically induce flaws into the decision structure of ML models trained on them. For example, a website logo was found in some images of the class 'truck' of the ImageNet dataset, which the state-of-the-art VGG-16 neural network would then use for its predictions [1].

Another advance brings successful explanation techniques to non-neural network architectures such as kernel-based models. The approach that we term 'neuralization' [6] finds for these non-neural network architectures a functionally equivalent neural network so that state-of-the-art explanation techniques such as LRP can be applied. The approach was successfully applied to various unsupervised models, in particular, kernel one-class SVMs [7] and various k-means clustering models [6], thereby shedding light into what input features make a data point anomalous or member of a given cluster.

Although significant progress has been made to improve the transparency of ML models such as deep neural networks, numerous challenges still need to be addressed both on the methods and theory side. In particular, there is a need for standardized and unbiased evaluation benchmarks for assessing the quality and usefulness of an explanation. Furthermore, an important future work will be to adopt a more holistic view on the problem of explanation, that considers how to make best use of the user's interpretation and feedback capabilities, and that also integrates the end goal of the explanation method, for example, achieving better and more informed decisions, or systematically improving and robustifying a machine learning model.

References

[1] C. J. Anders, T. Marinc, D. Neumann, W. Samek, K.-R. Müller, S. Lapuschkin, *Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed* arXiv:1912.11425 (2019).

[2] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, *"What is Relevant in a Text Document?": An Interpretable Machine Learning Approach*, PLOS ONE, **12**(8):e0181142 (2017).

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, *On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation*, PLOS ONE, **10**(7):e0130140 (2015).

[4] D. Balduzzi, M. Frean, L. Leary, J.P. Lewis, K. Wan-Duo Ma, B. McWilliams, *The Shattered Gradients Problem: If resnets are the answer, then what is the question?* ICML (2017), 342–350.

[5] Y. Ding, Y. Liu, H. Luan, M. Sun, *Visualizing and Understanding Neural Machine Translation*, ACL **1** (2017), 1150–1159.

[6] J. Kauffmann, M. Esders, G. Montavon, W. Samek, K.-R. Müller, *From Clustering to Cluster Explanations via Neural Networks*, arXiv:1906.07633 (2019).

[7] J. Kauffmann, K.-R. Müller, G. Montavon, *Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models*, Pattern Recognition, 107198 (2020).

[8] S. Lapuschkin, A. Binder, K.-R. Müller, W. Samek, *Understanding and Comparing Deep Neural Networks for Age and Gender Classification*, IEEE International Conference on Computer Vision Workshops (2019), 1629–1638.

[9] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, *Unmasking Clever Hans Predictors and Assessing What Machines Really Learn*, Nature Communications **10**:1096 (2019).

[10] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller. *Layer-Wise Relevance Propagation: An Overview*, in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer LNCS 11700, (2019).

[11] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, *Explaining NonLinear Classification Decisions with Deep Taylor Decomposition*, Pattern Recognition **65** (2017), 211–222.

[12] M. Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, KDD (2016), 1135–1144.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing properties of neural networks*, ICLR (2014).

[14] A. Thomas, H. Heekeren, K.-R. Müller, W. Samek, *Analyzing Neuroimaging Data Through Recurrent Deep Learning Models*, Frontiers in Neuroscience, **13**:1321 (2019).

[15] Y. Yang, V. Tresp, M. Wunderle, P. Fasching, *Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks*, ICHI (2018), 152-162.
[16] M. Zeiler, R. Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV **1** (2014), 818–833.

A feature sparse neural network

Robert Tibshirani
(joint work with Ismael Lemhadri, Feng Ruan)

We introduce LassoNet, a neural network model with global feature selection [1]. The model uses a residual connection to learn a subset of the most informative input features. Specifically, the model honors a hierarchy restriction that an input neuron only be included if its linear variable is important. This produces a path of feature-sparse models in close analogy with the lasso for linear regression, while effectively capturing complex nonlinear dependencies in the data. Using a single residual block, our iterative algorithm yields an efficient proximal map which accurately selects the most salient features. On systematic experiments, LassoNet achieves competitive performance using a much smaller number of input features. LassoNet can be implemented by adding just a few lines of code to a standard neural network.

We also apply LassoNet and the linear model lasso to convolution features from mammograms to classify cancer images from normal images. We find that the new methods are nearly as accurate as the state-of–the-art using ResNet, and the results are easier to interpret

References

[1] Ismael Lemhadri, Feng Ruan, Robert Tibshirani *LassoNet: Neural networks with Feature Sparsity*, arXiv:1907.12207

Distributed Machine Learning over Networks

Francis Bach
(joint work with Hadrien Hendrikx Sébastien Bubeck, Laurent Massoulié, Yin-Tat Lee)

The success of machine learning models is in part due to their capacity to train on large amounts of data. Distributed systems are the common way to process more data than one computer can store, but they can also be used to increase the pace at which models are trained by splitting the work among many computing nodes. In this work, we study the corresponding problem of minimizing a sum of functions which are respectively accessible by separate nodes in a network. New centralized and decentralized algorithms are analyzed, together with their convergence guarantees in deterministic and stochastic convex settings, leading to optimal algorithms for this particular class of distributed optimization problems.