Dr. Sebastian Lapuschkin

Opening the machine learning black box with Layer-wise Relevance Propagation

Abstract

Machine learning techniques such as (Deep) Neural Networks are successfully solving a plethora of tasks, e.g. in image recognition and text analysis, and provide novel predictive models for complex physical, biological and chemical systems. However, due to the nested complex and non-linear structure of many machine learning models, this comes with the disadvantage of them acting as a black box, providing little or no information about the internal reasoning. This black box character hampers acceptance and application of non-linear methods in many application domains, where understanding individual model predictions and thus trust in the model's decisions are critically important. In this thesis, we describe a novel method for explaining non-linear classifier decisions by decomposing the prediction function, called Layer-wise Relevance Propagation (LRP). We apply our method to Neural Networks, kernelized Support Vector Machines (with non-linear kernels) and Bag of Words feature extraction pipelines and evaluate LRP theoretically, qualitatively and quantitatively in comparison to other recent methods for interpreting model predictions. Using our method as a tool for comparative analyses between various pre-trained models we reveal different learned prediction strategies and flaws in datasets, predictors and the training thereof.

Zusammenfassung

Techniken des maschinellen Lernens wie (Tiefe) Neuronale Netze lösen eine Vielzahl an Aufgaben mit großem Erfolg, beispielsweise in der Bilderkennung und Textanalyse, und bieten neuartige Vorhersagemodelle für komplexe physikalische, biologische und chemische Zusammenhänge auf. Dies geht jedoch durch die verschachtelte und komplex-nichtlineare Struktur vieler Modelle des maschinellen Lernens mit dem Nachteil einher, dass diese Modelle sich wie Black Boxes verhalten und keine oder nur wenig Informationen über interne Schlussfolgerungen preisgeben. Dieser Black Box-Charakter beeinträchtigt die Anwendung und Akzeptanz von nichtlinearen Methoden in zahlreichen Anwendungsgebieten, in denen das Verstehen individueller Modellvorhersagen, und somit das Vertrauen in das Vorhersagemodell unumgänglich ist. Diese Dissertation behandelt eine neuartige Methode, genannt Layer-wise Relevance Propagation (LRP), zur Erklärung nichtlinearer Klassifikationsentscheidungen mittels der Zerlegung der Vorhersagefunktion. Wir wenden unsere Methode auf Neuronale Netze, Support Vector Maschinen (mit nichtlinearen Kernen) und Bag of Words Merkmalsextraktionssysteme an, und evaluieren LRP auf theoretischer, qualitativer und quantitativer Ebene im Vergleich zu weiteren aktuellen Methoden zur Interpretation von Modellvorhersagen. Unsere Methode als Analysewerkzeug nutzend decken wir vergleichend zwischen diversen vortrainierten Modellen verschiedene erlernte Vorhersagestrategien und Schwächen in Datensätzen, Prädiktionsmodellen und deren Training auf.

Patent Applications

Lapuschkin S, Samek W, Müller K-R, Binder A and Montavon G .
"Relevance Score Assignment for Artificial Neural Networks".
FhG patent ref.: *2014F55840*

Samek W, **Lapuschkin S**, Wiedemann S, Seegerer P, Yeom S-K, Müller K-R, Wiegand T .
"Pruning and/or Quantizing Machine Learning Predictors".
FhG patent ref.: *2019F62193*

Publications

Journal Articles

Alber M, **Lapuschkin S**, Seegerer P, Hägele M, Schütt K T, Montavon G, Samek W, Müller K-R, Dähne S and Kindermans P-J (2019).
"iNNvestigate Neural Networks!".
In: *Journal of Machine Learning Research* 20(93):1-8

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W and Müller K-R (2019).
"Unmasking Clever Hans Predictors and Assessing what Machines Really Learn".
In: *Nature Communications* 10:1069

Horst F, **Lapuschkin S**, Samek W, Müller K-R and Schöllhorn W I (2019).
"Explaining the Unique Nature of Individual Gait Patterns with Deep Learning".
In: *Scientific Reports* 9:2391

Montavon G, **Lapuschkin S**, Binder A, Samek W and Müller K-R (2017).
"Explaining NonLinear Classification Decisions with Deep Taylor Decomposition".
In: *Pattern Recognition* 65:211-222

Samek W, Binder A, Montavon G, **Lapuschkin S**, and Müller K-R (2017).
"Evaluating the Visualization of what a Deep Neural Network has Learned".
In: *IEEE Transactions of Neural Networks and Learning Systems*

Sturm I, **Lapuschkin S**, Samek W and Müller K-R (2016).
"Interpretable Deep Neural Networks for Single-Trial EEG Classification".
In: *Journal of Neuroscience Methods* 274:141-145

Lapuschkin S, Binder A, Montavon G, Müller K-R and Samek W (2016).
"The Layer-wise Relevance Propagation Toolbox for Artificial Neural Networks".
In: *Journal of Machine Learning Research* 17(114):1-5

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

Contributions to Conference Proceedings

Alber M, **Lapuschkin S**, Seegerer P, Hägele M, Schütt K T, Montavon G, Samek W, Müller K-R, Dähne S and Kindermans P-J (2018).
"How to iNNvestigate Neural Networks' Predictors!".
In: *Machine Learning Open Source Software: Sustainable Communities. NIPS Workshop*

Lapuschkin S, Binder A, Müller K-R and Samek W (2017).
"Understanding and Comparing Deep Neural Networks for Age and Gender Classification".

In: *Proceedings of the ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures (AMFG)* 2017:1629-1638

Srinivasan V, **Lapuschkin S**, Hellge C, Müller K-R and Samek W (2017).
"Interpretable Action Recognition in Compressed Domain".
In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2017:1692-1696

Bach S, Binder A, Müller K-R and Samek W (2016).
"Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth".
In: *Proceedings of the IEEE International Conference of Image Processing (ICIP)* 2016:2271-2275

Binder A, Samek W, Montavon G, **Bach S**, and Müller K-R (2016).
"Analyzing and Validating Neural Network Predictions".
In: *Proceedings of the ICML'16 Workshop on Visualization for Deep Learning* . *Best paper award winner*

Lapuschkin S, Binder A, Montavon G, Müller K-R and Samek W (2016).
"Analyzing Classifiers: Fisher Vectors and Deep Neural Networks".
In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016:2912-2920

Montavon G, **Bach S**, Binder A, Samek W and Müller K-R (2016).
"Deep Taylor Decomposition of Neural Networks".
In: *Proceedings of the ICML'16 Workshop on Visualization for Deep Learning* 2016:2912-2920

Samek W, Montavon G, Binder A, **Lapuschkin S** and Müller K-R (2016).
"Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation".
In: *Proceedings of the Interpretable ML for Complex Systems NIPS'16 Workshop*

Book Chapters

Binder A, **Bach S**, Montavon G, Müller K-R and Samek W (2016).
"Layer-wise Relevance Propagation for Deep Neural Network Architectures".
In: *Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering* 276:913-922. Springer Singapore

Binder A, Montavon G, **Lapuschkin S**, Müller K-R and Samek W (2016).
"Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers".
In: *Lecture Notes in Computer Science* 9887:63-71. Springer Berlin/Heidelberg

Preprints

Becker S, Ackermann M, **Lapuschkin S**, Müller K-R and Samek W (2018).
"Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals".
In: *CoRR abs/1807.03418*

Schwenk G and **Bach S** (2014).
"Detecting Behavioural and Structural Anomalies in Media-Cloud Applications".
In: *CoRR abs/1409.8035*