EXPLAINING NONLINEAR CLASSIFICATION DECISIONS WITH DEEP TAYLOR DECOMPOSITIONS *

MARCEL POMMER

Abstract. During the last decade Deep Neural Networks (DNNs) as well as other sophisticated machine learning models gained substantially on relevance, due to so far unreached performance in a variety of topics like image recognition, finance or natural language processing, to name only a few. Despite their great performance, those, mostly non linear models, lack of one important aspect, the explainability of the results. Montavon et al. introduce in their paper *Explaining NonLinear Classification Decisions with Deep Taylor Decompositions* [4] a new technique, the deep Taylor decomposition, to map the relevance of the output on the input features, i.e. quantify the influence of each input variable on the output. They demonstrate the results on two image recognition data sets, the MNIST [2] and the ILSVRC [5], creating heatmaps to display the relevance of each single pixel. I recreate the procedure in Python and apply the deep Taylor decomposition to the titanic dataset [1].

Key words. Explainability, Deep Neural Networks, Image Recognition

MSC codes. 62H35, 93B15

1. Introduction. The raise of machine learning, combined with steadily growing computational power revolutionized many so far hard to grasp tasks like image recognition and are highly used in countless areas like self driving cars and diagnoses of diseases. Those new techniques became quite famous during the last decade due to their over performance in nearly every field, however their complexity makes them mathematically hard to understand and explain, leading to one of their major drawbacks, missing explainability. The paper *Explaining NonLinear Classification Decisions with Deep Taylor Decompositions* [4] by Montavon et al. tries to tackle this problem by extending the explainability of deep neural networks using Taylor expansion, resulting in a mapping of non negative relevance from the output to each input feature. The authors apply their technique to a variety of examples from the MNIST [2] dataset as well as the ILSVRC [5] dataset. Figure 1 shows a tractable example, in which a neural network detects a '0' while distracted by a '5'. We denote the neurons with x_i and the respective contributions with R_i, resulting in a graphical representation, i.e. a heatmap, indicating which pixels contribute with which intensity to the decision of the neural network. Montavon et al. focus on image recognition, but

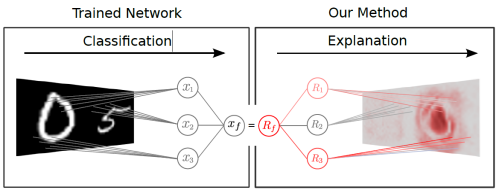


Fig. 1: Example: Detecting 0 with distracting numbers with a neural network

highlight, that the procedure can be broad casted to any input space and feature set. In contrast to the authors I will focus on a non image recognition task, the titanic (survival)[1] dataset which is very tractable and easy to understand and interpret.

*Submitted to professor Gutyniok on the 01.07.2022

2. Main results. The main section summarizes the general idea and presents the definitions and theorems while focusing on image recognition and following the methodology of Montavon et al. In the context of image classification, we define a d-dimensional input space $x \in \mathbb{R}^d$, where the image pixels (p) can be represented as $x = \{x_p\}$. The function $f(x) : \mathbb{R}^d \to \mathbb{R}^+$ quantifies either the probability of an object in the picture or the quantity of the object in question. The aim of the deep Taylor decomposition is to assign a relevance score $R_p(x)$ to each pixel p in the input space. The relevance score quantifies the explanatory power of each pixel, i.e. the higher the relevance score the more important was the pixel for the classification. The result can be displayed in an image or to say heatmap as shown in figure 1 the pixels which led to the classification decision are highlighted. In practice some conditions can help to further define and understand the relevance score. In the context of heatmaps, but also in other cases, the authors state three definitions.

DEFINITION 2.1 (conservative). *A heatmapping $R(x)$ is conservative if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model, that is*

$$\forall x : f(x) = \sum_p R_p(x)$$

In other words, the sum of the relevance of all pixels should align with the output. Definition 2.1 ensures that all relevance detected by the model can be explained by the input variables.

DEFINITION 2.2 (positive). *A heatmapping $R(x)$ is positive if all values forming the heatmap are greater or equal to zero, that is:*

$$\forall x, p : R_p(x) \geq 0$$

This property ensures, that relevance cannot be negative in the sense that two pixels cancel each other out. Since definition 2.1 and definition 2.2 are of essence for the evaluation of models we further define:

DEFINITION 2.3 (consistent). *A heatmapping $R(x)$ is consistent if it is conservative and positive.*

We will use definition 2.3 to evaluate heatmaps, however consistency is not necessarily a measure of quality, which can be seen in the following example of uniform distributed relevance over all d pixels where $\forall p : R_p(x) = \frac{1}{d} \cdot f(x)$. Although, the heatmap will comply with definition 2.3 it will result in an all gray/red image providing no further information on the relation between input and output.

3. Algorithms. The deep Taylor decomposition is based on the first order Taylor expansion at a root point $\tilde{x}$, such that $f(\tilde{x}) = 0$:

$$f(x) = f(\tilde{x}) + \left(\frac{\partial f}{\partial x}\Big|_{x=\tilde{x}}\right)^T \cdot (x - \tilde{x}) + \epsilon = 0 + \sum_p \frac{\partial f}{\partial x_p}\Big|_{x=\tilde{x}} \cdot (x_p - \tilde{x}_p) + \epsilon, \tag{3.1}$$

where the sum over all pixels derivative is defined as the redistributed relevance:

$$R(x) = \frac{\partial f}{\partial x}\Big|_{x=\tilde{x}} \odot (x - \tilde{x}),$$

and $\odot$ is the element wise multiplication. The search of the root point is a great challenge and far from obvious. In figure 2 we can see an image, where the root point is

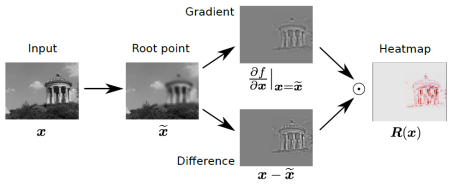


Fig. 2: Root point in an image.

simply a variant of the picture where the building is blurred. Since an image possibly can have more than one root point the choice is crucial and a good root point deviates from the original point x as few as possible, i.e. minimizing the objective:

$$\min_{\xi} ||\xi - x||^2 \text{ subject to } f(\xi) = 0 \text{ and } \xi \in \mathbb{X},$$

where $\mathbb{X}$ is the input domain.
Next we focus on the deep Taylor decomposition, where we consider the mapping of neurons in one layer to each neuron in the next layer, assuming a relation explainable by some relevance function $R_j(\{x_i\})$. If this assumption holds and we further identify a root point $\{\tilde{x}_i\}$ such that $R_j(\{\tilde{x}_i\}) = 0$ we apply the Taylor decomposition layer-wise:

$$\sum_j R_j = \left(\frac{\partial(\sum_j R_j)}{\partial\{x_i\}}\Big|_{\{\tilde{x}_i\}}\right)^T \cdot (\{x_i\} - \{\tilde{x}\}) + \epsilon = \sum_i \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}\}} \cdot (x_i - \tilde{x}_i) + \epsilon$$

If definition 2.3 holds for each local Taylor decomposition, the relevances is guaranteed to be conserved in each layer, i.e. $R_f = ... = \sum_j R_j = ... = \sum_p R_p$ and the relevance propagation rule is positive $R_f, ..., \{R_j\}, ...\{R_p\} \geq 0$.

Further I will present two different approaches, the ω^2-rule and the z-rule. As a starting point we consider a simple detection pooling neural network with a rectified linear activation function, i.e.

$$x_j = \max(0, \sum_i x_i w_{ij} + b_j), \; x_k = \sum_j x_j,$$

where $\{x_i\}$ is a d-dimensional input and $\theta = \{w_{ij}, b_j\}$ are weights and bias. To guarantee the existence of a root point in the origin we restrict $b_j \leq 0$. The relevance of the top layer is due to the pooling $R_k = \sum_j x_j$ and can be redistributed to the next layer according to the Taylor decomposition

$$R_j = \frac{\partial R_k}{\partial x_j}\Big|_{\{\tilde{x}_j\}} \cdot (x_j - \tilde{x}_j)$$

If either $\sum_j \tilde{x}_j = 0$ and $\forall j : \tilde{x}_j \geq 0$ we need to chose $\{\tilde{x}_j\} = 0$ resulting in $R_j = x_j$, since $\frac{\partial R_k}{\partial x_j} = 1$. Redistributing the relevance to the next layer using the Taylor decomposition leads to

$$R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)}) \tag{3.2}$$

which is the starting point for the further analysis and leaves the question of a suitable root point.

3.1. Unconstrained Input Space and ω^2-Rule. Considering an unconstrained input space we search the nearest root point in the Euclidean sense. The intersection of equation $\sum_i \tilde{x}_i^{(j)} w_{ij} + b_j = 0$ (root point) and the line of maximum descent $\{\tilde{x}_i\}^{(j)} = \{x_i\} + t w_j$ (w_j as gradient of R_j), where w_j is the weight vector and $t \in \mathbb{R}$ defines the nearest root point which is then given by $\{\tilde{x}_i\}^{(j)} = \{x_i - \frac{w_{ij}}{\sum_i w_{ij}^2}(\sum_i x_i w_{ij} + b_j)\}$. If we plug in $\{\tilde{x}_i\}^{(j)}$ in equation 3.2 we get the propagation rule for R_i

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$$

PROPOSITION 3.1 (ω^2-Rule consistency). *$\forall g \in G$, the deep Taylor decomposition with the ω^2-rule is consistent in the sense of definition 3.*

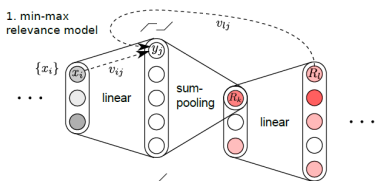
3.2. Constrained Input Space and the z-Rules. Since in many cases the input domain is restricted the authors present a rule for bounded input spaces as well. Montavon et al. consider $\mathcal{X} = \mathbb{R}^d_x$ and $\mathcal{B} = \{\{x_i\} : \forall_{i=1}^d l_i \leq x_i \leq h_i\}$, where $l_i \leq 0$ and $h_i \geq 0$ are the respective lower and higher bounds for each input feature. I will only cover the first case since logic and results are quiet similar and the domain corresponds to the rectified linear activation. Montavon et al. propose the segment $(\{x_i 1_{w_{ij}<0}\}, \{x_i\})$, since we already know of the existence of one root point at the origin and thus the direction of the segment is given by $v_i^{(j)} = x_i - x_i 1_{w_{ij}<0} = x_i 1_{w_{ij}\geq 0}$. If we follow the same logic as in section 3.1 but adjust the line of maximum descent to $\{\tilde{x}_i\}^{(j)} = \{x_i\} + t x_i 1_{w_{ij}\geq 0}$ we get the following relevance propagation rule

$$R_i = \sum_j \frac{x_i 1_{w_{ij}\geq 0}}{\sum_i x_i 1_{w_{ij}\geq 0}} R_j$$

PROPOSITION 3.2 (z-Rule consistency). *$\forall g \in G$, the deep Taylor decomposition with the z-rule is consistent in the sense of Definition 3.*

3.3. Deep Neural Networks. Since many neural networks use complex deep architectures the authors further show a tractable way for the mapping of relevance from higher to the lower layers if the mapping is not explicit and introduce the concept of relevance models. The Min-Max and the Training-Free relevance model are introduced in the paper, however I will only cover the first. We define

$$y_j = \max(0, \sum_i x_i v_{ij} + a_j), \;\; \hat{R}_k = \sum_j y_j,$$



where $a_j = \min(0, \sum_l R_l v_{lj} + d_j)$ is a negative bias where the sum runs over the detection neurons from the upper layer and R_l are the corresponding relevances. After estimation of the parameters $\{v_{ij}, v_{lj}, d_j\}$ by minimizing

$$min\langle(\hat{R}_k - R_k)^2\rangle \tag{3.3}$$

, where R_k and $\hat{R}_k$ are the true and predicted relevances, we end up with the same problem as in section 2. Due to the similar structure we can apply the same computations and derive $R_j = y_j$ and $R_i = \sum_j \frac{q_{ij}}{\sum_i q_{ij}} R_j$, where $q_{ij} \in \{w_{ij}^2, x_i w_{ij} 1_{w_{ij}>0}\}$ for the ω^2-rule and z^+-rule, respectively. In contrast to the problem before the resulting heatmap is only approximately conservative, due to the possible errors during the minimization of 3.3.

4. Experiment. As mentioned in the beginning the authors focus on image recognition and present examples from the MNIST [2] and the ILSVRC [5] data sets. I will only present one example from the MNIST [2] data set and conclude with an example from the titanic [1] data set.

4.1. MNIST Example. Figure 1 shows how the proposed algorithms can detect pixels with a high relevance for detecting a '0' next to a distracting number. Montavon et al. show several examples of other numbers as in figure 3 and compare several methods. It can be seen that the heatmap clearly identifies the important pixels, i.e. correctly assigns relevance to areas where the digit is located. Further analysis

shows, that especially the ω^2- and z-rule are conservative and positive and thus satisfy comply with definition 2.3.

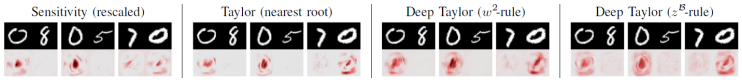


Fig. 3: Example: Detecting number from 0-3 while distracting by number from 4-9

4.2. Titanic Example. Neural networks are not only used for image recognition but for a variety of problems like the prominent titanic survival classification which I chose for my personal analysis. I build a simple neural network with one hidden layer and 5 neurons, a bias $b_j \leq 0$ and a sum pooling as output layer resulting in a accuracy of 79 % on the test data and 84 % on the trainings data. Table 1 summarizes the results for a (very young) passenger which was classified as surviver. The numbers clearly show, that th age has a huge influence on the survival probability, but features like the passenger ID does not.[1]

Table 1: Relevance of Titanic Input Features

feature	ID	Class	Sex	Age	Sibl.	Parch	Fare	C.	Q.	S.
rel. %	0.01	0.81	1.45	48.92	10.63	5.07	22.18	2.86	2.68	5.38

5. Conclusion. The increasing popularity of machine learning and, with it, complex deep neural networks raised the question of explainability, amongst others on context of legal questions for self driving cars or in pricing for insurance. Montavon et al. present in their paper *Explaining NonLinear Classification Decisions with Deep Taylor Decompositions* [4] practical algorithms which can help to tackle the problem of interpretability for a wide range of deep learning models. The authors furthermore substantiate their theory with examples on two very famous image recognition data sets, namely the MNIST [2] and the ILSVRC [5] data sets. I myself could reproduce reliable results on the titanic [1] data set.

REFERENCES

[1] J. P. Eaton, C. A. Haas, and J. Maxtone-Graham, *Titanic: Triumph and tragedy*, Patrick Stephens, Nr. Yeovil, Somerset, 2nd ed. ed., 1998.
[2] Y. LeCun, C. Cortes, and C. Burges, *Mnist handwritten digit database*, ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2 (2010).
[3] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans, *innvestigate neural networks!*, Journal of Machine Learning Research, 20 (2019), pp. 1–8, http://jmlr.org/papers/v20/18-540.html.
[4] G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller, *Explaining nonlinear classification decisions with deep taylor decomposition*, Pattern Recognition, 65 (2017), pp. 211–222, https://doi.org/10.1016/j.patcog.2016.11.008, http://arxiv.org/pdf/1512.02479v1.
[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, *Imagenet large scale visual recognition challenge*, International Journal of Computer Vision, 115 (2015), pp. 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[1] Code can be found on: https://github.com/mpommer/Deep-Taylor-Decomposition-Python