EXPLAINING NONLINEAR CLASSIFICATION DECISIONS WITH DEEP TAYLOR DECOMPOSITIONS
AN ANALYSIS*

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, classification 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* [1] 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 and the ILSVRC, creating heatmaps to display the relevance of the single pixel. I recreate the procedure in Python and apply the deep taylor decomposition to a non image recognition data set, the titanic dataset.

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 of our daily lives like image recognition for self driving cars and diagnoses of deseases. Those new techniques became quite famous during the last decade due to their overperformance in nearly ever field, however their complexity makes them mathematicaly hard to understand and explain, leading to one of their major drawbacks, missing explainability. The paper *Explaining NonLinear Classification Decisions with Deep Taylor Decompositions* [1] 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 [XXX] dataset as well as the ILSVRC [XXX] dataset. In figure 1 one can see 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 contribut with which intensity to the decision of the neural network. Montavon
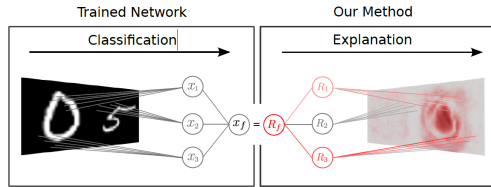

FIG. 1. *Example: Detecting 0 with distrecting numbers with a neural network*

et al. focus on image recognition in their paper, but highlight, that the procedure can be broadcasted to any input space and feature set. In my analysis of the application of the algorithm I will focus on the titanic [XXX] dataset which is very tractable and

easy to understand and interpret.

2. Main results. the main section summarizes the general idea and presents the definitions and theorems while focusing on image recognition and following the metodology of MOntavon. In the context of image classification, we define the d-dimensional input $x \in \mathbb{R}^d$, 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. If the result is plotted in an image or to say heatmap as displayed in figure 1 the pixels which led to the classification decision are highligthed. 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 <u>conservative</u> 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, so the probability or quantity of the object in question. 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 <u>positive</u> 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 a 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 <u>consistent</u> if it is conservative <u>and</u> 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 a uniform relevance over all d pixels where $\forall p : R_p(x) = \frac{1}{d}f(x)$. Although, the heatmap will comply with definition 2.3 it will result in an all black image giving 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 $f(\tilde{x}) = 0$:

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

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

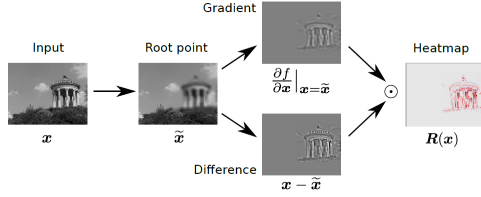$$R(x) = \left. \frac{\partial f}{\partial x} \right|_{x=\tilde{x}} \odot (x - \tilde{x}),$$

FIG. 2. *Example: Root point in an image.*

and $\odot$ is defined as element wise multiplication. The finding 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 simply a variant of the picture where the building is blurrred. 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 obective:

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

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 can identify a root point $\{\tilde{x}_i\}$ such that $R_j(\{\tilde{x}_i\}) = 0$ we can write equation 3.1:

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

If definition 2.3 holds, the relevances is guaranteed to be conserved in each layer, i.e. $R_f = ... = \sum_j R_j = ... = \sum_p R_p$ and $R_f, ..., \{R_j\}, ... \{R_p\} \geq 0$.

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 exististence of a root point in the origin $b_j \leq 0$. 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

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

If either $\sum_j \tilde{x}_j = 0$ and $\forall j : \tilde{x}_j \geq 0$ we need to chose $\{\tilde{x}_j\}$ 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

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

which is the starting point for the further analysis.

3.1. Unconstrained Input Space and ω^2-Rule. If we consider an unconstrained input space one can chose the nearest root point in the Euclidean sense. Considering the rectified linear activation the intersection of the equation $\sum_i \tilde{x}_i^{(j)} w_{i,j} + b_j$ and the line of maximum descent defined by the derivative $\{\tilde{x}_i\}^{(j)} = \{x_i\} + t w_j$, where w_i ist the weight vector and $t \in \mathbb{R}$ define the nearest root point which is then given by $\{\tilde{x}_i\}^{(j)} = \{x_i - \frac{w_{ij}}{\sum_i w_{ij}^2}(\sum_i x_i w_{ij} + b_j)\}$. Pluggin in equation 3.2 leads to

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

PROPOSITION 3.1 (w-Rule consistency). $\forall g \in G$, the deep Taylor decomposition with the w^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 in their paper two possible spaces, first $\mathcal{X} = \mathbb{R}_x^d$ and $\mathcal{B} = \{\{x_i\} : \forall_{i=1}^d l_i \leq x_i \leq h_i\}$, where l_i and h_i 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. Hence we already know of the existence of one root at the origin we search on the segment $(\{x_i 1_{w_{ij}<0}\}, \{x_i\})$ 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. It is defined as

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

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 paramters $\{v_{ij}, v_{lj}, d_j\}$ by minimizing

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

, 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 compute $R_j = y_j$ and $R_i = \sum_j \frac{q_{ij}}{\sum_i q_{ij}} R_j$, where $q_{ij} = \{v_{ij}^2, x_i v_{ij}\}$ for the w^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 and the ILSVRC data sets. I will only present one example from the MNIST data set and conclude with a self made example from the titanic 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 adistracting number. Montavon et all 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 w^2- and z-rule are conservative and positive an thus satisfy the condition for definition 2.3.
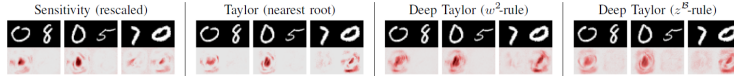

FIG. 3. *Example: Detecting 0 with distrecting other numbers present. Comparison of different methods.*

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 10 neurons resulting in a accuracy of 80 %. Table 1 summarizes the results and clearly shows, that features like sex and age have greater influence on the survival probability than features like the pasenger ID or the port of embarkation.[1]

TABLE 1
Relevance of titanic input features

feature	sex	age	parch	passenger id
relevance in %	10	10	10	10

5. Conclusion. During the last years the importance of explainability of complex machine learning models became more and more present, due to either simple questions of understanding or even leag issues in for example pricing of insurances. Montavon et al. present in their paper *Explaining NonLinear Classification Decisions with Deep Taylor Decompositions* practical algorithms which can help to tackle the problem of interpreteability 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 and the ILSVRC data sets. I myself could also reproduce very reliable results on the titanic data set which is in contrast to the other two a simple supervised binary classification problem.

REFERENCES

[1] 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.

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