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* 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, namely 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 graps tasks like image recognition in order to push the development of self driving cars, help to diagnose deseases or automate classification problems. One can think of a variety of different models like random forests, boosting or deep neural networks and maybe even of more applications in our daily lives, beginning with the advertisement displayed in our browser, or the netflix recommendations going to automated voice recgonition and translation to comunicate with people all over the world. Those new techniques became quite famous during the last decade due to their overperformance in nearly ever field, however due to their complexity, they are not yet fully mathematicaly explained and only few really understand those models in their full depths. This leads to one of the major drawback of those comparably new non linear models like deep neural networks. The paper *Explaining NonLinear Classification Decisions with Deep Taylor Decompositions* [XXX] by Montavon et al. tries to tackle this problem by extending the explainability of deep neural networks using taylor expansion. The main idea is to backpropagate the relevance of the output back to the input features. In the end, each input feature can be assigned a real valued non negative number, expressing its contribution to the output variable. The authors apply their technique to a variety of exmaples from the MNIST [XXX] dataset as well as the ILSVRC [XXX] dataset. The approach results in relevance distribution from the output variable to the input pixels, which can be graphically displayed in a heatmap, mapping relevance on pixels. In figure 1 one can see a tractable example, in which a neural network detects a '0' while distrected by a '5'. The contribution of the single neurons of (all) hidden layers to the output is independently distributed on a backward pass to the input pixels. We denote the neurons with x_i and the respective contributions with R_i. The result is a heatmap, indicating which pixels contributed with which intensity to the decision of the neural network. Montavon et al. focus on image recognition in their paper, but highlight, that the procedure can be broadcasted to any input space and feature set.

*Submitted to professor Gutyniok on the 01.07.2022

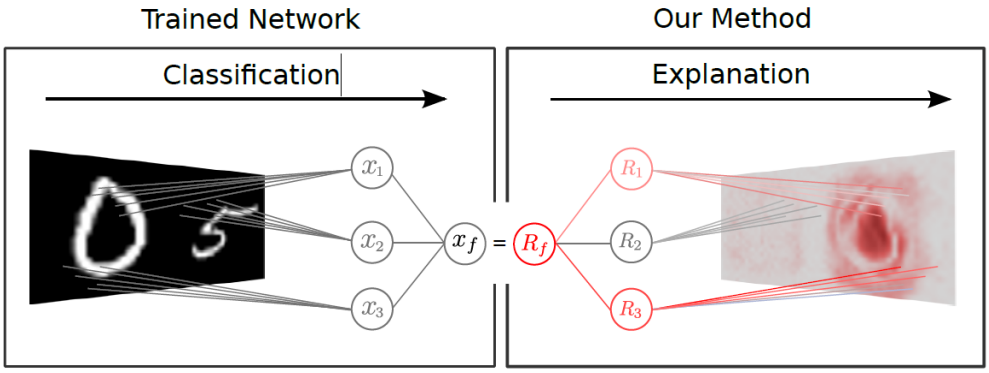

FIG. 1. *Example: Detecting 0 while distrecting other numbers are in place with a neural network*

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.

The paper is organized as follows. The main results are in **??**, the application on a simple neural network and algorithms in section **??**, examples on the MNIST data set and experiments on the titanic dataset in section **??**, and the conclusion follows in section **??**.

2. Main results. In this section I will describe the general idea of the deep taylor decomposition, present the definitions and theorems as well as some handy examples. Since the authors focus on image recognition in their analysis I will follow this methodology, however in section **??** I will transfer all results to a simple regression task. 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. This means, that a value $f(x) > 0$ indicates that either the probability of the obect being in the picture is bigger than 0 or, that at least one occurence of the object was detected. 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. The creation of a heatmap is only possible in the context of image recognition. If, as in section **??**, we have a classical classification problem the output is a table with the feature label and the corresponding relevance. 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 describe three properties.

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, so the probability or quantity of the object in question. Coming back to figure 1, if the output of the neural network defines a probability of 90 % the sum of the

relevance of all pixels should be 90 % or 0.9 as well. Definition **??** ensures that all relevance detected by the model can be explained by the input variables. There is no classification without explainability by the inputs.

DEFINITION 2.2 (positiv). *A heatmapping $R(x)$ is positiv 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. In other words, an input feature either has a positive impact on the desicion made, or no impact at all, but there are no contradictionary evidence. In the following we verify those two properties for resulting heatmaps, which is why we introduce theroem **??**, which defines a heatmap as *consistent* if definition **??** and definition **??** are both fullfilles, or in other word:

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

We will use definition **??** to access the correctness of heatmaps, but it has to be mentioned that many relevance rules might confirm with the definition of *consistency*, however its not a measure for the quality of the algorithm, which can be seen in the following example of a uniformly relevance over all pixels:

$$\forall p : R_p(x) = \frac{1}{d} f(x),$$

where d denotes the number of pixels. Although, the heatmap will comply with definition **??** 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 $f(\tilde{x}) = 0$:

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

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 defined as element wise multiplication. The finding of the root point is agreat challenge an far from obvious. In figure 2 we can see a picture of the monopterus in Munich, 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:

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

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

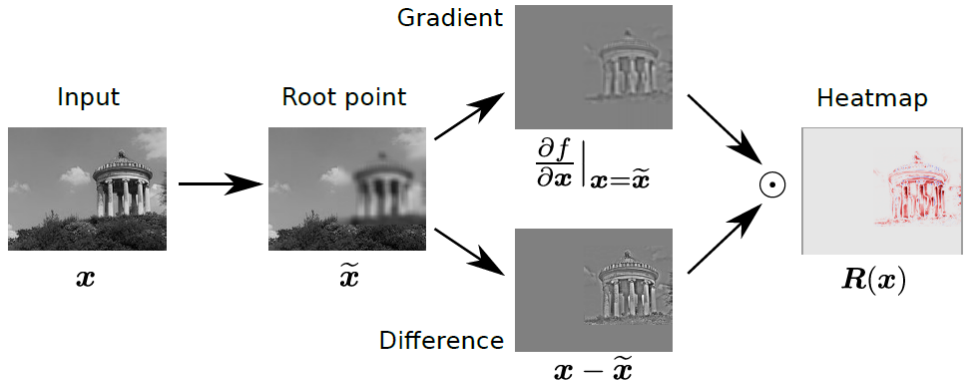

FIG. 2. *Example: Root point in an image.*

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 **??**:

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

If definition **??** 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 reluctified linear activation function, i.e.

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

where $\{x_i\}$ is a d-dimensional input and $\theta = \{w_{i,j}, b_j\}$ are weights and bias. Tu 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

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

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

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

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 $\{\tilde{x}_i\}^{(j)} = \{x_i\} + tw_j$, where w_i ist the weight vector and $t \in \mathbb{R}$. The root point is then given by $\{\tilde{x}_i\}^{(j)} = \{x_i - \frac{w_{ij}}{\sum_i w_{ij}^2}(\sum_i x_i w_{ij} + b_j)\}$ which leads by pluggin in the propagation rule 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 like in the cae of images where usually the value of a pixel is in $[0, 255]$ it is only logical to consider a rule for bounded input spaces as well.Montavon et. al present in their paper two possible spaces, first $\mathbb{X} = \mathbb{R}_x^d$, so the space of positive but unbounded values and $\mathbb{B} = \{\{x_i\} : \forall_{i=1}^d l_i \leq x_i \leq h_i\}$, 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 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 is in section but change 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. Many neural network solutions are highly complicated and use coplex deep architectures. To also apply the rules from **??** and **??** we need to make the mapping from the higher to the lower layers, i.e. a mapping from given neurons to the relevance of a neuron in a higher layer for which the authors introduce 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 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 $min\langle(\hat{R}_k - R_k)^2\rangle$, where R_k and $\hat{R}_k$ are the trueand predicted relevance, we end up with the same problem as in section **??**. So we end up with the same rules and compute $R_j = y_j$ and $R_i = \sum_j \frac{q_{ij}}{\sum_i q_{ij}} R_j$, where $q_{ij} = \{v_{ij}^2, x_i v_{ij}\}$ depending on the model. In contrast to the problem before we are due to solving of the minimization problem only approximately conservative.

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.

5. Conclusion.