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].

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,
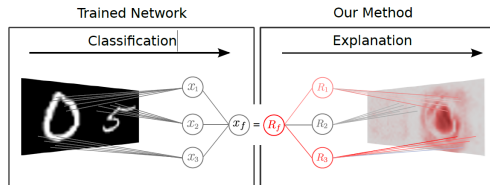

Fig. 1: Example: Detecting 0 with distracting numbers with a neural network

but highlight, that the procedure can be broad casted to any input space and feature set. In my analysis of the application of the algorithm I will focus on the titanic [1] 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 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. 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 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* <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. 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 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* <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/gray 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 $f(\tilde{x}) = 0$:

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

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 a great challenge and far from obvious. In figure 2 we can see an image, where the root point
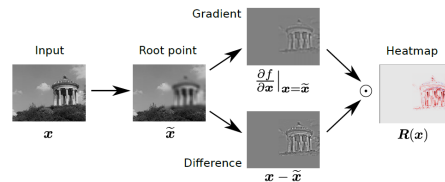

Fig. 2: Root point in an image.

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

85 $$\min_{\xi} ||\xi - x||^2 \text{ subject to } f(\xi) = 0 \text{ and } \xi \in \mathbb{X},$$
86

87 where $\mathbb{X}$ is the input domain.
88 Next we focus on the deep Taylor decomposition, where we consider the mapping of
89 neurons in one layer to each neuron in the next layer, assuming a relation explainable
90 by some relevance function $R_j(\{x_i\})$. If this assumption holds and we further can
91 identify a root point $\{\tilde{x}_i\}$ such that $R_j(\{\tilde{x}_i\}) = 0$ we can write equation 3.1:

92 $$\sum_j R_j = \left(\frac{\partial(\sum_j R_j)}{\partial\{x_i\}}\Big|_{\{\tilde{x}_i\}}\right)^T \cdot (\{x_i\} - \{\tilde{x}\}) + \epsilon = \sum_i \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}\}} \cdot (x_i - \tilde{x}_i) + \epsilon$$
93

94 If definition 2.3 holds, the relevances is guaranteed to be conserved in each layer, i.e.
95 $R_f = ... = \sum_j R_j = ... = \sum_p R_p$ and $R_f, ..., \{R_j\}, ...\{R_p\} \geq 0$.

96

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

100 $$x_j = \max(0, \sum_i x_i w_{ij} + b_j), \ x_k = \sum_j x_j,$$
101

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

106 $$R_j = \frac{\partial R_k}{\partial x_j}\Big|_{\{\tilde{x}_j\}} \cdot (x_j - \tilde{x}_j)$$
107

108 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 $R_j =$
109 x_j, since $\frac{\partial R_k}{\partial x_j} = 1$. Redistributing the relevance to the next layer using the Taylor
110 decomposition leads to

111 (3.2) $$R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)})$$
112

113 which is the starting point for the further analysis.

114 **3.1. Unconstrained Input Space and ω^2-Rule.** If we consider an uncon-
115 strained input space one can chose the nearest root point in the Euclidean sense.
116 Considering the rectified linear activation the intersection of the equation $\sum_i \tilde{x}_i^{(j)} w_{i,j} +$
117 $b_j = 0$ (root point condition) and the line of maximum descent defined due to the
118 rectified linear unit defined as $\{\tilde{x}_i\}^{(j)} = \{x_i\} + tw_j$, where w_j ist the weight vec-
119 tor and $t \in \mathbb{R}$ defines the nearest root point which is then given by $\{\tilde{x}_i\}^{(j)} =$
120 $\{x_i - \frac{w_{ij}}{\sum_i w_{ij}^2}(\sum_i x_i w_{ij} + b_j)\}$. If we plug in $\{\tilde{x}_i\}^{(j)}$ in equation 3.2 we get the propa-
121 gation rule for R_i

122 $$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$$
123

124 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 $\mathcal{X} = \mathbb{R}_x^d$ 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 $v_i^{(j)} = x_i - x_i 1_{w_{ij}<0} = x_i 1_{w_{ij}\geq0}$. If we follow the same logic as in section 3.1 but adjust the line of maximum descent to $\{\tilde{x}_i\}^{(j)} = \{x_i\} + tx_i 1_{w_{ij}\geq0}$ we get the following relevance propagation rule

$$R_i = \sum_j \frac{x_i 1_{w_{ij}\geq0}}{\sum_i x_i 1_{w_{ij}\geq0}} R_j$$

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

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

, 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 derive $R_j = y_j$ and $R_i = \sum_j \frac{q_{ij}}{\sum_i q_{ij}} R_j$, where $q_{ij} = \{w_{ij}^2, x_i w_{ij} 1_{w_{ij}>0}\}$ 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,
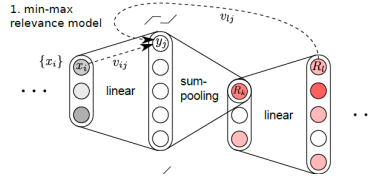
170 that especially the ω^2- and z-rule are conservative and positive an thus satisfy the condition of definition 2.3.
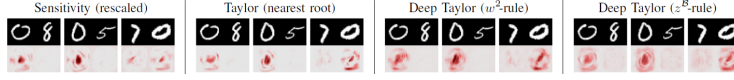

Fig. 3: Example: Detecting 0 with distracting other numbers present.

171

172 **4.2. Titanic Example.** Neural networks are not only used for image recognition
173 but for a variety of problems like the prominent titanic survival classification which
174 I chose for my personal analysis. I build a simple neural network with one hidden
175 layer and 5 neurons, a bias $b_j \leq 0$ and a sum pooling as output layer resulting in a
176 accuracy of 79 % on the test data and 84 % on the trainingsdata. Table 1 summarizes
177 the results for a passenger which was classified as surviver with 0.67 in a ver young age.
178 The numbers clearly shows, that age has greater influence on the survival probability,
179 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

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

189 REFERENCES

190 [1] J. P. Eaton, C. A. Haas, and J. Maxtone-Graham, *Titanic: Triumph and tragedy*, Patrick
191 Stephens, Nr. Yeovil, Somerset, 2nd ed. ed., 1998.
192 [2] Y. LeCun, C. Cortes, and C. Burges, *Mnist handwritten digit database*, ATT Labs [Online].
193 Available: http://yann.lecun.com/exdb/mnist, 2 (2010).
194 [3] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T.
195 Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne,
196 and Pieter-Jan Kindermans, *innvestigate neural networks!*, Journal of Machine Learning
197 Research, 20 (2019), pp. 1–8, http://jmlr.org/papers/v20/18-540.html.
198 [4] G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller, *Explaining nonlinear clas-*
199 *sification decisions with deep taylor decomposition*, Pattern Recognition, 65 (2017), pp. 211–
200 222, https://doi.org/10.1016/j.patcog.2016.11.008, http://arxiv.org/pdf/1512.02479v1.
201 [5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpa-
202 thy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, *Imagenet large scale visual*
203 *recognition challenge*, International Journal of Computer Vision, 115 (2015), pp. 211–252,
204 https://doi.org/10.1007/s11263-015-0816-y.

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