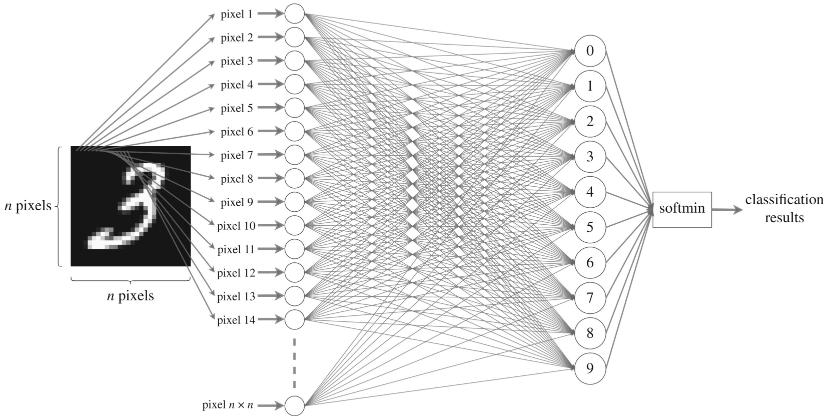
Explaining NonLinear Classification Decisions with Deep Taylor Decomposition analysis

Marcel Pommer

LMU München

3. Juni 2022

Explainability



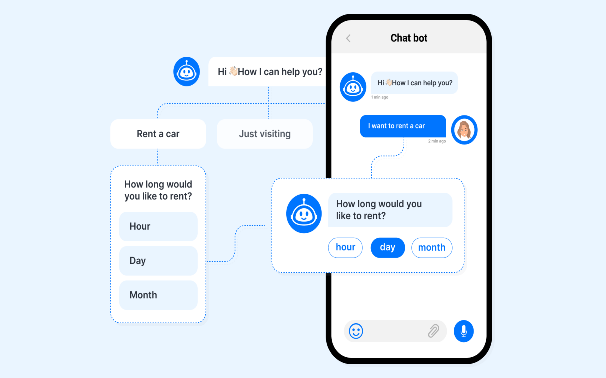
Deep neural networks have great performance on a variety of problems **but** how can we justify decisions made by complex deep architectures?

Table of Contents

Table of Contents

Introduction

Deep neural networks revolutionized amongst others the field of

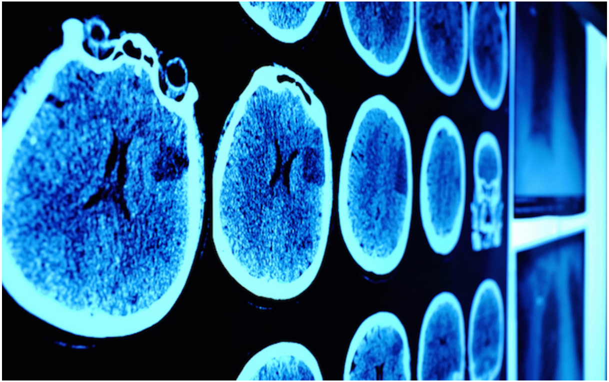


- Image recognition
- Natural language processing
- Human action recognition
- Physics
- Finance
- ...

With one major drawback → **lack of transparency**

Interpretable Classifier

Explanation of non-linear classification in terms of the inputs
$\rightarrow$ A classifier should not only provide a result but also a reasoning



We do not only need to know if the patient has cancer but also where exactly it is located

General Idea

To accomplish the task of explainability we map relevance from the output to the input features

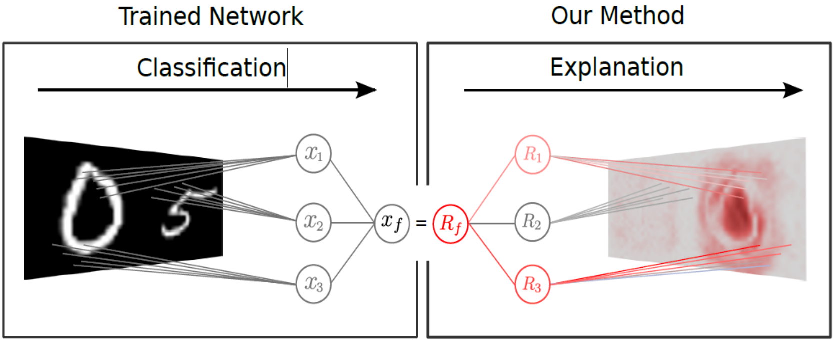


Figure: NN detecting 0 while distrected by 5

Table of Contents

Mathematical Framework

In the context of image classification we define the following mathematical framework

- Positive valued function $f : \mathbb{R}^d \to \mathbb{R}^+$, where the output $f(x)$ defines either the probability that the object is present or the quantity of the object in question

$\rightarrow$ $f(x) > 0$ expresss the presence of the object

- Input $x \in \mathbb{R}^d$, composable in a set of pixel values $x = \{x_p\}$
- Relevance score $R_p(x)$ indicating the relevance of each pixel

$\rightarrow$ The relevance score can be displayed in a heatmap denoted by $R(x) = \{R_P(x)\}$

Definitions

Definition 1

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

$$\forall x : f(x) = \sum_p R_p(x) \tag{1}$$

Definition 2

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

$$\forall x, p : R_p(x) \geq 0 \tag{2}$$

Definitions

Further we test all algorithms for compliance with definition 1 and 2 why we introduce the definition of consistency which forth on is a measure of correctness of a technique

Definition 3

A heatmapping $R(x)$ is <u>consistent</u> if it is *conservative* and *positive*. That is, it is consistent if it complies with Definitions 1 and 2.

However consistency is not a measure of quality which can be seen easily on the following example which complies with definition 3

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

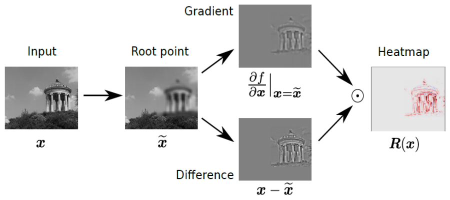
where d denotes the number of pixels

Taylor Expansion

First order taylor expansion at root point $\tilde{x}$

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

The challenge of finding a root point



- Potentially more than one root point
- Remove object but deviate as few as possible

$\rightarrow$ $\min_\xi ||\xi - x||^2$ subject to $f(\xi) = 0$

Sensitivity Analysis

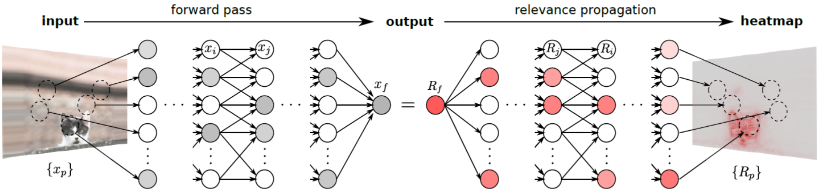
Choose a root point at infinitisimally small distance from the actual point, i.e. $\xi = x - \delta\frac{\partial f}{\partial x}$

If we assume a locally constant function we get

$$
\begin{aligned}
f(x) &= f(\xi) + \left(\frac{\partial f}{\partial x}\bigg|_{x=\xi}\right)^T \cdot (x - (x - \delta\frac{\partial f}{\partial x})) + 0 \\
&= f(\xi) + \delta\left(\frac{\partial f}{\partial x}\right)^T \frac{\partial f}{\partial x} + 0 \\
&= f(\xi) + \sum_p \underbrace{\delta\left(\frac{\partial f}{\partial x}\right)^2}_{R_p} + 0
\end{aligned}
$$

- The heatmap is positive but not conservative
- Measure local effect

Deep Taylor Decomposition

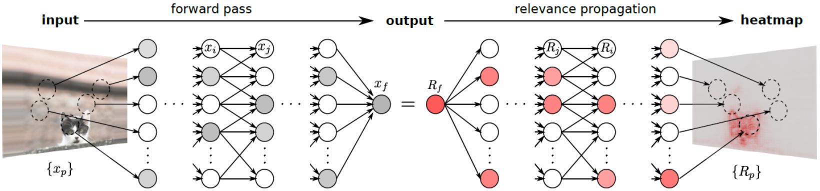


We view each layer as a seperate function and write the Taylor decomposition of $\sum_j R_j$ at $\{x_i\}$ as

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

$$= \sum_i \sum_j \frac{\partial R_j}{\partial x_i}\Bigg|_{\{\tilde{x}\}} \cdot (x_i - \tilde{x}_i) + \epsilon$$

Deep Taylor Decomposition



- If each local Taylor decomposition is *conservative* then the chain of equalitis is also *conservative* (Layer-wise relevance conservation)

$\rightarrow$ $R_f = ... = \sum_i R_i = ... = \sum_p R_p$

- If each local Taylor decomposition is *positive* then the chain of equalitis is also *positive*

$\rightarrow$ $R_f, ..., \{R_i\}, ..., \{R_P\} \geq 0$

- If each local Taylor decomposition is *consistent* then the chain of equalitis is also *consistent*

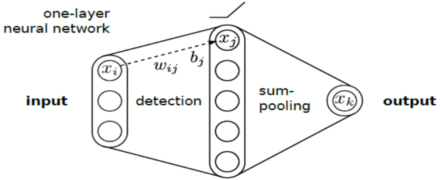
Table of Contents

Setting

Consider a simple detection-pooling one layer neural network with

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

$$x_k = \sum_j x_j, \;\; b_j \leq 0, \forall j$$



1. $R_k = x_k$ and thus $R_k = \sum_j x_j$
2. Chose a root point and redistribute R_k on neurons x_j $\rightarrow R_j = \frac{\partial R_k}{\partial x_j}\Big|_{\{\tilde{x}_j\}} \cdot (x_j - \tilde{x}_j)$, with $\{x_j\} = 0$
3. Since $\frac{\partial R}{\partial x_j} = 1$ we obtain $R_j = x_j$
4. Apply Taylor decomposition another time and get $R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)})$

Derivation of Propagation Rules

Given $R_j = \max(0, \sum_i x_i w_{ij} + b_j)$ and $b_j < 0$ and a search direction $\{v_i\}^{(j)}$ in the input space such that

$$\{\tilde{x}\}^{(j)} = \{x_i\} + t\{v_i\}^{(j)} \Leftrightarrow t = \frac{\tilde{x}_i^{(j)} - x_i}{v_i^{(j)}} \tag{3}$$

If the data point itself is not a root point, i.e. $\sum_i x_i w_{ij} + b_j > 0$ the nearest root along $\{v_i\}^{(j)}$ is given by the intersection of equation (3) and $\sum_i \tilde{x}_i^{(j)} w_{ij} + b_j = 0$ which can be resolved to

$$\begin{aligned} 0 &= \sum_i \{x_i\} w_{ij} + b_j + \sum_i v_i^{(j)} t \\ x_i - \tilde{x}_i^{(j)} &= \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}} v_i^{(j)} \end{aligned}$$

Derivation of Propagation Rules

Starting from the Taylor expansion we can plug in

$$x_i - \tilde{x}_i^{(j)} = \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}} v_i^{(j)}$$

To get

$$R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i^{(j)}\}} \cdot (x_i - \tilde{x}_i^{(j)}) = \sum_j w_{ij} \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}} v_i^{(j)}$$

$$= \sum_j w_{ij} \frac{v_i^{(j)} w_{ij}}{\sum_i v_i^{(j)} w_{ij}} R_j \tag{4}$$

The relevance propagation rule can now easily be calculated by defining

1. Define a segment with search directio $\{v_i\}^{(j)}$
2. The line lies inside the input domain and contains a root point
3. Inject search direction in equation (4)

w^2-rule $\mathcal{X} = \mathbb{R}^d$

Choose root point which is nearest in euclidean sense

1. Search direction $\{v_i\}^{(j)} = w_{ij}$
2. No domain restriction
3. Inject search direction in equation (4)

w^2-rule

$$R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)}) = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$$

w^2-rule $\mathcal{X} = \mathbb{R}^d$

Proposition 1

For all function $g \in G$, the deep Taylor decomposition with the w^2-rule is consistent.

Proof

1. *Conservative*

$$\sum_i R_i = \sum_i \Big(\sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j \Big)$$
$$= \sum_j \frac{\sum_i w_{ij}^2}{\sum_i w_{ij}^2} R_j = \sum_j R_j = \sum_j x_j = f(x)$$

2. *Positive*

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j = \sum_j \underbrace{w_{ij}^2}_{>0} \underbrace{\frac{1}{\sum_i w_{ij}^2}}_{>0} \underbrace{R_j}_{\geq 0} \geq 0$$

z^+-rule $\mathcal{X} = \mathbb{R}_+^d$

Search for a root point on the segment $(\{x_i 1_{w_{ij}<0}\}, \{x_i\}) \subset \mathbb{R}_+^d$

1. Search direction $\{v_i\}^{(j)} = x_i - x_i 1_{w_{ij}<0} = x_i 1_{w_{ij}\geq 0}$
2. If $\{x_i\} \in \mathbb{R}_+^d$ so is the whole domain, further for $w_{ij}^- = \min(0, w_{ij})$

$$\begin{aligned} R_j(\{x_i 1_{w_{ij}<0}\}) &= \max(0, \sum_i x_i 1_{w_{ij}<0} w_{ij} + b_j) \\ &= \max(0, \sum_i x_i w_{ij}^- + b_j) = 0 \end{aligned}$$

3. Inject search direction in equation (4)

z^+-rule

$$R_i = \sum_j \frac{x_i w_{ij}^+}{\sum_{i'} x_{i'} w_{i'j}^+} R_j$$

z^+-rule $\mathcal{X} = \mathbb{R}_+^d$

Proposition 2

For all function $g \in G$ and data points $\{x_i\} \in \mathbb{R}_+^d$, the deep Taylor decomposition with the z^+-rule is consistent.

Proof

If $\sum_i x_i w_{ij}^+ > 0$ the same proof as for the w^2-rule applies, if $\sum_i x_i w_{ij}^+ = 0$ follows that $\forall i : x_i w_{ij} \leq 0$ and

$$R_j = x_j = 0$$

and there is no relevance to redistribute to the lower layers.

z^b-rule $\mathcal{X} = \mathcal{B}$

Often we have a bounded input space $\mathcal{B} = \{\{x_i\} : \forall_{i=1}^{d} l_i \leq x_i \leq h_i\}$ and a segment $(\{l_i 1_{w_{ij}>0} + h_i 1_{w_{ij}<0}\}, \{x_i\}) \subset \mathcal{B}$

1. Search direction $\{v_i\}^{(j)} = x_i - x_i 1_{w_{ij}<0} = x_i 1_{w_{ij}\geq 0}$
2. If $\{x_i\} \in \mathcal{B}$ so is the whole domain, further for $w_{ij}^- = \min(0, w_{ij})$ and $w_{ij}^+ = \max(0, w_{ij})$

$$\begin{aligned} R_j(\{l_i 1_{w_{ij}>0} + h_i 1_{w_{ij}<0}\}) &= \max(0, \sum_i l_i 1_{w_{ij}>0} w_{ij} + h_i 1_{w_{ij}<0} w_{ij} + b_j) \\ &= \max(0, \sum_i l_i w_{ij}^+ + h_i w_{ij}^- + b_j) = 0 \end{aligned}$$

3. Inject search direction in equation (4)

z^b-rule

$$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_{i'} x_{i'} w_{i'j} - l_i w_{i'j}^+ - h_i w_{i'j}^-} R_j$$

z^b-rule $\mathcal{X} = \mathcal{B}$

Proposition 3

For all function $g \in G$ and data points $\{x_i\} \in \mathcal{B}$, the deep Taylor decomposition with the z^b-rule is consistent.

Proof

Since the proof is similar to the proofs of proposition 1 and 2 but lenghty I refer to the literature.

Example MNIST: Setting

Training of a neural network to detect a handwritten digit between 0-3 next to a distractor digit from 4-9 given teh following setting:

- images of size 28 x 56 pixels
- $28x56 = 1568$ input neurons $\{x_i\}$, one hidden layer with 400 neurons $\{x_j\}$ and one output x_k
- weights are random initializated $\{w_{ij}\}$ and bias $\{b_j\}$ is initialized to zero and non negative during training
- Training with 300000 iterations of stochastic gradient descent with a bach size of 20
-

Example MNIST: Heatmaps

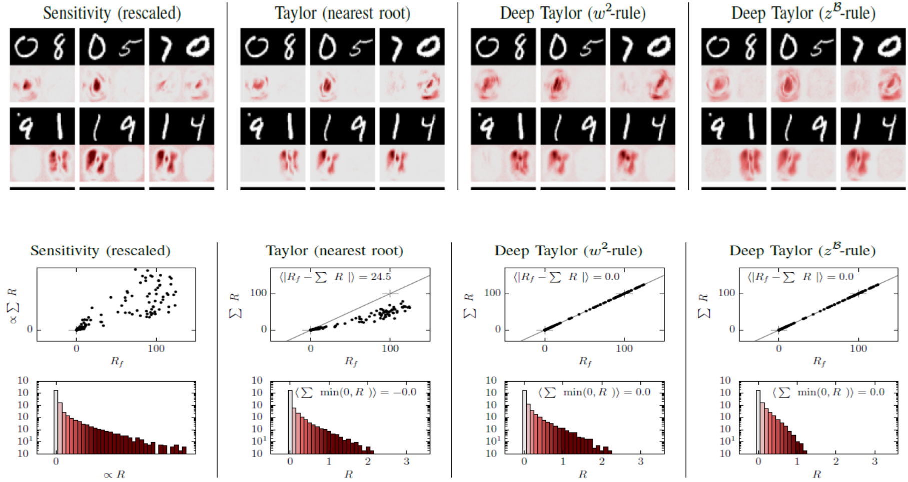


Figure: Heatmap and Empirical Results of Consistency

Table of Contents

Min-Max Relevance Model

TRainable relevnce model designed to incorporate bottom-up and top-down information

$$y_i = \max(0, \sum_i x_i v_{ij} + a_j)$$

$$\hat{R}_k = \sum_j y_j,$$

where $a_j = \min(0, \sum_l R_l V_{lj} + d_j)$ is a negative bias.
$\rightarrow$ Compute $\{v_{ij}, v_{lj}, d_j\}$ by minimizing

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

Deep Networks

Many problems require very complex deep architectures

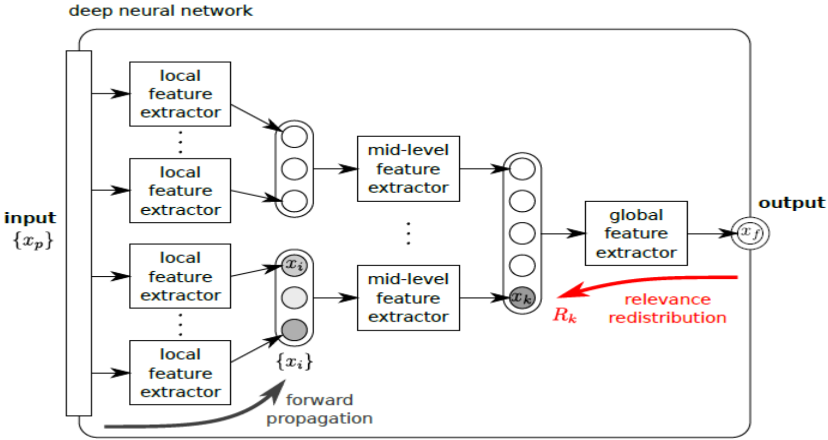


Figure: Example Deep Network

Table of Contents

Table of Contents

Ein alternatives privates Motiv

Das alternative private Motiv: Jeder Agent ist neben dem Allgemeinwohl daran interessiert, dass sich seine Empfehlung ex post als richtig erweist.

Behauptung

Der in Kapitel drei definierte Mechanismus implementiert das PT für jedes Profil von Präferenzen, das strikt wachsend im PT und dem oben definierten privaten Motiv ist.

Beweis

Der Beweis erfolgt weitestgehend analog zum Beweis von Proposition 2. Im Folgenden die Unterschiede:

- Es gilt $\pi_1 \geq V(1)$, aber $\pi_{2,1} < 1$. Trotzdem steht das private Motiv von Agent 1 dem öffentlichen nicht entgegen.

→ Agent 1 maximiert $\pi_{2,1} < 1$, indem er π_1 maximiert, damit wählt er S so informativ wie möglich.

- Analog zum Beweis von Proposition 2 folgt damit $S_{NT} = \varnothing$.
- Mit Hilfslemma 1 folgt ebenfalls, dass $S_c = \varnothing$, da ein Agent durch einen Wechsel von der Strategie "c", zur Strategie "T", $\pi_{2,i}$ von $\frac{1}{2}$ auf p erhöht.
- Da alle Agenten $i \notin S$ ausschließlich daran interessiert sind, dass sich ihre Empfehlung ex post als richtig erweist, spielen sie "T".

→ Damit folgt, dass Agent 1 $S = N \setminus \{1\}$ wählt und alle Agenten in S spielen "T". Agent 1 folgt der Mehrheit und im Fall eines Unentschieden folgt er seinem eigenen Signal.

Table of Contents

Literatur

[GR98] Jacob Glazer and Ariel Rubinstein. Motives and implementation: On the design of mechanism to elicit opinions. Journal of economic Theory, 79(2):157-173, 1998.