

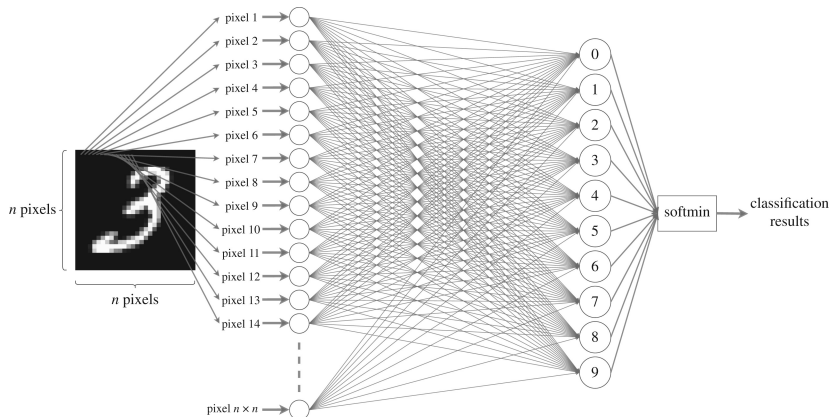
Explaining NonLinear Classification Decisions with Deep Taylor Decomposition analysis

Marcel Pommer

LMU München

29. Mai 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

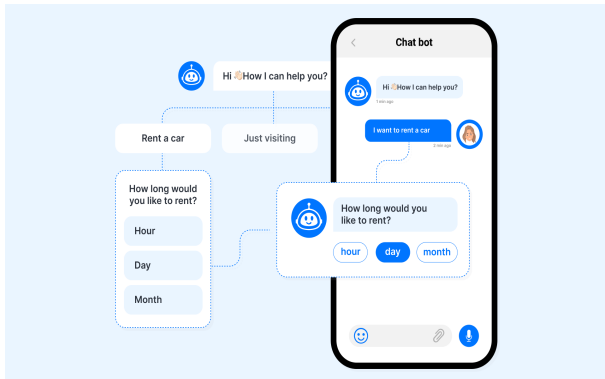
- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z -rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

Table of Contents

- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z -rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

Introduction

Deep neural networks revolutionized amongst others the field of



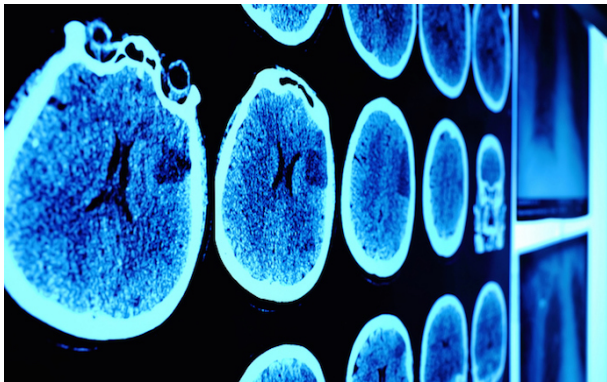
- Image recognition
- Natural language processing
- Human action recognition
- Physics
- Finance
- ...

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

Interpretable Classifier

Explanation of non-linear classification in terms of the inputs

→ 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

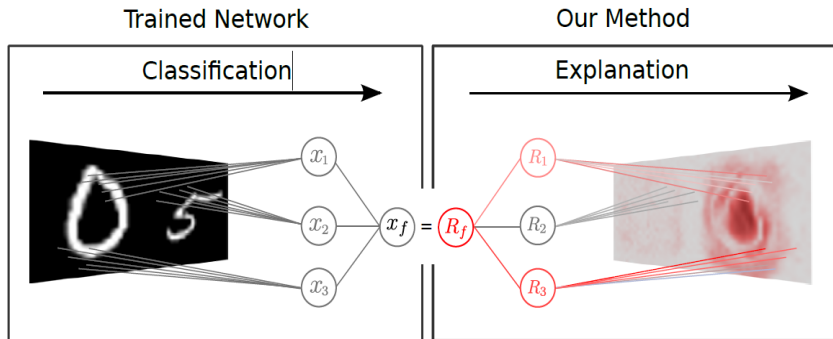


Figure: NN detecting 0 while distracted by 5

Table of Contents

- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z -rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

Mathematical Framework

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

- Positive valued function $f : \mathbb{R}^d \rightarrow \mathbb{R}^+$, where the output $f(x)$ defines either the probability that the object is present or the quantity of the object in question
- $f(x) > 0$ 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
- 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

$$\forall x : f(x) = \sum_p R_p(x) \quad (1)$$

Definition 2

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

$$\forall x, p : R_p(x) \geq 0 \quad (2)$$

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 consistent 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

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

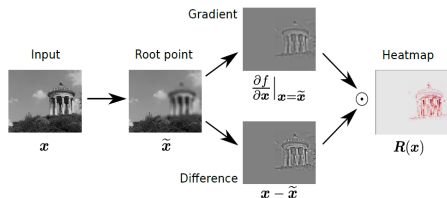
where d denotes the number of pixels

Taylor Expansion

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

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

The challenge of finding a root point



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

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

Sensitivity Analysis

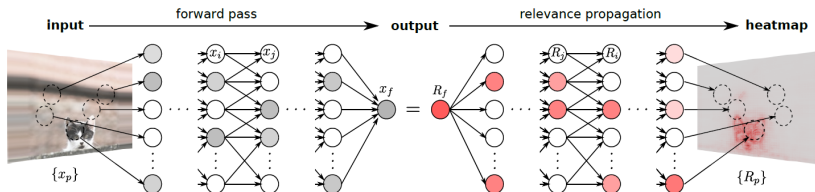
Choose a root point at infinitesimally small distance from the actual point,
i.e. $\xi = x - \delta \frac{\partial f}{\partial x}$

If we assume a locally constant function we get

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

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

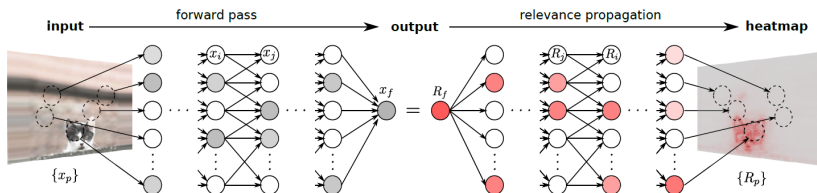
Deep Taylor Decomposition



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

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

Deep Taylor Decomposition



- If each local Taylor decomposition is *conservative* then the chain of equalities is also *conservative* (Layer-wise relevance conservation)
- $R_f = \dots = \sum_i R_i = \dots = \sum_p R_p$
- If each local Taylor decomposition is *positive* then the chain of equalities is also *positive*
- $R_f, \dots, \{R_i\}, \dots, \{R_P\} \geq 0$
- If each local Taylor decomposition is *consistent* then the chain of equalities is also *consistent*

Table of Contents

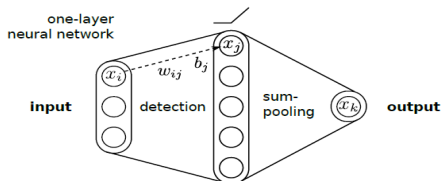
- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z -rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

Setting

Consider a simple detection-pooling one layer neural network with

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

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



- 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

$$\rightarrow R_j = \left. \frac{\partial R_k}{\partial x_j} \right|_{\{\tilde{x}_j\}} \cdot (x_j - \tilde{x}_j), \text{ with } \{x_j\} = 0$$

- ③ Since $\frac{\partial R}{\partial x_j} = 1$ we obtain $R_j = x_j$
- ④ Apply Taylor decomposition another time and get

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

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

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

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

$$0 = \sum_i \{x_i\} w_{ij} + b_j + \sum_i v_i^{(j)} t$$

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

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

$$\begin{aligned} R_i &= \sum_j \frac{\partial R_j}{\partial x_i} \Big|_{\{\tilde{x}_i^{(j)}\}} (x_i - \tilde{x}_i^{(j)}) = \sum_j w_{ij} \frac{\sum_i x_i w_{ij} + b_j}{\sum_i v_i^{(j)} w_{ij}} v_i^{(j)} \\ &= \sum_j w_{ij} \frac{v_i^{(j)} w_{ij}}{\sum_i v_i^{(j)} w_{ij}} R_j \end{aligned} \quad (4)$$

The relevance propagation rule can now easily be calculated by defining

- ① Define a segment with search direction $\{v_i\}^{(j)}$
- 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*

$$\begin{aligned} \sum_i R_i &= \sum_i \left(\sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j \right) \\ &= \sum_j \frac{\sum_i w_{ij}^2}{\sum_i w_{ij}^2} R_j = \sum_j R_j = \sum_j x_j = f(x) \end{aligned}$$

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

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

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

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 lengthy I refer to the literature.

Table of Contents

- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z -rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

Table of Contents

- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z-rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

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 $S_{NT} = \emptyset$.
- Mit Hilfslemma 1 folgt ebenfalls, dass $S_c = \emptyset$, 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

- 1 Introduction
 - Interpretable Classifier
 - General Idea
- 2 Pixel-Wise Decomposition
 - Mathematical Framework and Definitions
 - Methodology
- 3 Application to one-layer networks and root finding
 - w^2 -rule
 - z -rule
 - z^b -rule
- 4 Die Implementierung des PT ist unmöglich, wenn alle Agenten ausschließlich am Allgemeinwohl interessiert sind
- 5 Diskussion: Ein Vorschlag zur Umformulierung des privaten Motivs
 - Ein alternatives privates Motiv
- 6 Literatur

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.