Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition

Grégoire Montavon., et al. (2015)

Presented by Sarah Kim

December 3, 2016
Outline

1. Introduction
2. Pixel-wise decomposition of a function
3. Application to one-layer networks
4. Application to deep networks
1. Introduction

- Deep networks models are successful in terms of performance, but they act like a black box in the sense that it is not clear how and why they arrive at a particular classification decision.

- Interpretable classifier explains its nonlinear classification decision in terms of the inputs.

- View the neural network as a function and obtain decompositions, in a similar way to the error backpropagation algorithm.
1. Introduction

Figure 1: Overview of method. The method produces a pixel-wise heatmap explaining why a neural network classifier has come up with a particular decision. The heatmap is the result of a deep Taylor decomposition of the neural network function.
Outline

1. Introduction

2. Pixel-wise decomposition of a function

3. Application to one-layer networks

4. Application to deep networks
Notations

▶ “pixels” as the input variables denoting $x_p$.

▶ “heatmap”: designate the set of redistributed relevances onto pixels.

▶ $f: \mathbb{R}^d \rightarrow \mathbb{R}^+$ and the input $x \in \mathbb{R}^d$ can be an image.

The image can be decomposed as $x = \{x_p\}$ where $p$ denotes a particular pixel.

▶ $f(x)$ quantifies the presence of a certain type of object(s) in the image.

▶ $R_p(x)$: relevance score, indicating for an image $x$ to what extent the pixel $p$ contributes to explaining the classification decision $f(x)$.

Then heatmap denoted by $R(x) = \{R_p(x)\}$. 
Definitions

Definition 1. A heatmapping $R(x)$ is conservative if

$$f(x) = \sum_{p} R_p(x) \quad \forall x.$$

Definition 2. A heatmapping $R(x)$ is positive if

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

Definition 3. A heatmapping $R(x)$ is consistent if it is conservative and positive.
A. Taylor decomposition

Define a root point $\tilde{x}$ with $f(\tilde{x}) = 0$. Then the first-order Taylor approximation of the function is given as

$$f(x) = f(\tilde{x}) + \left( \frac{\partial f}{\partial x} \bigg|_{x=\tilde{x}} \right)^\top \cdot (x - \tilde{x}) + \varepsilon$$

$$= \sum_p \frac{\partial f}{\partial x_p} \bigg|_{x=\tilde{x}} \cdot (x_p - \tilde{x}_p) + \varepsilon$$

For simplicity, we will consider only the first-order terms for heatmapping, so

$$R_p(x) = \frac{\partial f}{\partial x_p} \bigg|_{x=\tilde{x}} \cdot (x_p - \tilde{x}_p).$$

The heatmap can be written as

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

where “$\odot$” means element-wise product.
A. Taylor decomposition

Figure 2: The construction of a Taylor-based heatmap and a hypothetical function $f$ detecting the presence of objects of class “building” in the image.
A. Taylor decomposition

Choice of the root point \(\tilde{x}\)

- A good root point should selectively remove information from some pixels, while keeping the surroundings unchanged.
- The nearest root \(\tilde{x}\) can be obtained by

\[
\min_{\xi} \|\xi - x\|^2 \quad \text{subject to} \quad f(\xi) = 0 \quad \text{and} \quad \xi \in \mathcal{X},
\]

where \(\mathcal{X}\) is the input domain.
B. Sensitivity analysis

- A special case of Taylor decomposition
- Choose \( \xi = x - \delta \cdot \partial f/\partial x \) with \( \delta \) small.
- Assume the function is locally linear, then we get

\[
\begin{align*}
  f(x) &= f(\xi) + \left( \frac{\partial f}{\partial x} \bigg|_{x=\xi} \right)^\top \cdot \left( x - \left( x - \delta \frac{\partial f}{\partial x} \right) \right) + 0 \\
  &= f(\xi) + \delta \left( \frac{\partial f}{\partial x} \right)^\top \left( \frac{\partial f}{\partial x} \right) \quad \text{by assumption that } f \text{ is locally linear} \\
  &= f(\xi) + \sum_p \delta \left( \frac{\partial f}{\partial x_p} \right)^2
\end{align*}
\]

- \( R_p = \delta \left( \frac{\partial f}{\partial x_p} \right)^2 \geq 0 \)

\( \Rightarrow \) The resulting heatmap is positive, but not conservative.
Outline

1. Introduction

2. Pixel-wise decomposition of a function

3. Application to one-layer networks

4. Application to deep networks
Application to one-layer networks

Network structure

\[ \{x_i, i = 1, \ldots, d\} : \text{input} \]

\[ h_j = \max(0, \sum_{i=1}^{d} x_i w_{ij} + b_j) \text{ for } j = 1, \ldots, n \]

\[ y_k = \sum_{j=1}^{n} h_j \]

\[ (\text{Additional constraint}) \quad b_j \leq 0 \text{ for all } j \]
Application to one-layer networks

To perform the deep Taylor decomposition, we define a amount of total relevance as $R^y_k = y_k$. Then we have $R^y_k = \sum h_j$.

By Taylor decomposition, redistributed relevances $R_j$ can be written as:

$$R_j^h = \frac{\partial R^y_k}{\partial h_j} \bigg|_{h_j \approx h_j} \cdot (h_j - \tilde{h}_j).$$

For a root point $\{\tilde{h}_j\}$, $R^y_k(\{\tilde{h}_j\}) = \sum_{j=1}^n \tilde{h}_j = 0$ and $\tilde{h}_j \geq 0$ for all $j$,

hence $\tilde{h}_j = 0$, $\forall j$.

$$\Rightarrow R_j^h = h_j \quad \text{and} \quad R^y_k = \sum_j h_j = \sum_j R_j^h$$
Application to one-layer networks

Then the redistributed total relevance is

$$\sum_j R_j^h = \sum_j h_j = \sum_j \max(0, \sum_i x_i w_{ij} + b_j).$$

To obtain $\{R_i^x\}$, again by Taylor decomposition on $\sum_j R_j^h$,

$$R_i^x = \sum_{j=1}^n \frac{\partial R_j^h}{\partial x_i} \{\tilde{x}_i\}^{(j)} \cdot (x_i - \tilde{x}_i^{(j)}),$$

where $\{\tilde{x}_i\}^{(j)}$ is a root point for each relevance $R_j^h$. 
Choosing a root point

General scheme

Here we consider $R_j^h = \max(0, \sum_i x_i w_{ij} + b_j)$, where $b_j < 0$ and we search for a root in a particular search direction $\{v_i\}^{(j)}$ in the input space $\mathcal{X}$:

$$\{\tilde{x}_i\}^{(j)} = \{x_i\} + t\{v_i\}^{(j)}$$

Let

$$C_1 = \{ j : \sum_i x_i w_{ij} + b_j \leq 0 \} = \{ j : R_j^h = 0 \}$$

$$C_2 = \{ j : \sum_i x_i w_{ij} + b_j > 0 \} = \{ j : R_j^h > 0 \}$$

- For $j \in C_1$, $\tilde{x}_i^{(j)} = x_i$
- For $j \in C_2$, solve $\{\tilde{x}_i\}^{(j)} = \{x_i\} + t\{v_i\}^{(j)}$ with $\sum_i \tilde{x}_i^{(j)} w_{ij} + b_j = 0$. Then we get

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

General scheme

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

\[ = \sum_{j \in C_1} \frac{\partial R_j^h}{\partial x_i} \bigg|_{\{\tilde{x}_i\}} \cdot 0 + \sum_{j \in C_2} w_{ij} \cdot \frac{\sum_i x_i w_{ij} + b_j v_i^{(j)}}{\sum_i v_i^{(j)} w_{ij}} \]

\[ = \sum_j \frac{v_j^{(j)} w_{ij}}{\sum_i v_i^{(j)} w_{ij}} R_j^h \]
Choosing a root point

1. Unconstrained input space (\( \mathcal{X} = \mathbb{R}^d \)) and the \( w^2 \)-Rule:
   
   we can choose \( \{x_i\}^{(j)} \) that is nearest in the Euclidean sense to the actual data point \( \{x_i\} \). Then \( v_i^{(j)} = w_{ij} \), hence
   
   \[
   R_i^x = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j^h.
   \]

2. Constrained input space (\( \mathcal{X} \subseteq \mathbb{R}^d \)) and \( z \)-Rules:
   
   restrict the search domain to a segment (or subset) of \( \mathcal{X} \) and choose the nearest root of \( R_j^h \) on the segment.
   
   ➤ If \( \mathcal{X} = \mathbb{R}^d \), \( R_i^x = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^h \), where \( z_{ij} = x_i w_{ij} \).
   
   ➤ If \( \mathcal{X} = \mathbb{R}^d_+ \), \( R_i^x = \sum_j \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} R_j^h \), where \( z_{ij}^+ = x_i w_{ij}^+ \).
   
   ➤ If \( \mathcal{X}(= \mathcal{B}) = \{ \{x_i\} : \forall_{i=1}^d l_i \leq x_i \leq h_i \} \), where \( l_i \leq 0 \) and \( h_i \geq 0 \),
   
   \[
   R_i^x = \sum_j \frac{z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_{i'} z_{i'j} - l_i w_{i'j}^+ - h_i w_{i'j}^-} R_j^h.
   \]
Choosing a root point

Figure 3: Illustration of root points (empty circles) found for a given data point (full circle).
Training to predict whether a MNIST handwritten digit of class 0-3 is present in the input image, next to a distractor digit of a different class 4-9.

\[
\text{output} = \begin{cases} 
0 & \text{if there is no digit to detect in the image,} \\
100 & \text{if there is one.} 
\end{cases}
\]

We minimize the mean-square error between the true scores \( \{0, 100\} \), and the neural network output \( y_k \).

- The input image is of size \( 28 \times 56 \) pixels and is coded between \(-0.5\) (black) and \(+1.5\) (white).
- (Dimensions) Input: \( 28 \times 56 \), Hidden: \( 400 \times 4 \), Output: \( 4 \) \( (y_0, y_1, y_2, y_3) \)
- For the \( z^B \)-rule, \( l_i = -0.5 \) and \( h_i = 1.5 \) for all \( i \).
Figure 4: Comparison of heatmaps produced by various decompositions.
Figure 5: Quantitative analysis of each demoposition technique.
Outline

1. Introduction

2. Pixel-wise decomposition of a function

3. Application to one-layer networks

4. Application to deep networks
Training-free relevance model

A relevance model is a function that maps a set of neuron activations at a given layer to the relevance of a neuron in a higher layer and whose output can be redistributed onto its input variables.
Training-free relevance model

We consider

\[ h_j = \max(0, \sum_i h_i^L w_{ij} + b_j) \]

\[ h_k^U = \|\{h_j\}\|_p, \]

where \(\|\cdot\|_p\) can represent a variety of pooling operation such as sum-pooling or max-pooling. Assume that the upper-layer has been explained by the \(z^+\)-rule,

\[
R_k^U = \sum_l \frac{h_k^U w_{kl}^+}{\sum_{k'} h_{k'}^U w_{k' l}^+} R_l
\]

\[
= (\sum_j h_j) \cdot \frac{\|\{h_j\}\|_p}{\|\{h_j\}\|_1} \cdot \sum_l \frac{w_{kl}^+ R_l}{\sum_{k'} h_{k'}^U w_{k' l}^+}
\]

\[
= (\sum_j h_j) \cdot c_k \cdot d_k(\{R_l\}),
\]

where the pooling ratio \(c_k > 0\) and the top-down term \(d_k(\{R_l\}) > 0\).
Training-free relevance model

The training-free relevance model has the same structures as the network in Sec 3, by Taylor decomposition, we get

\[
R_j = \frac{h_j}{\sum_{j'} h_{j'}} R_k^U
\]

\[
R_i^L = \sum_j \frac{q_{ij}}{\sum_{i'} q_{i'j}} R_j,
\]

where \( q_{ij} = w_{ij}^2 \), \( q_{ij} = h_i^l w_{ij}^+ \), \( q_{ij} = h_i^l w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^- \), if choosing the \( w^2- \), \( z^+ - \), \( z^3 - \) rules respectively.
Experiment on MNIST
Experiment on ILSVRC
Experiment on ILSVRC