Explaining Predictions of Non-Linear Classifiers in NLP

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek

Explaining Classifier Predictions

input (e.g., image or text)

$x_d$  

relevance (one value per input dimension)

$R_d$  

prediction score for a class

$f(x)$  

predict

$\text{Predict}$

$\text{Explain}$

$R$  

Explaining Neural Network Predictions

LRP

Layer-wise Relevance Propagation

Layer $0$, Layer $1$, Layer $L-1$, Layer $L$

Model: word-based CNN

$\text{CosNet} \Rightarrow \text{Bal2D} \Rightarrow \text{max-Pool}\Rightarrow \text{FC}$ horizontal concat. of $d$-dim pre-trained word-embeddings ($w_1, w_2, \ldots, w_d$) as input

1-dim convolutional operator covering the entire word-embedding space

Experimental Setup

Task: Document classification

$20\, \text{Newspaper dataset} \ (113\, \text{short})$ (transcribed documents)

newspaper subcategories into 20 fine-grained categories

Document Highlighting

Quantitative Eval.: Word Deleting

SA

Sensitivity Analysis

or Gradient-Magnitude

Word-level Relevance w.r.t. our CNN Model

$R(w_t) = \sum_{i=1}^{D} R_{i,t}$

Backpropagation

$R_0 = f(x)$

Properties:

1) layer-wise conservation of score:

$\sum R_0 = \ldots = \sum R_{i} = \sum R_{i} = f(x)$

2) signed relevance: indicate input regions that support or inhibit a specific classification decision

3) relevance indicates contribution of input dim to actual classification decision (static)

Intuition

$f(x) = \max(0, x_1) + \max(0, x_2)$

From word-vectors to document-vectors

SA

$R_0 = R_2$

LRP

$R_1 \approx R_2$

SUM

$R_1 \approx 3 \cdot R_2$

SA w.v.

LRP w.v.

LRF w.v.

SA

TFIDF

$\text{LRP} \Rightarrow \text{LRP}$

$\text{SA} \Rightarrow \text{SA}$

$\text{SUM} \Rightarrow \text{SUM}$

$\text{TFIDF} \Rightarrow \text{TFIDF}$

More details on LRP: www.heatmapping.org