Layer-Wise Relevance Propagation: An Overview

Paper link: https://tinyurl.com/su2z69y

Authors: Gr´egoire Montavon1, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert M¨uller

Abstract

- Developing methods for explainable AI is an area of active research
- If future, it’s will be inevitable to highlight the input features the machine learning model uses to support the prediction outcome for critical use cases
- Layer-wise Relevance Propagation (LRP) is one of the technique that brings such explainability

Layer-wise Relevance Propagation - Part I

- A technique that leverages the graph structure of the deep neural network
- The procedure is subject to the conservation property and behaviour is analogous to Kirchoff's conservation laws in electrical circuits
- Let j and k be neurons at two consecutive layers of the neural network. Propagating relevance scores is defined as:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k.$$

Layer-wise Relevance Propagation - Part II

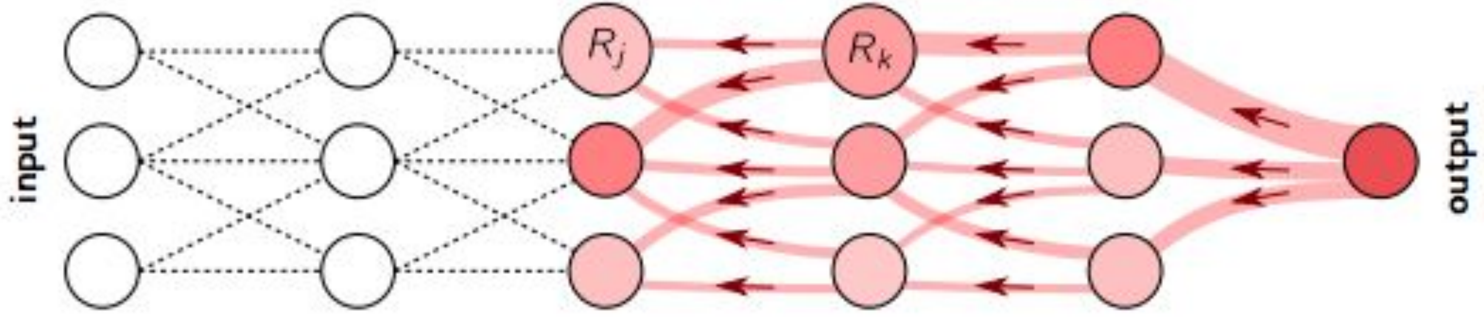


Fig. 10.2. Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer.

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

Basic Rule

$$R_j = \sum_k \frac{a_j \cdot \rho(w_{jk})}{\epsilon + \sum_{0,j} a_j \cdot \rho(w_{jk})} R_k,$$

Epsilon Rule

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k$$

Gamma Rule

Layer-wise Relevance Propagation - Part III

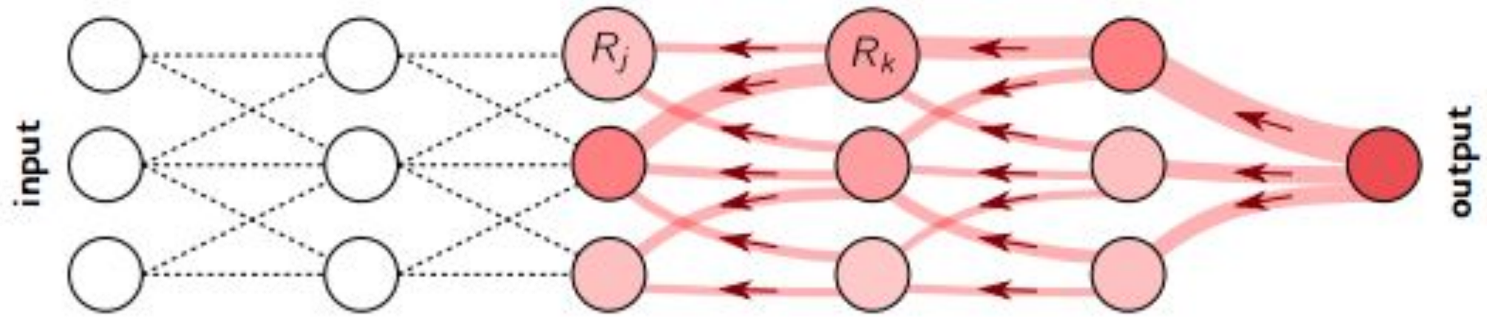


Fig. 10.2. Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer.

Propagation operations:

$$
\begin{aligned}
&\forall_k: \; z_k = \epsilon + \textstyle\sum_{0,j} a_j \cdot \rho(w_{jk}) && \text{(forward pass)}\\
&\forall_k: \; s_k = R_k / z_k && \text{(element-wise division)}\\
&\forall_j: \; c_j = \textstyle\sum_k \rho(w_{jk}) \cdot s_k && \text{(backward pass)}\\
&\forall_j: \; R_j = a_j \, c_j && \text{(element-wise product)}
\end{aligned}
$$

LRP as Deep Taylor Decomposition (Interpretation)

$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \sum_{i=1}^{d} (x_i - \widetilde{x}_i) \cdot [\nabla f(\widetilde{\boldsymbol{x}})]_i + \ldots$$

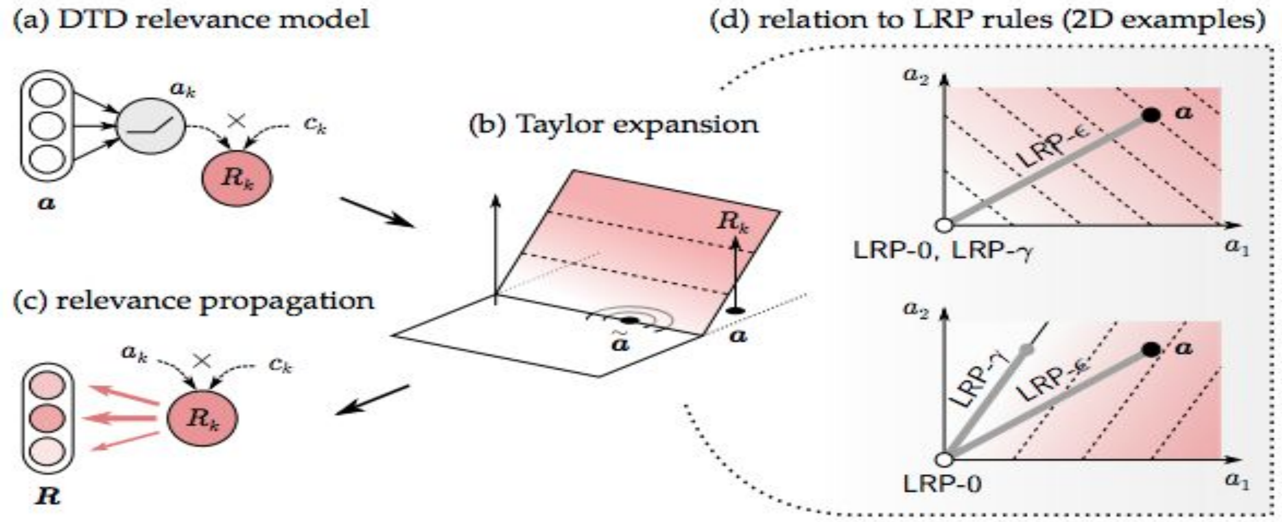


Fig. 10.3. Illustration of DTD: (a) graph view of the relevance model, (b) function view of the relevance model and reference point at which the Taylor expansion is performed, (c) propagation of first-order terms on the lower layer.

Some Use Cases

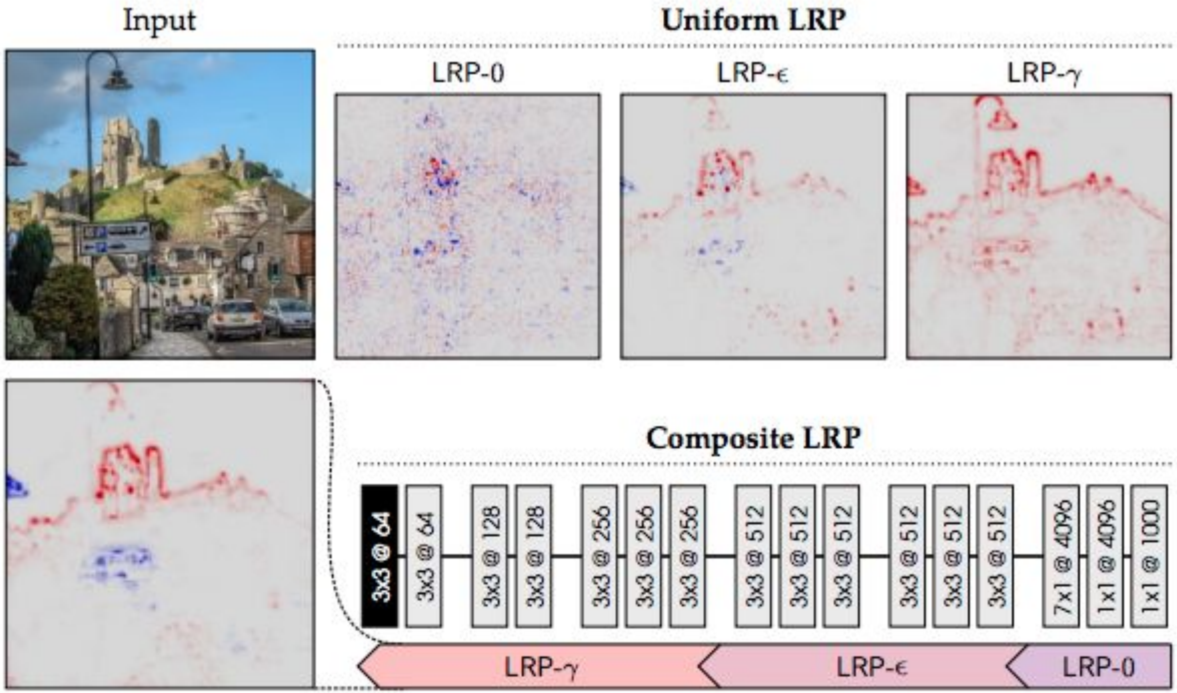
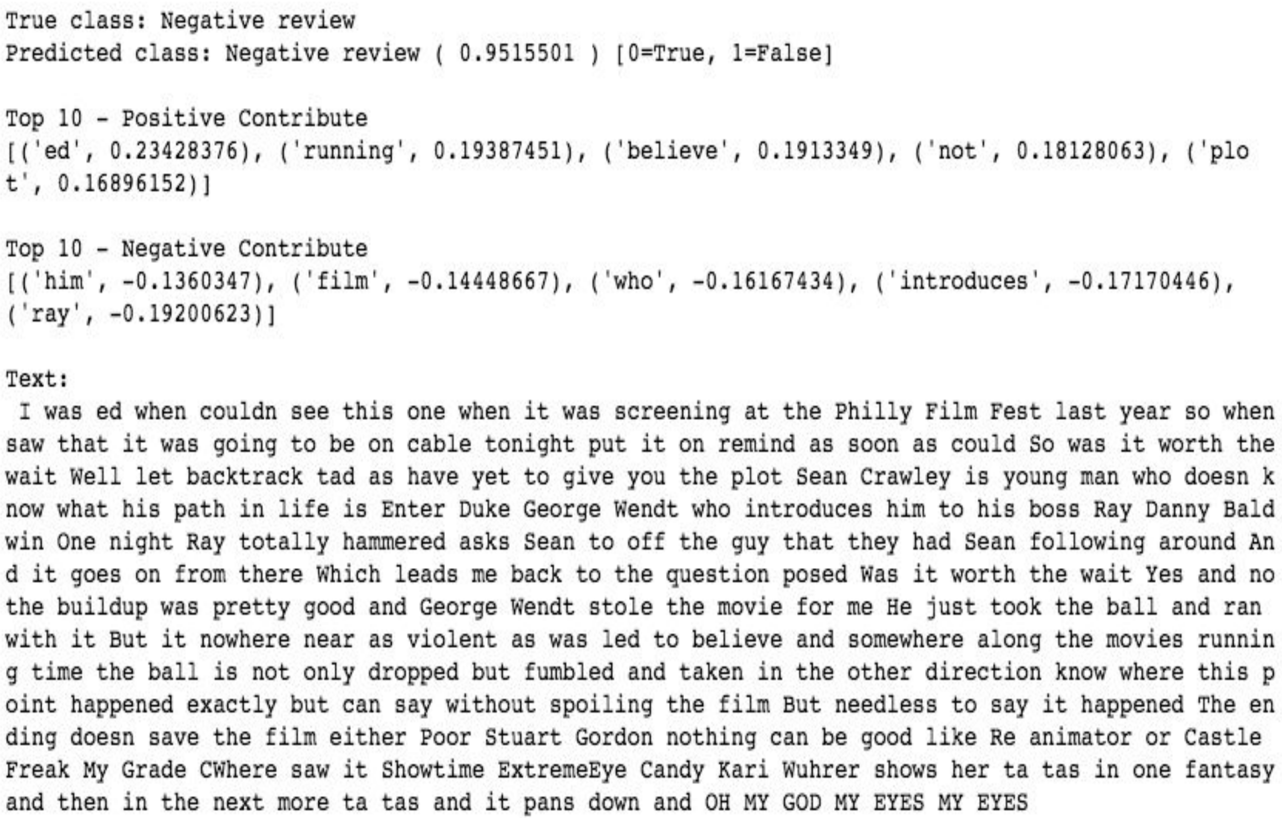


Fig. 10.4. Input image and pixel-wise explanations of the output neuron 'castle' obtained with various LRP procedures. Parameters are $\epsilon = 0.25\,\text{std}$ and $\gamma = 0.25$.

Implementation - Movie Reviews Sentiment Analysis

Link: https://tinyurl.com/ybybgz4x

- Glove Embedding
- Simple Dense Keras network
- Accuracy - 91%



```
True class: Negative review
Predicted class: Negative review ( 0.9515501 ) [0=True, 1=False]

Top 10 - Positive Contribute
[('ed', 0.23428376), ('running', 0.19387451), ('believe', 0.1913349), ('not', 0.18128063), ('plo
t', 0.16896152)]

Top 10 - Negative Contribute
[('him', -0.1360347), ('film', -0.14448667), ('who', -0.16167434), ('introduces', -0.17170446),
('ray', -0.19200623)]

Text:
 I was ed when couldn see this one when it was screening at the Philly Film Fest last year so when
saw that it was going to be on cable tonight put it on remind as soon as could So was it worth the
wait Well let backtrack tad as have yet to give you the plot Sean Crawley is young man who doesn k
now what his path in life is Enter Duke George Wendt who introduces him to his boss Ray Danny Bald
win One night Ray totally hammered asks Sean to off the guy that they had Sean following around An
d it goes on from there Which leads me back to the question posed Was it worth the wait Yes and no
the buildup was pretty good and George Wendt stole the movie for me He just took the ball and ran
with it But it nowhere near as violent as was led to believe and somewhere along the movies runnin
g time the ball is not only dropped but fumbled and taken in the other direction know where this p
oint happened exactly but can say without spoiling the film But needless to say it happened The en
ding doesn save the film either Poor Stuart Gordon nothing can be good like Re animator or Castle
Freak My Grade CWhere saw it Showtime ExtremeEye Candy Kari Wuhrer shows her ta tas in one fantasy
and then in the next more ta tas and it pans down and OH MY GOD MY EYES MY EYES
```