LRP

Seimandi Juliette - Thoirey Romane - Robin Gilles

TABLE OF CONTENTS

What is LRP?

- Layer-wise Relevance Propagation (LRP) : cIt’s an explaination tech applied to neural networks, where pictures video, text
- The goal : Generate explaination of the classification decisions made by the algo through analyzing deeply the model neuron by neuron
- The functionnment is based on backward propagation

Visual introduction

[Explainable AI Demos (fraunhofer.de)]

How works LRP?

Propragation :

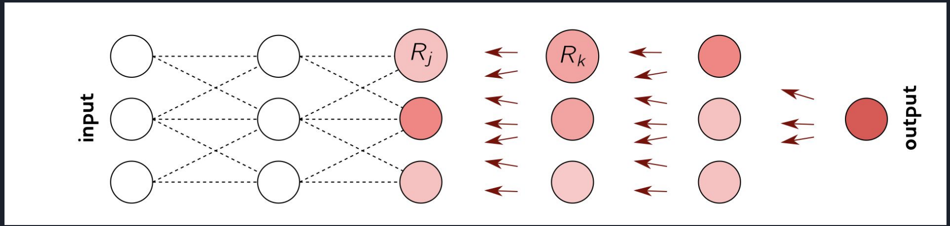
$$X_i^{L+1} = \phi(\boldsymbol{X}^L \boldsymbol{W}_i^L + \boldsymbol{b}_i^L)$$

Taylor decomposition :

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2!} f^{(2)}(x_0) + \cdots + \frac{h^n}{n!} f^{(n)}(x_0) + h^n \varepsilon(h)$$
$$= \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x_0) + h^n \varepsilon(h)$$

Propagating relevance scores

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



Rules of LRP

Basic Rule (LRP-0):

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

Epsilon Rule (LRP-ε) :

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

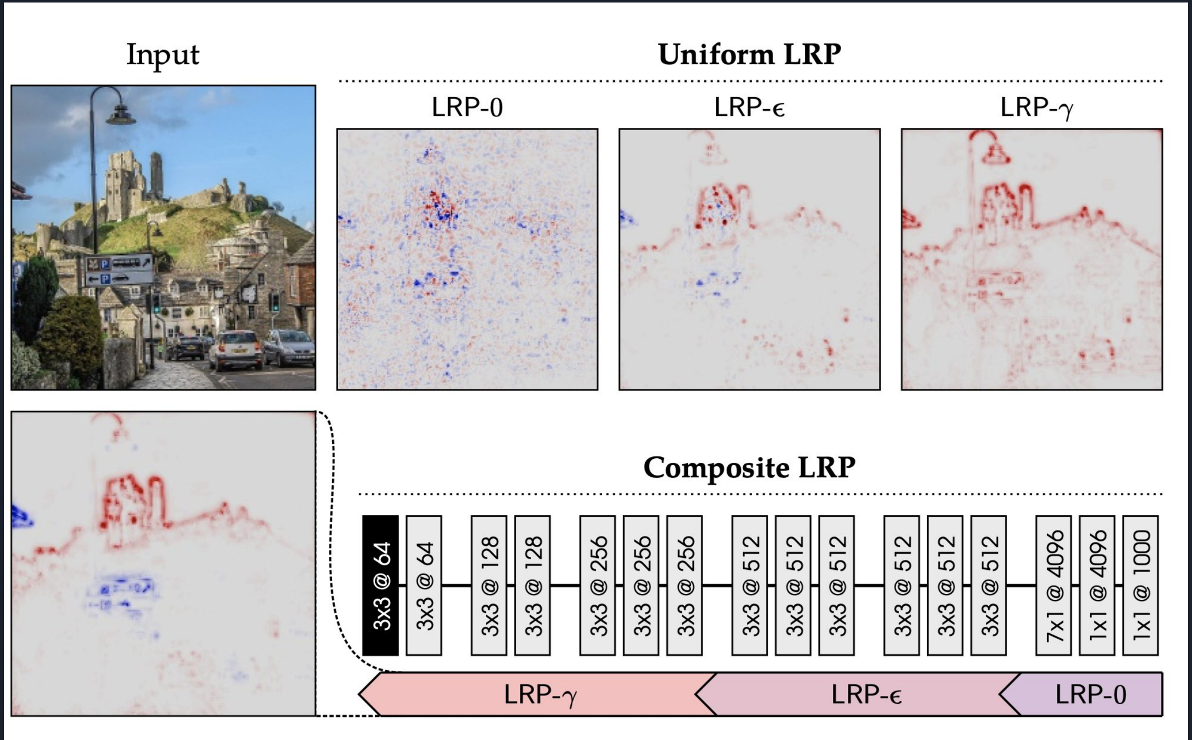
Gamma Rule (LRP-γ) :

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

Name	Formula	Usage	DTD
LRP-0 [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$	upper layers	✓
LRP-ϵ [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$	middle layers	✓
LRP-γ	$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$	lower layers	✓
LRP-$\alpha\beta$ [7]	$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-}\right) R_k$	lower layers	$\times^\star$
flat [30]	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	lower layers	$\times$
w^2-rule [36]	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$	first layer ($\mathbb{R}^d$)	✓
$z^{\mathcal{B}}$-rule [36]	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$	first layer (pixels)	✓

Implementation

Conclusion :


Input
Uniform LRP
LRP-0
LRP-ϵ
LRP-γ
Composite LRP
3x3 @ 64
3x3 @ 64
3x3 @ 128
3x3 @ 128
3x3 @ 256
3x3 @ 256
3x3 @ 256
3x3 @ 512
3x3 @ 512
3x3 @ 512
3x3 @ 512
3x3 @ 512
3x3 @ 512
7x1 @ 4096
1x1 @ 4096
1x1 @ 1000
LRP-γ
LRP-ϵ
LRP-0