

Extending Layerwise Relevance Propagation using Semiring Annotations

Antoine Groudiev
L3, ENS Ulm

Silviu Maniu – Supervisor
SLIDE Team, LIG

Tuesday, July 9th

Plan

Introduction

- Problem statement

- Layerwise Relevance Propagation

- Semiring-based provenance annotations

Extending LRP

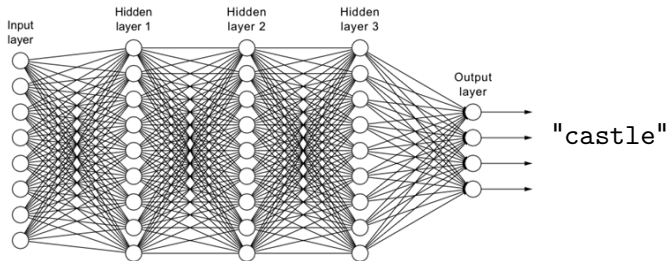
Applications

- Image mask computation

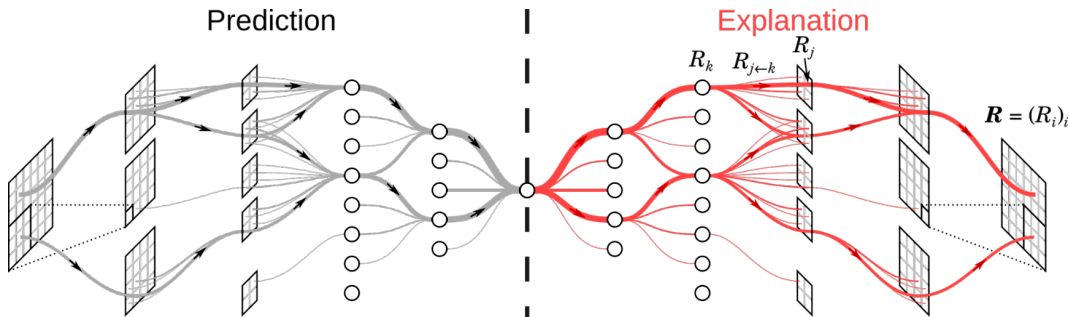
- Network pruning using LRP ranking

- Comparison to image perturbation

Problem statement



Layerwise Relevance Propagation [5]



Layerwise Relevance Propagation

Initialization

Initialization:

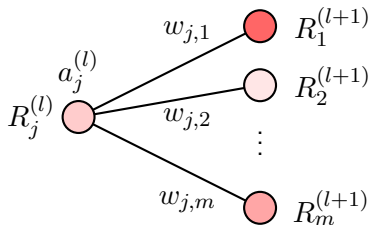
$$R_i^{(L)} = \begin{cases} a_i^{(L)} & \text{if } i = y \text{ (the class we want)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$\begin{bmatrix} 0 \end{bmatrix}$	\rightarrow "goldfish"
$\begin{bmatrix} 0 \end{bmatrix}$	\rightarrow "street sign"
\vdots	
$\begin{bmatrix} 1 \end{bmatrix}$	\rightarrow "castle"
\vdots	
$\begin{bmatrix} 0 \end{bmatrix}$	\rightarrow "printer"

Layerwise Relevance Propagation

LRP-0 rule:

$$R_j^{(l)} = \sum_k \frac{a_j^{(l)} w_{j,k}}{\sum_{j'} a_{j'}^{(l)} w_{j',k}} R_k^{(l+1)} \quad (2)$$



Other rules exist (LRP- ϵ , LRP- γ , z^B)

Multilayer Perceptron on MNIST dataset

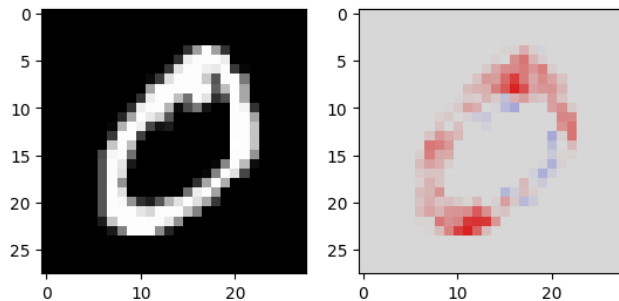


Figure: Reference image and relevance for the class 0

VVG-16 on ImageNet dataset



Figure: Reference image

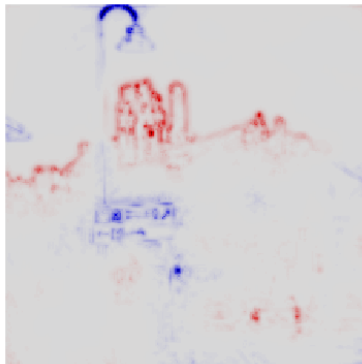


Figure: Relevance for the class "castle"

Semiring-based provenance annotations [4, 6]

Definition (Semiring)

A semiring $(\mathbb{K}, \oplus, \otimes, \mathbf{0}, \mathbf{1})$ is such that:

- \otimes distributes over \oplus ,
- $(\mathbb{K}, \oplus, \mathbf{0})$ is a commutative monoid,
- $(\mathbb{K}, \otimes, \mathbf{1})$ is a monoid such that $\mathbf{0}$ is absorbing

Example

The following structures are semirings:

- Real semiring: $(\mathbb{R}, +, \times, 0, 1)$
- Boolean semiring: $(\{\perp, \top\}, \vee, \wedge, \perp, \top)$
- Counting semiring: $(\mathbb{N}, +, \times, 0, 1)$
- Viterbi semiring: $([0, 1], \max, \times, 0, 1)$

Plan

Introduction

Problem statement

Layerwise Relevance Propagation

Semiring-based provenance annotations

Extending LRP

Applications

Image mask computation

Network pruning using LRP ranking

Comparison to image perturbation

Semiring generalization of the LRP rule

Consider a semiring $(\mathbb{K}, \oplus, \otimes, \mathbf{0}, \mathbf{1})$

Conversion functions for activations, weights:

$$\Theta_a : \mathbb{R} \longrightarrow \mathbb{K}$$

$$\Theta_w : \mathbb{R} \longrightarrow \mathbb{K}$$

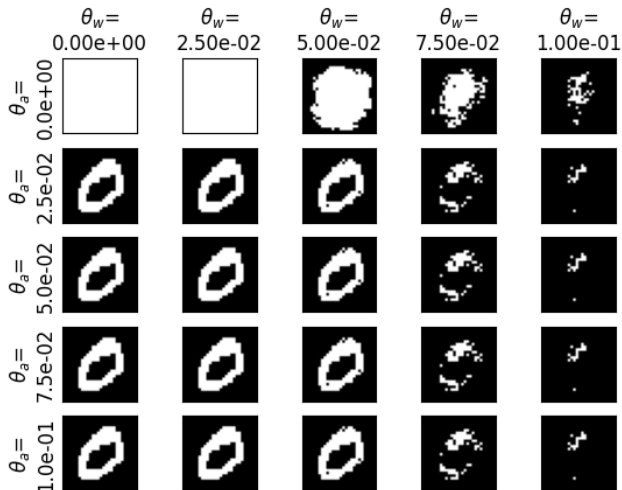
Initialization:

$$R_i^{(L)} = \begin{cases} \Theta_a \left(a_i^{(L)} \right) & \text{if } i = y \\ \mathbf{0} & \text{otherwise} \end{cases} \quad (3)$$

Propagation rule:

$$R_j^{(l)} = \bigoplus_k \Theta_a \left(a_j^{(l)} \right) \otimes \Theta_w \left(w_{j,k}^{(l)} \right) \otimes R_k^{(l+1)} \quad (4)$$

Influence of the thresholds

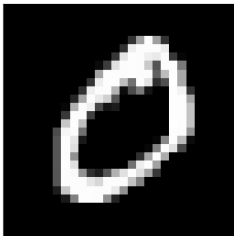


Viterbi Semiring

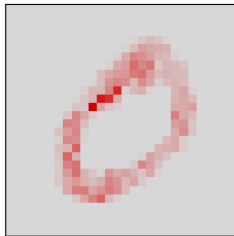
$$([0, 1], \max, \times, 0, 1)$$

$$R_j^{(l)} = \max_k \underbrace{\left(\frac{|a_j^{(l)} w_{j,k}^{(l)}|}{\max_{j'} |a_{j'}^{(l)} w_{j',k}^{(l)}|} \right)}_{\in [0,1]} \cdot R_k^{(l+1)}$$

Reference



Viterbi semiring



Plan

Introduction

Problem statement

Layerwise Relevance Propagation

Semiring-based provenance annotations

Extending LRP

Applications

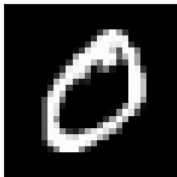
Image mask computation

Network pruning using LRP ranking

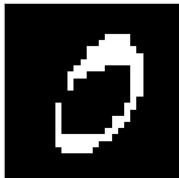
Comparison to image perturbation

Class-wise mask – Boolean semiring

Reference



Class-wise AND (5 examples)



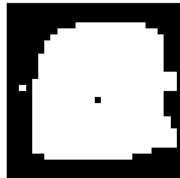
Class-wise OR
(5 examples)



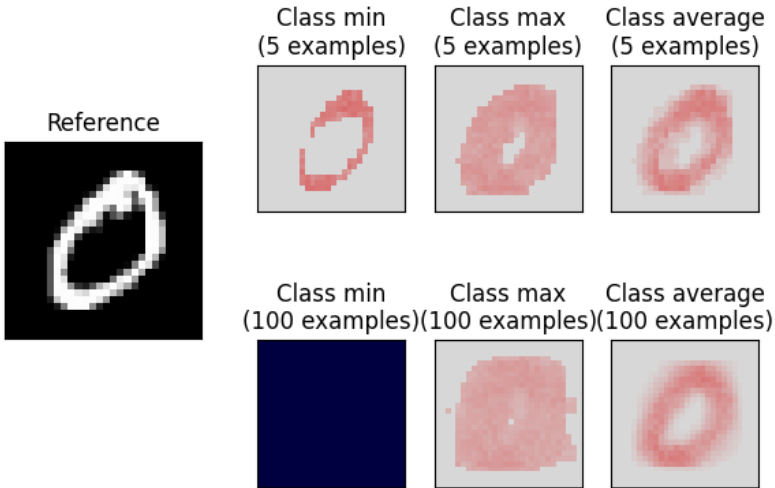
Class-wise AND
(100 examples)



Class-wise OR
(100 examples)



Class-wise mask – Counting semiring

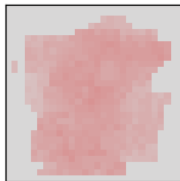


All classes mask – Counting semiring

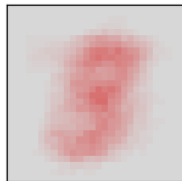
All classes min
(50 examples)



All classes max
(50 examples)



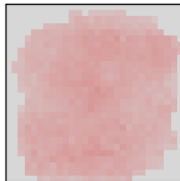
All classes average
(50 examples)



All classes min
(1000 examples)



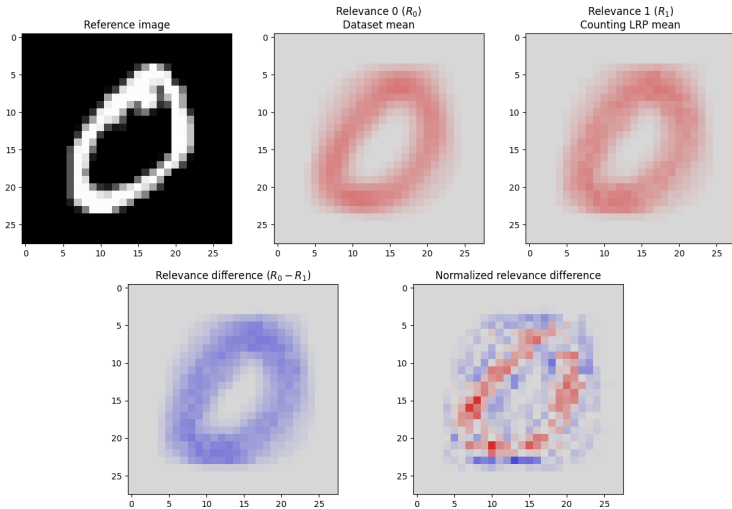
All classes max
(1000 examples)



All classes average
(1000 examples)



Comparison to dataset mean



Network pruning using LRP ranking

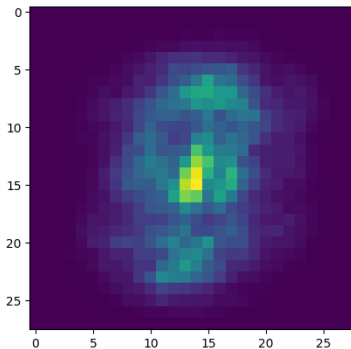
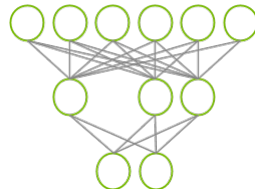
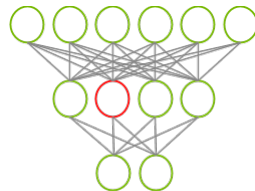
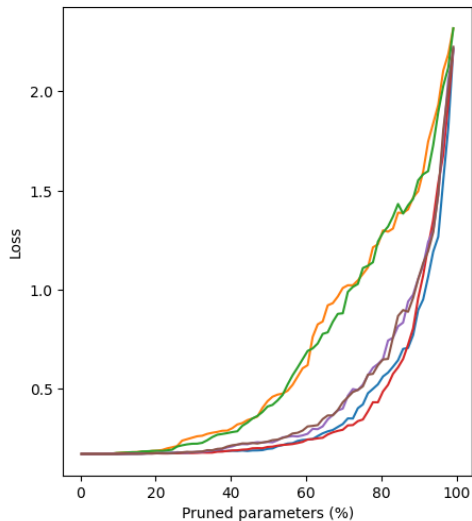


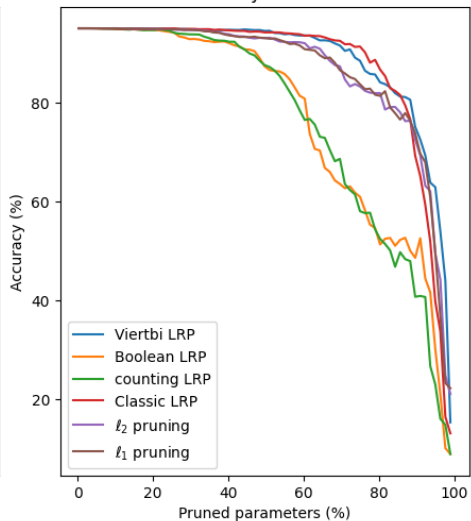
Figure: Relevance mean over the training dataset
(Input layer)



Loss evolution



Accuracy evolution



Comparison to image perturbation [2]

Accuracies per attack zone
Kernel size: 4 — Step: 1

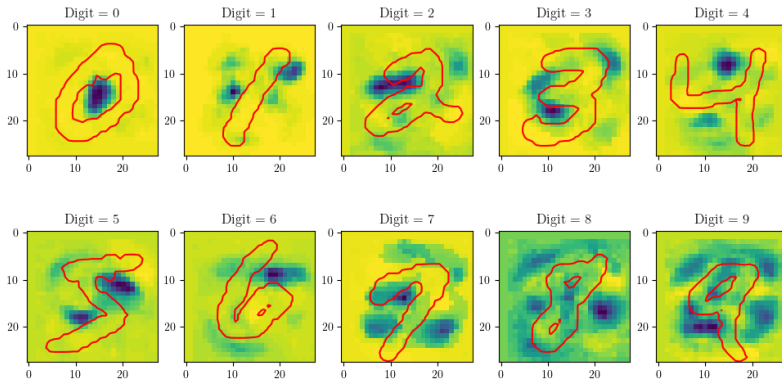


Figure: Accuracies per attack zone

