Extending Layerwise Relevance Propagation using Semiring Annotations

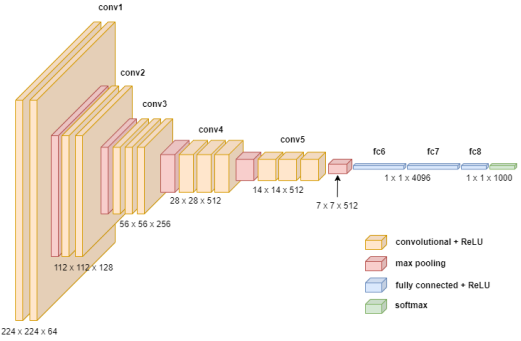
Antoine Groudiev
L3, ENS Ulm

Silviu Maniu – Supervisor
SLIDE Team, LIG

Tuesday, July 9th

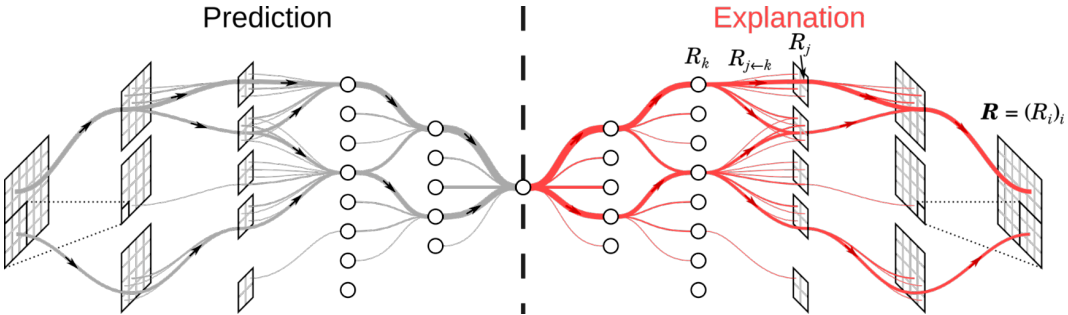
Plan

Problem statement



"castle"

Layerwise Relevance Propagation



Layerwise Relevance Propagation

Initialization

Initialization:

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

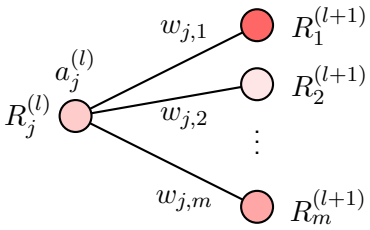
$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \mathbf{1} \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \longrightarrow \texttt{"goldfish"} \\ \longrightarrow \texttt{"street sign"} \\ \\ \longrightarrow \texttt{"castle"} \\ \\ \longrightarrow \texttt{"printer"} \end{array}$$

Layerwise Relevance Propagation

Propagation

LRP-0 rule:

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



Other rules exist (LRP-ϵ, LRP-γ, $z^{\mathcal{B}}$)

LRP Results visualization

Multilayer Perceptron on MNIST dataset

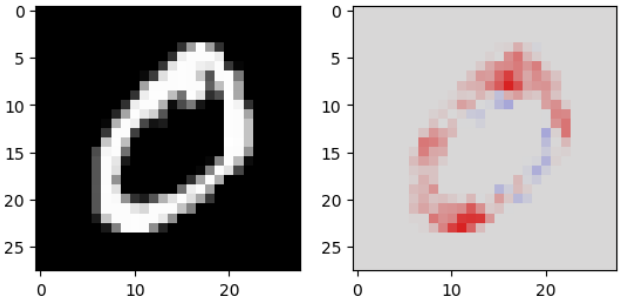


Figure: Reference image and relevance for the class 0

LRP Results visualization

VVG-16 on ImageNet dataset

Figure: Reference image

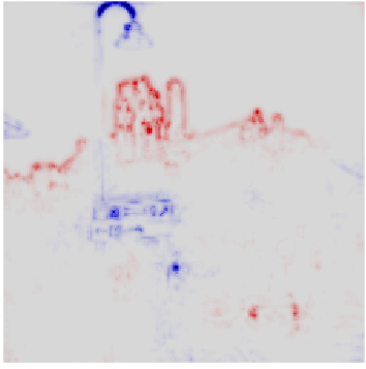
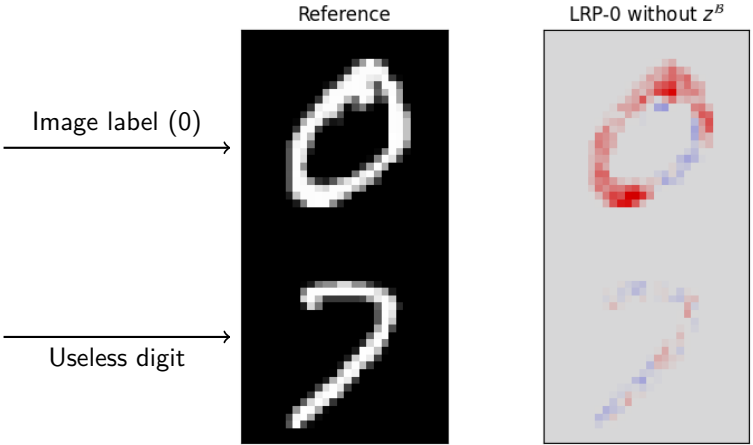


Figure: Relevance for the class `castle`

Pertinence of LRP results



Semiring-based provenance annotations

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

Plan

Simplifying LRP rule

Remove the denominator:

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

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:

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

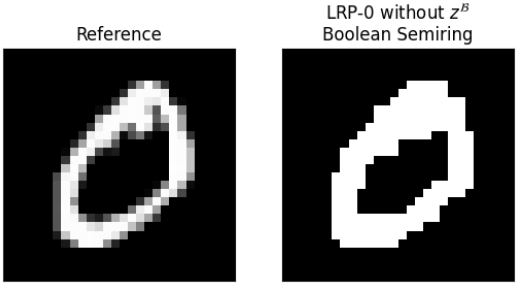
Propagation rule:

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

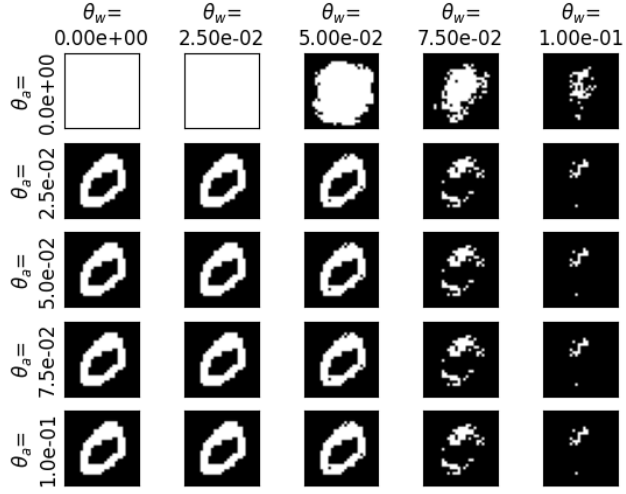
Boolean Semiring

$(\{\bot, \top\}, \vee, \wedge, \bot, \top)$

$$\Theta_a = a \longmapsto \begin{cases} \top & \text{if } a \geq \theta_a \\ \bot & \text{otherwise} \end{cases} \qquad \Theta_w = w \longmapsto \begin{cases} \top & \text{if } w \geq \theta_w \\ \bot & \text{otherwise} \end{cases}$$



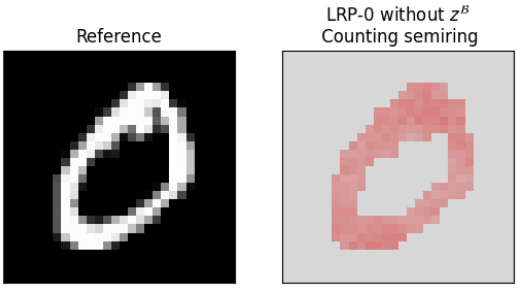
Influence of the thresholds



Counting Semiring

$(\mathbb{N}, +, \times, 0, 1)$

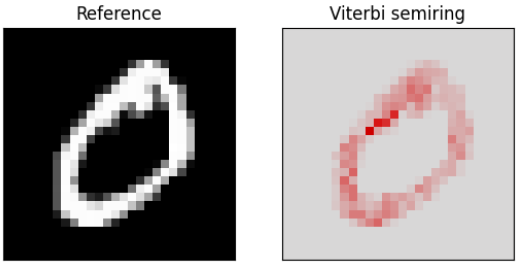
$$\Theta_a = a \longmapsto \begin{cases} 1 & \text{if } a \geq \theta_a \\ 0 & \text{otherwise} \end{cases} \qquad \Theta_w = w \longmapsto \begin{cases} 1 & \text{if } w \geq \theta_w \\ 0 & \text{otherwise} \end{cases}$$



Viterbi Semiring

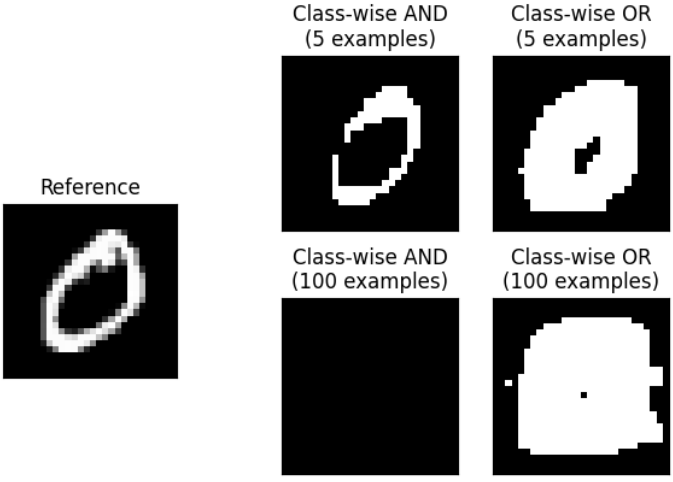
$([0,1], \max, \times, 0, 1)$

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



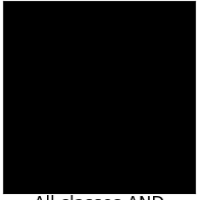
Plan

Class-wise mask – Boolean semiring

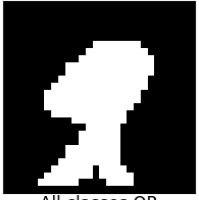


All classes mask – Boolean semiring

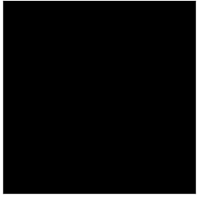
All classes AND
(50 examples)



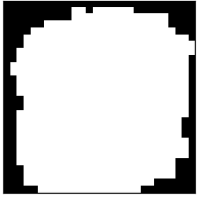
All classes OR
(50 examples)



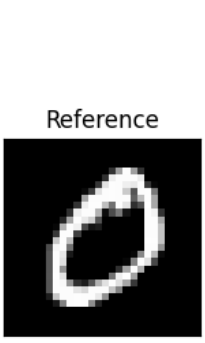
All classes AND
(500 examples)

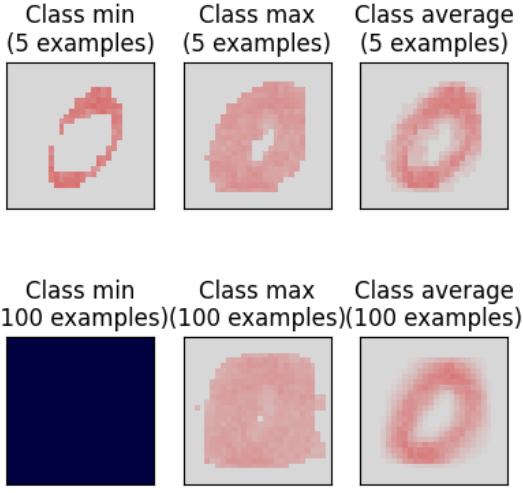


All classes OR
(500 examples)



Class-wise mask – Counting semiring

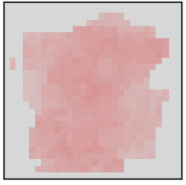




All classes mask – Counting semiring

All classes min (50 examples)

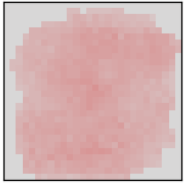
All classes max (50 examples)



All classes average (50 examples)

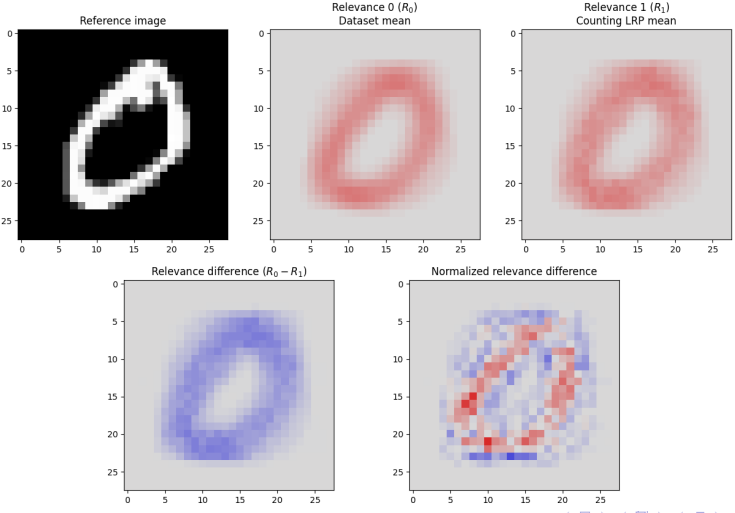
All classes min (1000 examples)

All classes max (1000 examples)



All classes average (1000 examples)

Comparison to dataset mean



Network pruning using LRP ranking

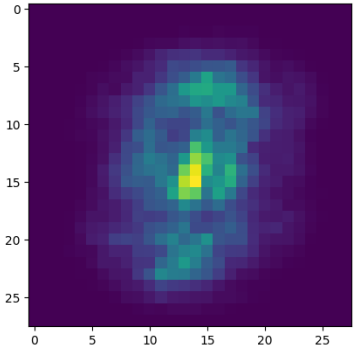
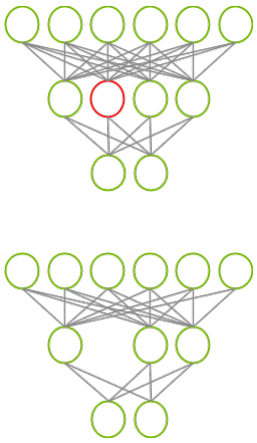
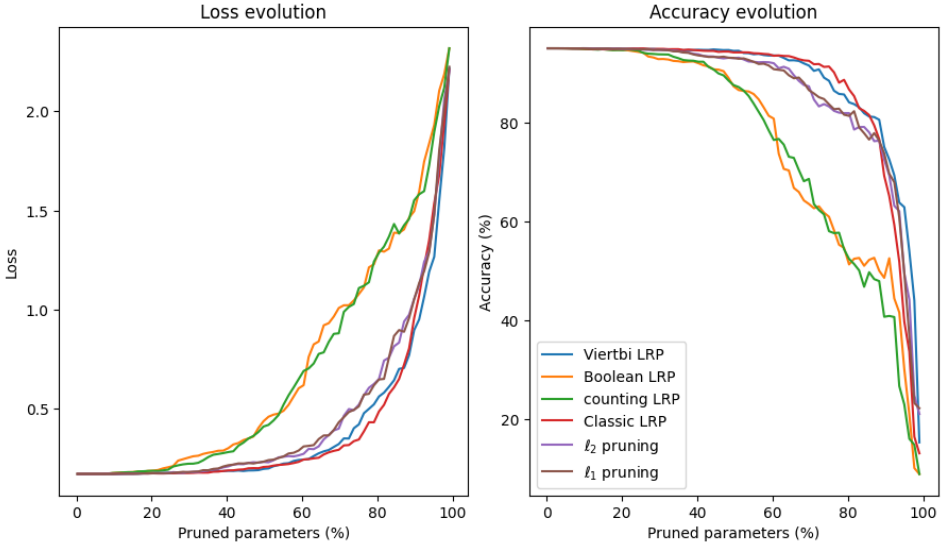


Figure: Relevance mean over the training dataset (Input layer)




Loss evolution
Loss
Pruned parameters (%)
Accuracy evolution
Accuracy (%)
Pruned parameters (%)
Viertbi LRP
Boolean LRP
counting LRP
Classic LRP
ℓ_2 pruning
ℓ_1 pruning

Comparison to image perturbation

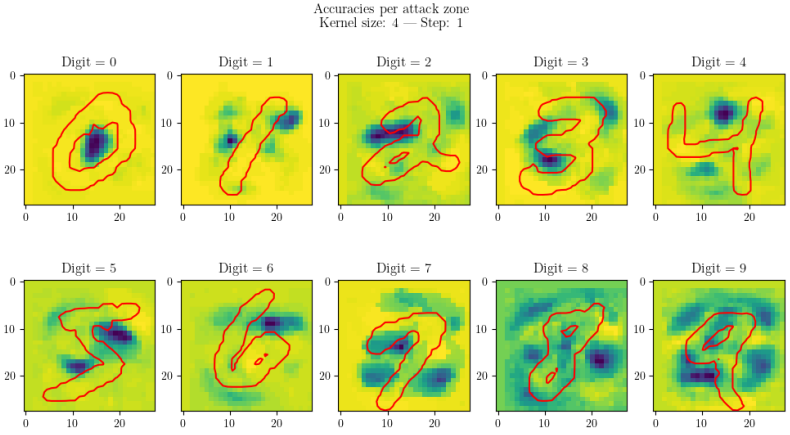


Figure: Accuracies per attack zone

[1] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: *PLOS ONE* (2015), pp. 1–46. DOI: 10.1371/journal.pone.0130140. URL: https://doi.org/10.1371/journal.pone.0130140.

[2] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3429–3437. URL: https://arxiv.org/abs/1704.03296.

[3] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2 (2020), pp. 665–673.

[4] Todd J Green, Grigoris Karvounarakis, and Val Tannen. "Provenance semirings". In: *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2007, pp. 31–40.

[5] Grégoire Montavon et al. "Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019, pp. 193–209. URL: https://doi.org/10.1007/978-3-030-28954-6_10.

[6] Yann Ramusat, Silviu Maniu, and Pierre Senellart. "Provenance-Based Algorithms for Rich Queries over Graph Databases". In: *EDBT 2021 - 24th International Conference on Extending Database Technology*. 2021. URL: https://inria.hal.science/hal-03140067.