

Extending Layerwise Relevance Propagation using Semiring Annotations

Antoine Groudiev
L3, ENS Ulm

Silviu Maniu – Supervisor
SLIDE Team, LIG

August 1, 2024



Plan

Introduction

- Problem statement

- Layerwise Relevance Propagation

- Semiring-based provenance annotations

Extending LRP

- Rules modification

- MNIST dataset

- VGG-16 network

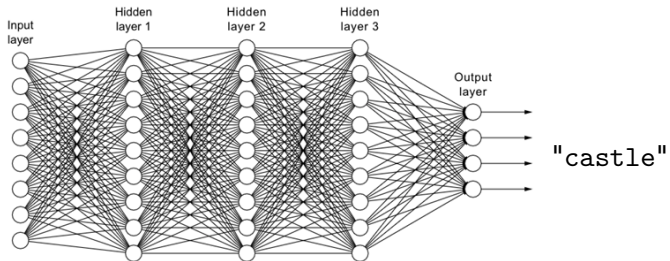
Applications

- Image mask computation

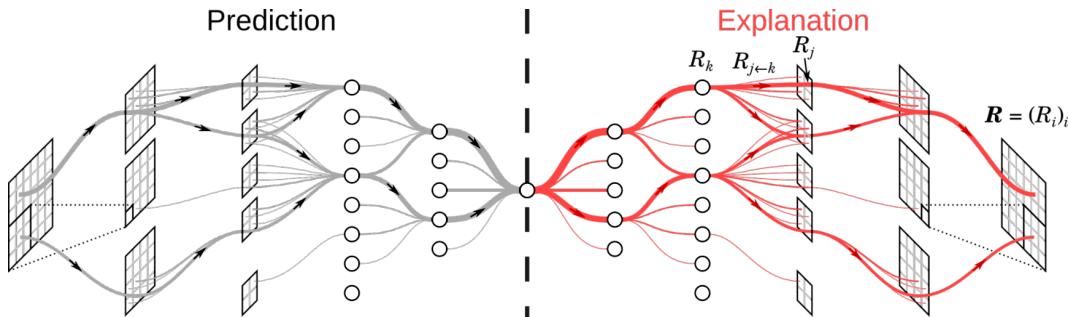
- Network pruning using LRP ranking

- Comparison to image perturbation

Problem statement



Layerwise Relevance Propagation [9]



Layerwise Relevance Propagation

Initialization

Initialization:

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

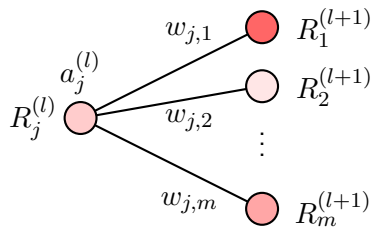
$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 4.2 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \rightarrow \text{"goldfish"} \\ \rightarrow \text{"street sign"} \\ \\ \rightarrow \text{"castle"} \\ \\ \rightarrow \text{"printer"} \end{matrix}$$

Layerwise Relevance Propagation

Propagation

LRP-0 rule:

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



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

LRP Results visualization

Multilayer Perceptron on MNIST dataset

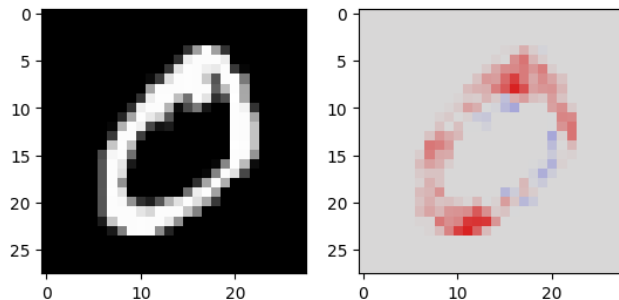


Figure: Reference image and relevance for the class 0

Semiring-based provenance annotations [5, 10]

Definition (Semiring)

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

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

Rules modification

MNIST dataset

VGG-16 network

Applications

Image mask computation

Network pruning using LRP ranking

Comparison to image perturbation

Semiring generalization of the LRP rule

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

Conversion function:

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

Initialization:

$$R_i^{(L)} = \begin{cases} \mathbb{1} & \text{if } i = y \\ \mathbb{0} & \text{otherwise} \end{cases} \quad (3)$$

Propagation rule:

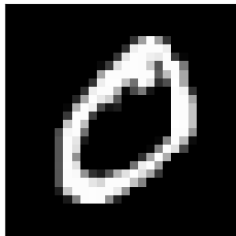
$$R_j^{(l)} = \bigoplus_k \Theta \left(\frac{a_j^{(l)} w_{j,k}}{\sum_{j'} a_{j'}^{(l)} w_{j',k}} \right) \otimes R_k^{(l+1)} \quad (4)$$

Counting Semiring

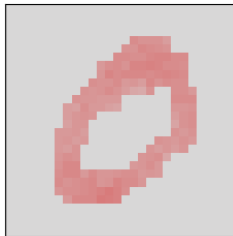
$(\mathbb{N}, +, \times, 0, 1)$

$$\Theta = x \mapsto \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Reference



Counting semiring

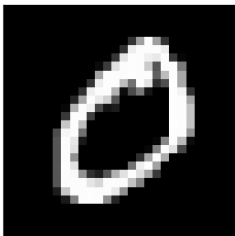


Viterbi Semiring

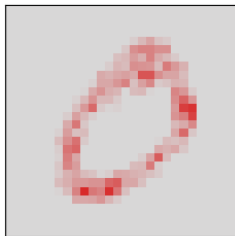
$([0, 1], \max, \times, 0, 1)$

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

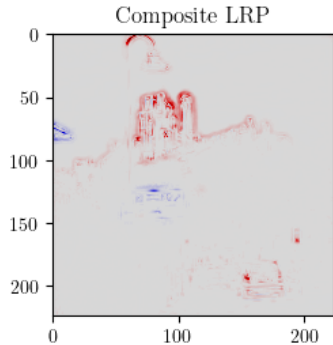
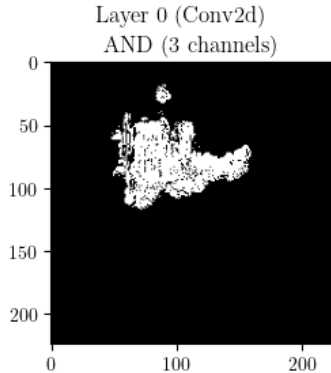
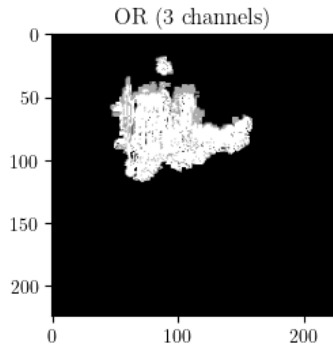
Reference



Viterbi semiring



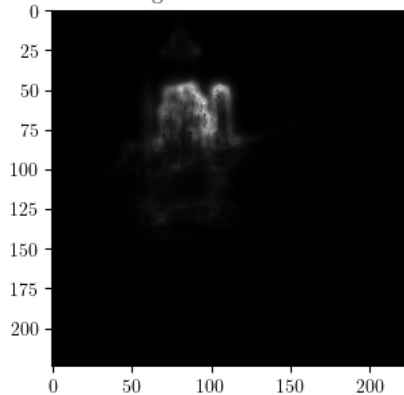
Boolean semiring



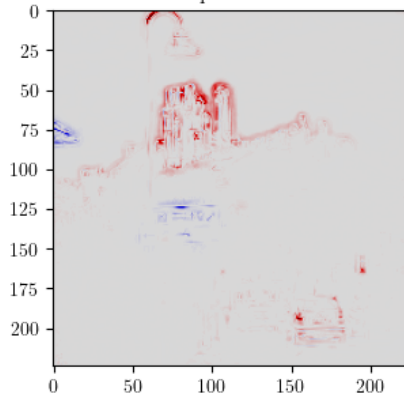
Counting semiring

Layer 0 (Conv2d)

Counting - Sum over 3 channels



Composite LRP



Plan

Introduction

- Problem statement

- Layerwise Relevance Propagation

- Semiring-based provenance annotations

Extending LRP

- Rules modification

- MNIST dataset

- VGG-16 network

Applications

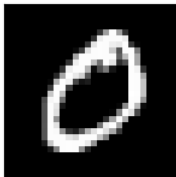
- Image mask computation

- Network pruning using LRP ranking

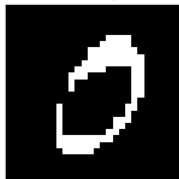
- Comparison to image perturbation

Class-wise mask – Boolean semiring

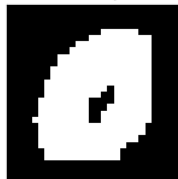
Reference



Class-wise AND (5 examples)



Class-wise OR
(5 examples)



Class-wise AND
(100 examples)



Class-wise OR
(100 examples)



All classes mask – Boolean semiring

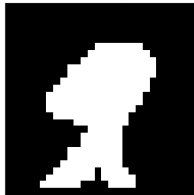
All classes AND
(50 examples)



All classes AND
(500 examples)



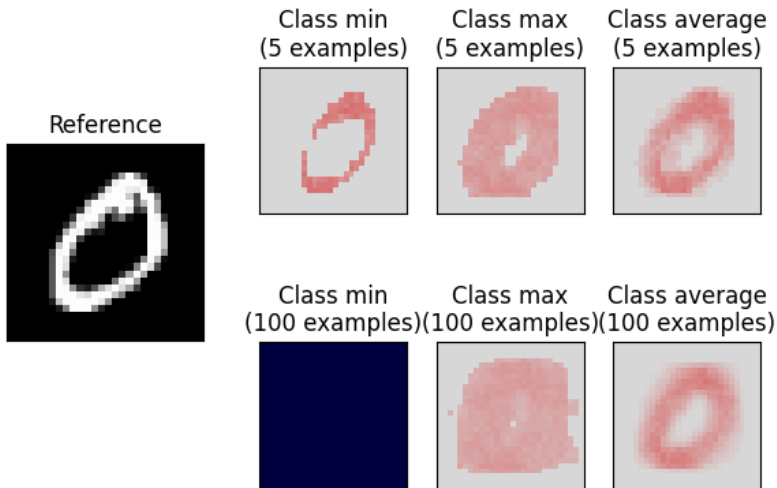
All classes OR
(50 examples)



All classes OR
(500 examples)



Class-wise mask – Counting semiring

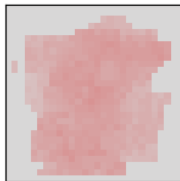


All classes mask – Counting semiring

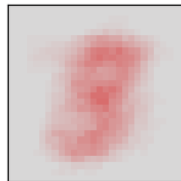
All classes min
(50 examples)



All classes max
(50 examples)



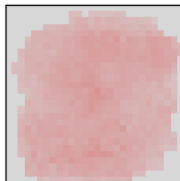
All classes average
(50 examples)



All classes min
(1000 examples)



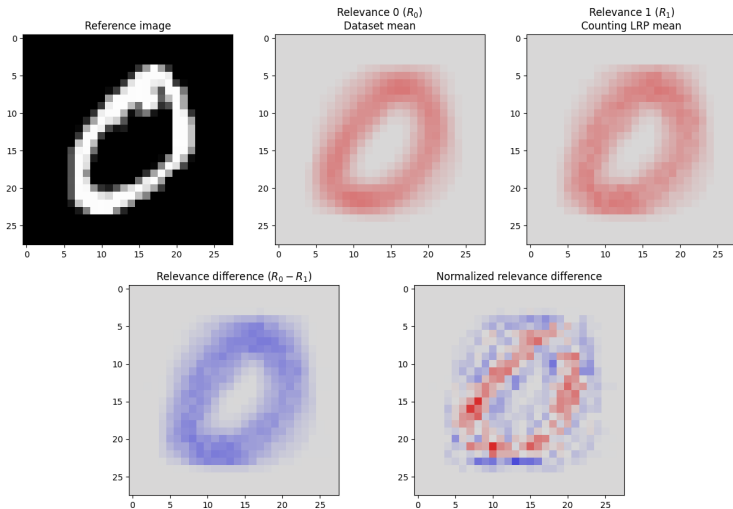
All classes max
(1000 examples)



All classes average
(1000 examples)



Comparison to dataset mean



Network pruning using LRP ranking

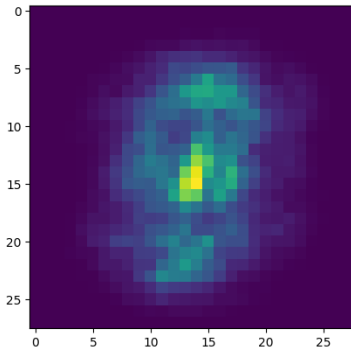
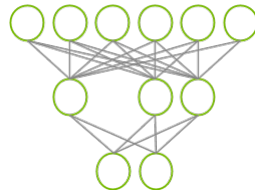
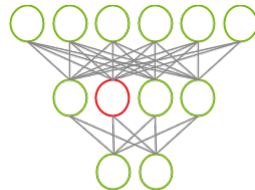
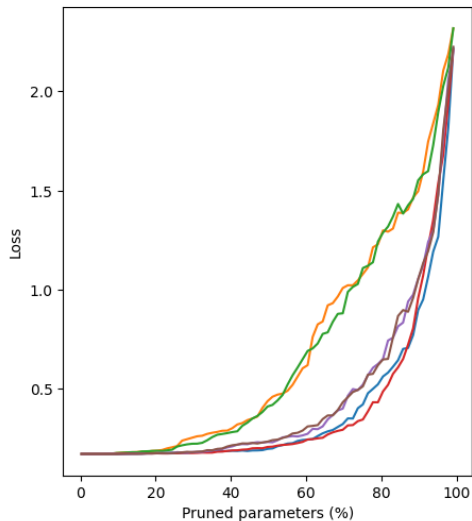


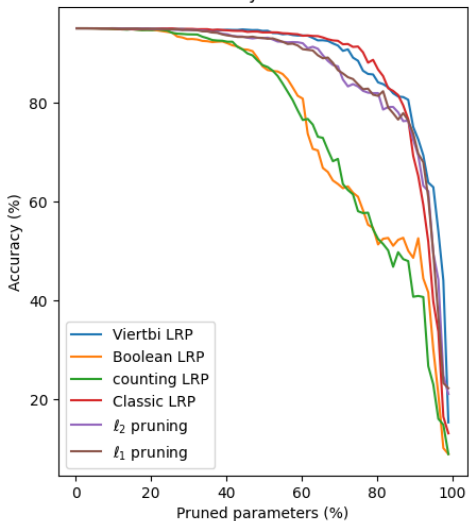
Figure: Relevance mean over the training dataset
(Input layer)



Loss evolution



Accuracy evolution



Comparison to image perturbation [3]

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

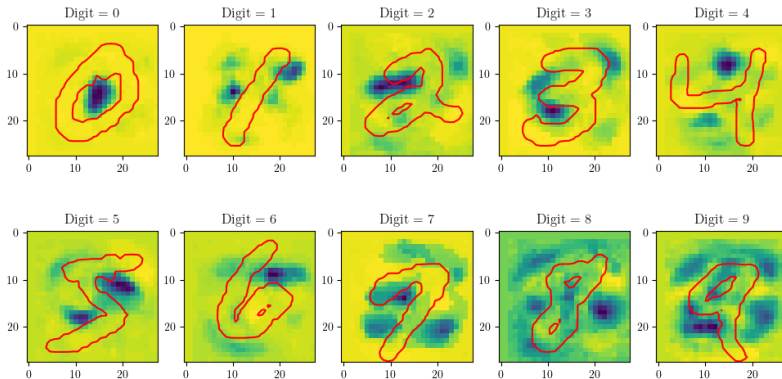


Figure: Accuracies per attack zone

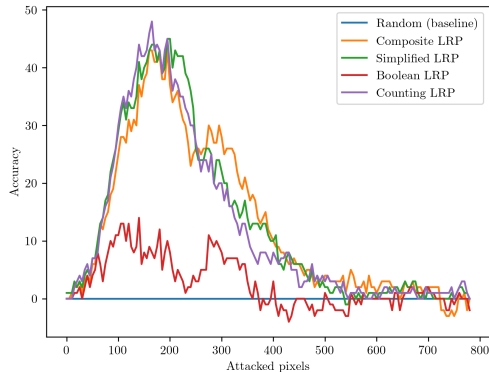
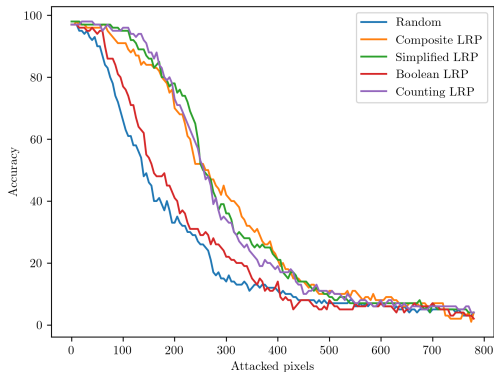


Figure: Accuracy drop for multiple pixels attacks strategies.



- [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. URL: <https://doi.org/10.1371/journal.pone.0130140>.
- [2] Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. “A Survey on Deep Neural Network Pruning-Taxonomy, Comparison, Analysis, and Recommendations”. In: (2023). URL: <https://arxiv.org/abs/2308.06767>.
- [3] 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>.
- [4] Robert Geirhos et al. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2 (2020), pp. 665–673. URL: <https://arxiv.org/abs/2004.07780>.
- [5] 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.
- [6] Yann LeCun. *The MNIST database of handwritten digits*. 1998. URL: <http://yann.lecun.com/exdb/mnist/>.



- [7] Aravindh Mahendran and Andrea Vedaldi. “Understanding deep image representations by inverting them”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5188–5196. URL: <https://arxiv.org/abs/1412.0035>.
- [8] Pavlo Molchanov et al. “Pruning Convolutional Neural Networks for Resource Efficient Inference”. In: (2017). URL: <https://arxiv.org/abs/1611.06440>.
- [9] 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.
- [10] 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>.
- [11] Wojciech Samek et al. “Evaluating the visualization of what a deep neural network has learned”. In: *IEEE transactions on neural networks and learning systems* 28.11 (2016), pp. 2660–2673. URL: <https://arxiv.org/abs/1509.06321>.
- [12] Ramprasaath R Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626. URL: <https://arxiv.org/pdf/1610.02391>.



- [13] Pierre Senellart. “Provenance and probabilities in relational databases”. In: *ACM SIGMOD Record* 46.4 (2018), pp. 5–15. URL: <https://inria.hal.science/hal-01672566>.
- [14] Pierre Senellart et al. “ProvSQL: Provenance and probability management in postgresql”. In: *Proceedings of the VLDB Endowment (PVLDB)* (2018). URL: <https://inria.hal.science/hal-01851538>.
- [15] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. URL: <https://arxiv.org/abs/1409.1556>.
- [16] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I* 13. Springer. 2014, pp. 818–833. URL: <https://arxiv.org/abs/1311.2901>.