

Plan

Introduction

Problem statement

Layerwise Relevance Propagation

Semiring-based provenance annotations

Extending LRP

Semiring generalization of the LRP rule

Results over the MNIST dataset

Applications

Image mask computation

Network pruning using LRP ranking

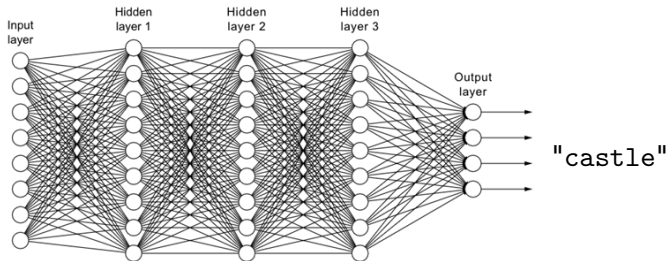
Comparison to image perturbation

Handling CNNs

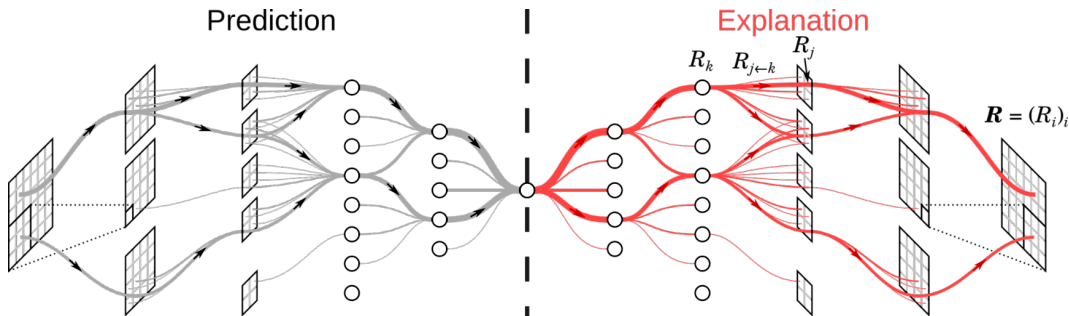
Computing relevance for convolutional layers

Results for the VGG-16 network

Problem statement



Layerwise Relevance Propagation [11]



Layerwise Relevance Propagation

Initialization

Initialization:

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

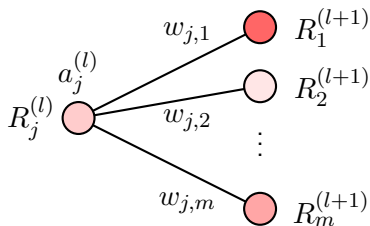
$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 4.2 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \rightarrow \text{"goldfish"} \\ \rightarrow \text{"street sign"} \\ \\ \rightarrow \text{"castle"} \\ \\ \rightarrow \text{"printer"} \end{matrix}$$

Layerwise Relevance Propagation

Propagation

LRP-0 rule:

$$R_j^{(l)} = \sum_k \frac{a_j^{(l)} w_{j,k}}{\sum_{j'} a_{j'}^{(l)} w_{j',k}} \cdot R_k^{(l+1)} \quad (\text{LRP-0})$$



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

Semiring-based provenance annotations [7, 12]

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

Semiring generalization of the LRP rule

Results over the MNIST dataset

Applications

Image mask computation

Network pruning using LRP ranking

Comparison to image perturbation

Handling CNNs

Computing relevance for convolutional layers

Results for the VGG-16 network

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:

$$R_i^{(L)} = \begin{cases} \mathbb{1} & \text{if } i = y \\ \mathbb{0} & \text{otherwise} \end{cases} \quad (2)$$

Propagation rule:

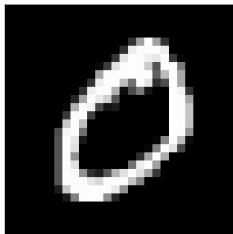
$$R_j^{(l)} = \bigoplus_k \Theta \left(\frac{a_j^{(l)} w_{j,k}}{\sum_{j'} a_{j'}^{(l)} w_{j',k}} \right) \otimes R_k^{(l+1)} \quad (\mathbb{K}\text{-LRP})$$

Boolean Semiring

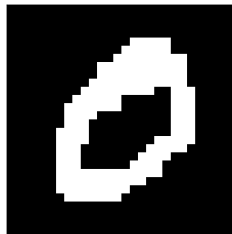
$(\{\perp, \top\}, \vee, \wedge, \perp, \top)$

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

Reference



Boolean Semiring

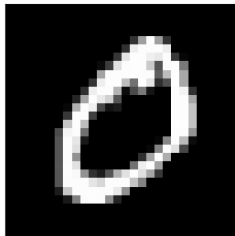


Counting Semiring

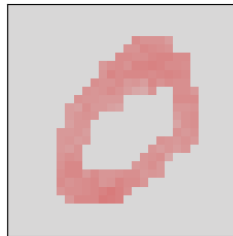
$(\mathbb{N}, +, \times, 0, 1)$

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

Reference




Counting semiring



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

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



Plan

Introduction

Problem statement

Layerwise Relevance Propagation

Semiring-based provenance annotations

Extending LRP

Semiring generalization of the LRP rule

Results over the MNIST dataset

Applications

Image mask computation

Network pruning using LRP ranking

Comparison to image perturbation

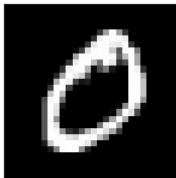
Handling CNNs

Computing relevance for convolutional layers

Results for the VGG-16 network

Class-wise mask – Boolean semiring

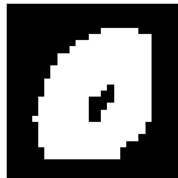
Reference



Class-wise AND (5 examples)



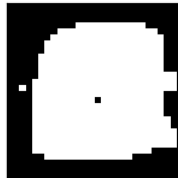
Class-wise OR
(5 examples)



Class-wise AND
(100 examples)




Class-wise OR
(100 examples)



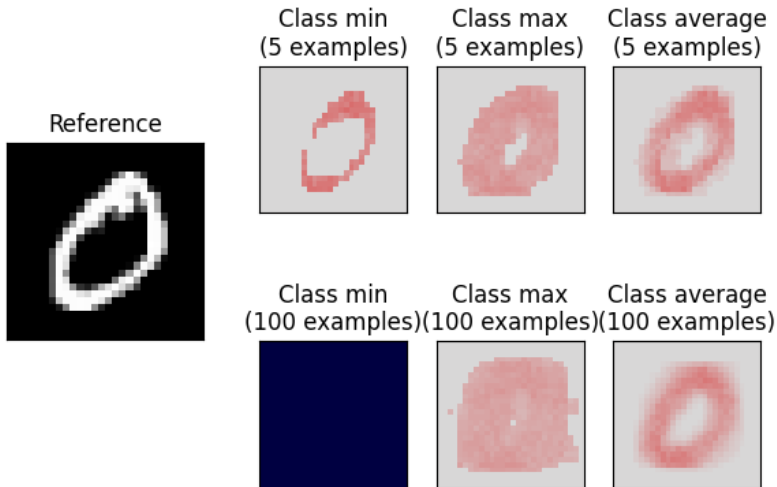
All classes mask – Boolean semiring

1. *Journal of Management Studies*, 1997, 34, 1, 1-14.

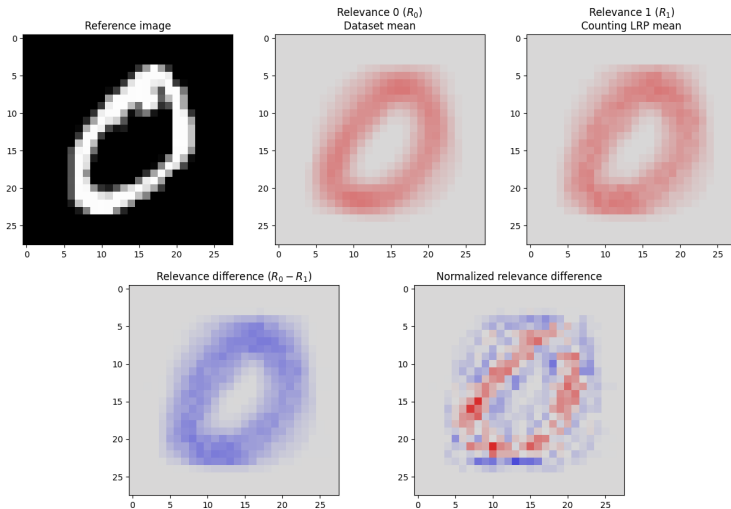
Abstract



Class-wise mask – Counting semiring



Comparison to dataset mean



Network pruning using LRP ranking

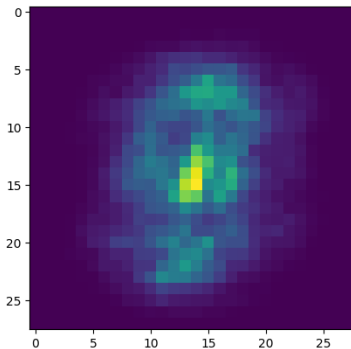
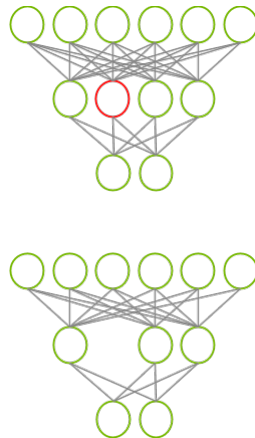


Figure 4: Relevance mean over the training dataset
(Input layer)



The graph plots Loss (Y-axis, 0.0 to 2.0+) against Pruned parameters (%) (X-axis, 0 to 100). Five methods are compared: a baseline (black), a proposed method (blue), and three other methods (orange, green, red). All methods start at a loss of approximately 0.15 at 0% pruning. The proposed method (blue) shows the slowest increase in loss as pruning progresses, maintaining a loss below 0.5 until about 70% pruning, then rising to approximately 2.2 at 100% pruning. The other methods (orange, green, red) show a more rapid increase in loss, reaching approximately 2.2 at 100% pruning. The green method shows the highest loss for most of the pruning range between 60% and 90%.

| Pruned parameters (%) | Baseline (Black) | Proposed (Blue) | Other 1 (Orange) | Other 2 (Green) | Other 3 (Red) |
|-----------------------|------------------|-----------------|------------------|-----------------|---------------|
| 0 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |
| 20 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |
| 40 | 0.20 | 0.18 | 0.25 | 0.25 | 0.20 |
| 60 | 0.30 | 0.25 | 0.60 | 0.70 | 0.35 |
| 80 | 0.60 | 0.50 | 1.30 | 1.40 | 0.60 |
| 100 | 2.20 | 2.20 | 2.20 | 2.20 | 2.20 |

[illegible]

Comparison to image perturbation [5]

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

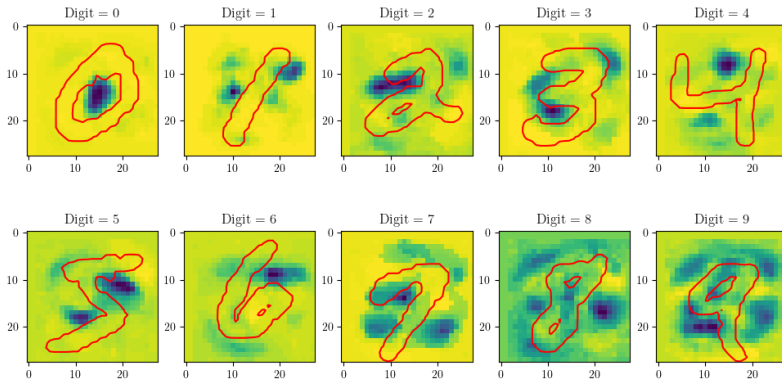


Figure 5: Accuracies per attack zone

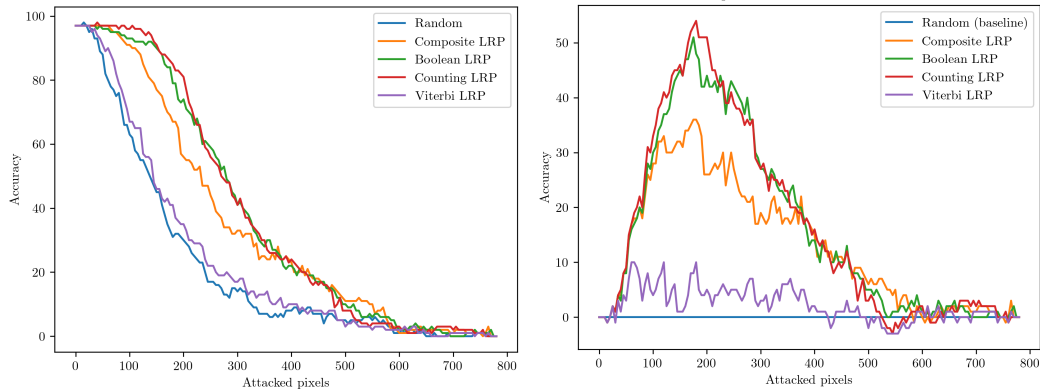


Figure 6: Accuracy drop for multiple pixels attacks strategies.

Plan

Introduction

Problem statement

Layerwise Relevance Propagation

Semiring-based provenance annotations

Extending LRP

Semiring generalization of the LRP rule

Results over the MNIST dataset

Applications

Image mask computation

Network pruning using LRP ranking

Comparison to image perturbation

Handling CNNs

Computing relevance for convolutional layers

Results for the VGG-16 network

Computing relevance for convolutional layers

$$R_j^{(l)} = \underbrace{\bigoplus_k \underbrace{\Theta \left(\frac{a_j^{(l)} w_{j,k}}{\sum_{j'} a_{j'}^{(l)} w_{j',k}} \right)}_{\text{Convolution over layer } l}}_{\text{Convolution over layer } l+1} \otimes R_k^{(l+1)} \quad (\mathbb{K}\text{-LRP})$$

VGG-16 network

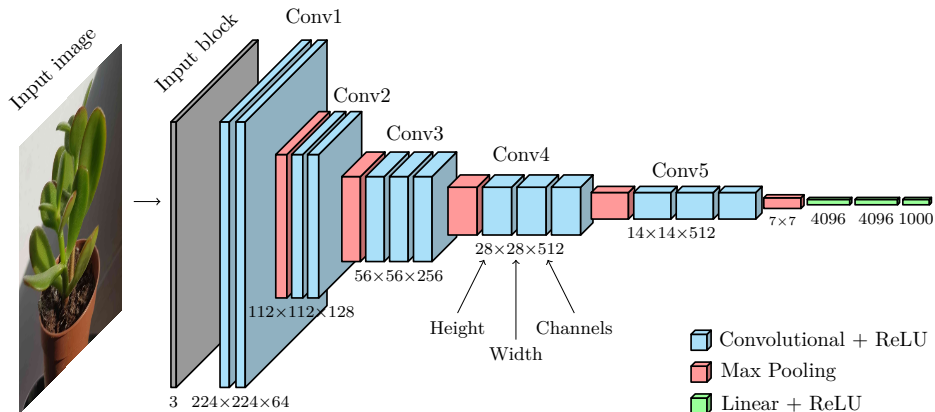
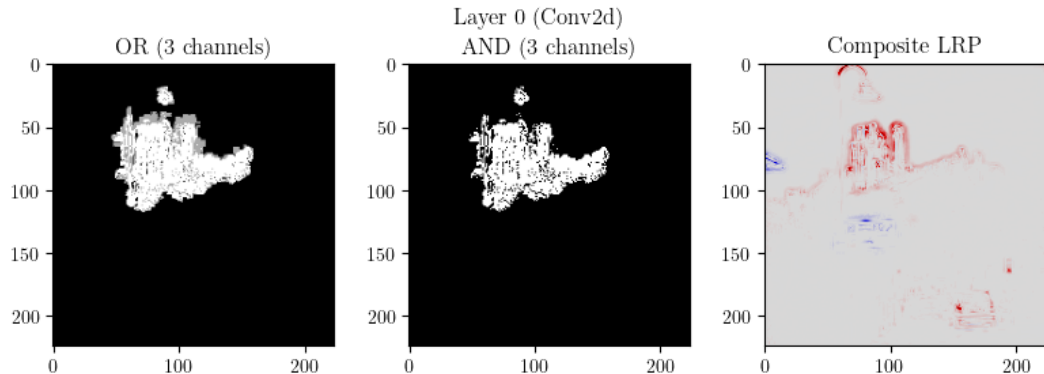


Figure 7: Architecture of the VGG-16 network.

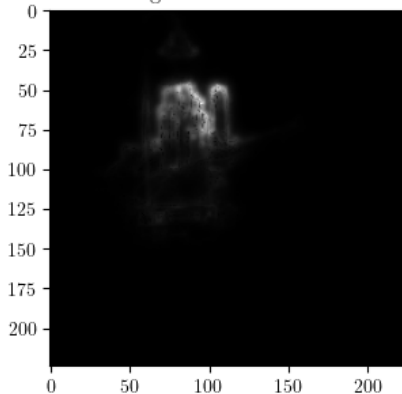
Results for VGG-16: Boolean semiring



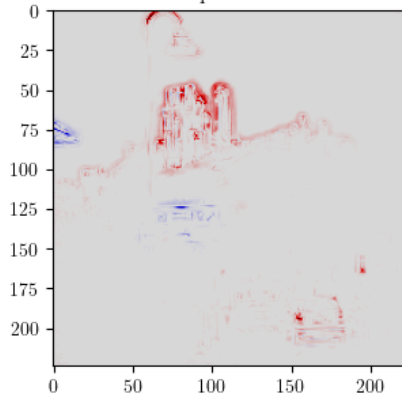
Results for VGG-16: Counting semiring

Layer 0 (Conv2d)

Counting - Sum over 3 channels



Composite LRP



- [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] Nick Cammarata et al. “Thread: Circuits”. In: *Distill* (2020). <https://distill.pub/2020/circuits>. DOI: 10.23915/distill.00024.
- [3] 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>.
- [4] Marina Danilevsky et al. “A survey of the state of explainable AI for natural language processing”. In: *arXiv preprint* (2020). URL: <https://arxiv.org/abs/2010.00711>.
- [5] 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>.
- [6] 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>.

- [7] 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.
- [8] Yann LeCun. *The MNIST database of handwritten digits*. 1998. URL: <http://yann.lecun.com/exdb/mnist/>.
- [9] 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>.
- [10] Pavlo Molchanov et al. “Pruning Convolutional Neural Networks for Resource Efficient Inference”. In: (2017). URL: <https://arxiv.org/abs/1611.06440>.
- [11] 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.
- [12] 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>.

- [13] 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>.
- [14] 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>.
- [15] 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>.
- [16] 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>.
- [17] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. URL: <https://arxiv.org/abs/1409.1556>.
- [18] 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>.