Reasoning and Provenance on Neural Networks

Antoine Groudiev
L3, DI, ENS

Silviu Maniu
Slide Team, LIG, UGA

13th June 2024

Abstract

Recently, neural networks allowed computers to solve numerous problems from diverse machine learning fields, such as natural language processsing and computer vision. Compared to traditional algorithms, machine learning models have proven both more successful and more difficult to interpret. Neural networks are considered as black boxes unable to easily explain itself, that is justifying the reasons that led him to make a prediction. Layer-wise Relevance Propagation (LRP) is a technique that has been introduced to provide explanability by identifying the input features relevant to the output choice. In parallel, research in the databases field developed annotations techniques to compute provenance for queries. In this paper, we extend LRP propagation rules to semiring-based provenance annotations of the network, for LRP to gain in expressivity.

1 Introduction

1.1 Problem statement

Deep neural networks have proven successful for solving with high accuracy machine learning problems. The expressivity of the class of functions generated by neural networks, combined with the relative simplicity of their training, make such models versatile tools to learn the relationship between the inputs and outputs of a dataset.

However, this versatility comes at the cost of poor interpretability: a neural network simply represents a function from one high-dimensional space to another, but provides no justification nor explanation for a given execution. If metrics such as the accuracy over a testing set provide confidence in the fact that the model is able to correctly classify inputs similar to the training set, no guarantee is given that the model generalizes well. Real-world examples show that networks can overfit the input data, or even take shortcuts instead of learning the intended solution [2]. For the user to have confidence in its predictions, a neural network should therefore be able to highlight the patterns in the input data that it actually learned.

1.2 Layer-wise Relevance Propagation

1.3 Semiring-based provenance annotations

References

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

[3] 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.

[4] 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.