Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth
M.Sc. Computer Science
Stony Brook University
crcollinswor@cs.stonybrook.edu

Abstract

Layer-wise Relevance Propagation (LRP) is a technique to analyze a classifiers decision making. It is often categorized as a method of explainable Artificial Intelligence (AI). LRP evaluates a classification decision, and is able to quantify each inputs' and intermediate representations' "importance" to said decision. This paper makes two main strides in this area. First, we test the applicability of LRP results by removing image areas quantified as important, and image areas quantified as unimportant and compare via subsequent classification outputs. We found that LRP is effective at identifying important features. Second, we use the results of LRP over a sampled dataset and remove inputs deemed unimportant. The ability to classify remained mostly unchanged; we successfully applied LRP to minimize the computational requirements of the network and optimize effeiciency.

1 Introduction

Neural Networks (NN) have provided machine learning a classifier with unique benefits and detriments. They are simultaneously scalable and tunable function optimizers capable of seemingly endless applicability, and black boxes which provide no introspection about their decisions. However, recent methods such as Layerwise Relevance Propagation (LRP) have been making strides in bringing explainability to NNs. LRP is employed after training a model, at which point it finds out the input features the network learned. This is done by taking the network prediction, and propagating it back through the network - accounting for previous activations and weights - to the inputs. Relevance values are obtained for the input nodes, which show the inputs that are most relevant to the predictions.

2 Background

Introduced to the field of Explainable AI (XAI) in 2015 [1], LRP was introduced to provide insight into how machine learning models actually use inputs to classify. It closes the gap between classification and interpretability both for multilayered NNs and Bag of Words over nonlinear kernels. The original paper derived solutions for these two particular classification architectures and evaluated them for meaningfulness using pixel-wise decomposition - a concept used to judge the contribution of each pixel to the output in the image classification task.

2.1 Origin

Conceptually, LRP assumes that each node can be assigned a "relevance" value, representing how much this node contributes to a certain classification. Directly from the paper, "Rd < 0 contributes evidence against the presence of a structure which is to be classified while Rd > 0 contributes evidence for its presence", where Rd is the relevance value of node d. Assuming conservation of the sum of relevances in each layer, the authors are able to compute relevances for each layer from the output to the input nodes. This relevance assignment technique worked so well, it was attempted to be patented in early 2018 [4], however, its status is still pending. This paper defined an LRP rule - the Basic Rule (Coined LRP-0 by [2]). This rule is the simplest rule, and therefore easiest to understand. LRP-0 works by providing a function to compute the relevance of a previous node given the activations and nodes following it. In layman's terms, the function computes a node's relevance by weighting (by the connection weights) the normalized (divided by all inputs to the next node) activations of the nodes of the next layer. This was the main finding of the paper, but the authors were far from finished.

2.2 Recent Advancements

Müller et al. [2] continued researching and provided an overview of all the rules they came up with in 2019. Complete with LRP-0, LRP-ϵ (LRP-0 with a regularization term in the denominator), LRP-γ (a rule that assigns more bias to positive effects than negative effects on relevance scores), LRP rules to deal with the non-smoothness of ReLU, and more, this paper served as a summary of findings of the entire project. They even mention an efficient implementation of their relevance score computation algorithm, and LRP for unsupervised models. Even though this paper wasn't able to provide complete insights and detailed analysis of LRP, it does provide a high-level overview with many possible variations for further research.

3 Experimental Setup

The experimental setup is defined by two main components.

- Dataset
- The classification method, in which a variety of network structures are used to predict on the MNIST dataset
- Testing LRP upon important vs unimportant feature removal

3.1 Dataset

To ensure stable results, a pre-cleaned high quality and accredited dataset was chosen. The MNIST dataset was used, and implemented in Python via PyTorch - specifically Torchvision's dataloader.

3.2 Classifier Setup

text ...

3.3 Feature Removal Setup

The purpose of feature removal is to test the implications of LRP. That is, we wish to ensure that highly "relevant" features are most if not solely important to image classification. To test this, we set up a system of feature removal. First, an image representation of a digit is classified. Using that classification, and as LRP specifies, we compute and assign relevance scores to each of the input features (pixels). Then, from this image, we perform two separate transformations. First, we obtain an important-feature image by removing 15% of unimportant (lowest relevance scoring) features. Similarly, we obtain an unimportant-feature image by removing 15% of the most important features from the original LRP image. Finally, these images are classified, and confidence values are computed for the prediction. This confidence, as well as the correctness of the prediction, are noted. Both of these metrics are used to draw conclusions about the interpretability of LRP, as well as how LRP can be used to lower data size and improve network efficiency.

4 Training

Text ... reference figure : fig 1

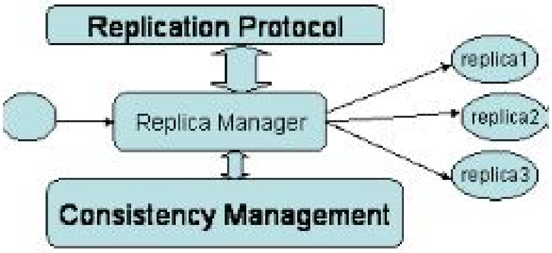

Figure 1: Replication Based Technique

text ...

5 Feature Removal Results

text ...

6 Evaluation

text ...

7 Conclusion

text ...

Major Factors	Replication Based	Check point	Fusion Based
Working	Request is redirected to replica	State saved on Stable storage Used for recovery	Back up machine
Consistency	Linearizi bility, Sequential	avoiding updated messages	Among Back up machine
Multiple Faults Handling	Depends upon No of Replica	Depends upon Checkpoints Scheduling	Depends Upon No of machines

text ...

8 Future Works

text ...

9 References

References

[1] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W 2015. *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.*, PLoS ONE 10(7): e0130140.

[2] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Lapuschkin S, Samek W. 2019. *Layer-Wise Relevance Propagation: An Overview.*, PLoS ONE 10.1007/978-3-030-28954-6_10.

[3] Deng, L. 2012. *The mnist database of handwritten digit images for machine learning research.*, IEEE Signal Processing Magazine 29(6), 141–142.

[4] Bach, S., Samek, W., Mueller, K.R., Binder, A. and Montavon, G. Fraunhofer Gesellschaft zur Forderung der Angewandten Forschung eV and Technische Universitaet Berlin. 2018. *Relevance score assignment for artificial neural networks.*, U.S. Patent Application 15/710,455.