Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

Stony Brook University

Introduction

Neural Networks 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 Layer-wise Relevance Propagation (LRP) have made 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.

Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

Stony Brook University

Motivation

Learning the contribution of each layer of the network and each part of the given input to the final output will help in

- Data Augumentation
- Better Feature Selection Methods
- Creating better Neural Network Architectures.

Implementation

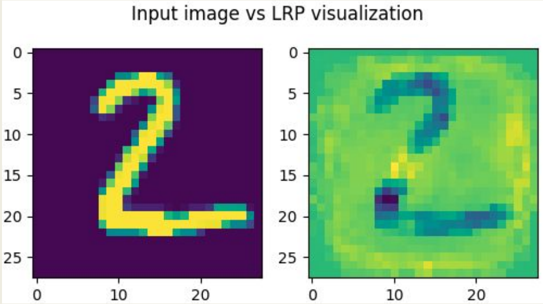
- Set up a pre-trained MNIST Classifier
- Build a wrapper for the built-in PyTorch classify function call such that we perform LRP subsequent to a classification
- Use various LRP rules to test relevant decision features.
  - LRP-0
  - LRP-ϵ
  - LRP-γ
  - LRP-Composite
- Test areas of importance

Layerwise Relevance Propagation - Implications for Feature Extraction

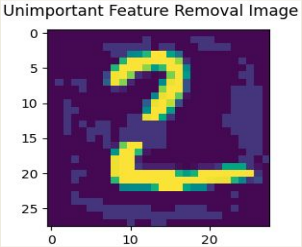
Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

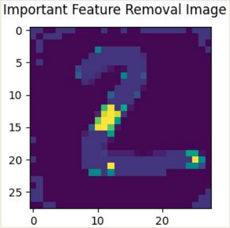
Stony Brook University

Feature Removal


Input image vs LRP visualization

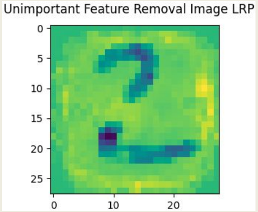
Original Confidence: 0.5657

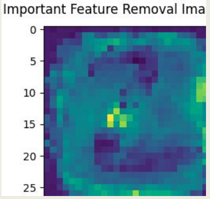

Unimportant Feature Removal Image


Important Feature Removal Image

Confidence: 0.4547

Confidence: 0.2009
and misclassified as 6


Unimportant Feature Removal Image LRP


Important Feature Removal Ima

Confidence Formula: (Max - 2nd-Max)/Max where max and 2nd max refer the the argmax of predicted class values and second argmax respectively.

Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

Stony Brook University

LRP Rules

LRP-0

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

LRP-γ

$$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^{+})}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^{+})} R_k$$

LRP-ϵ

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$$

LRP-Composite

$$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^{+}}{\sum_{0,j} (a_j w_{jk})^{+}} - \beta \frac{(a_j w_{jk})^{-}}{\sum_{0,j} (a_j w_{jk})^{-}} \right) R_k$$

Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

Stony Brook University

	Strongest Correct Classification	Strongest Incorrect Classification
LRP - 0	0.94473	0.07645
LRP-ϵ	0.96633	0.09543
LRP-γ	0.89288	0.08678
Composite LRP	0.79616	0.08536

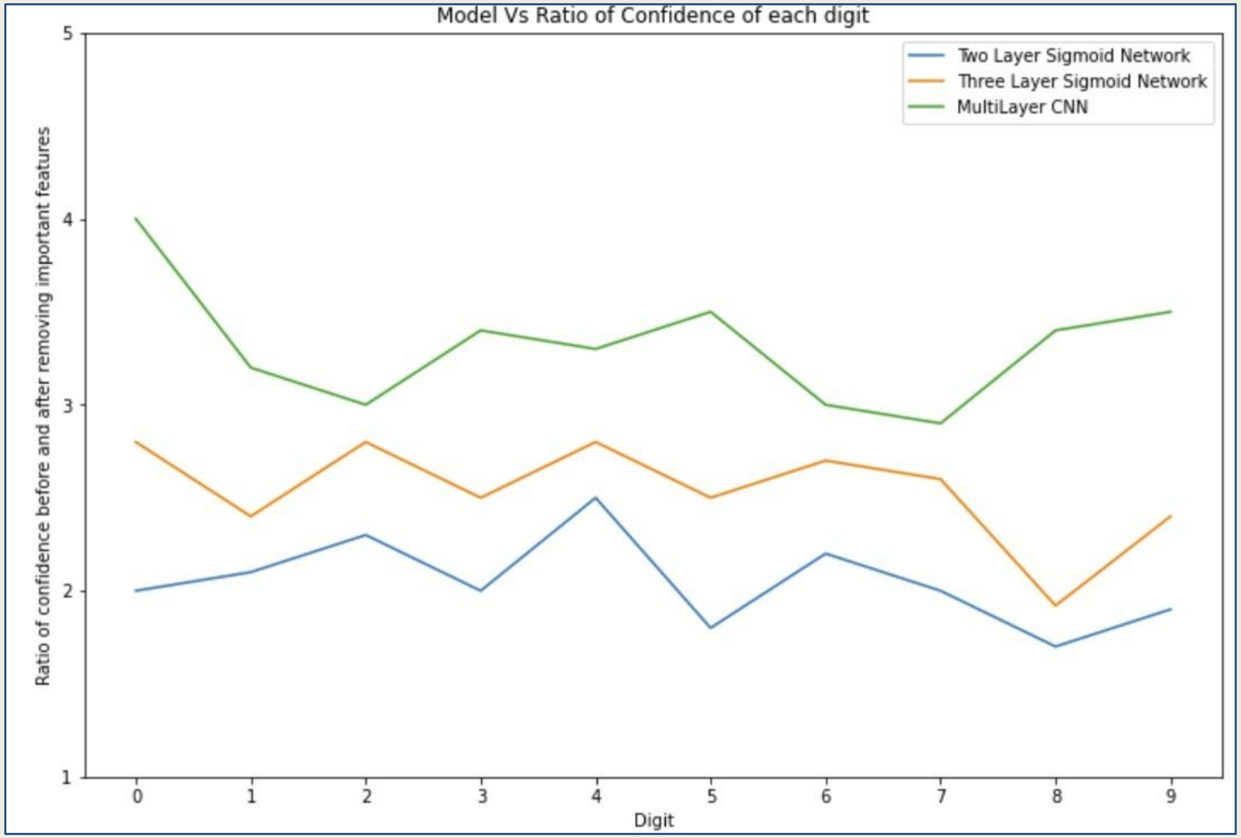
Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

Stony Brook University

Generalized Network Structure

2 Layer ReLu DNN vs 2 Layer ReLu DNN vs CNN



Layerwise Relevance Propagation - Implications for Feature Extraction

Carwyn Collinsworth | Gopi Sumanth Sanka | Akhil Arradi

Stony Brook University

Future Work

- Will be generalizing the model further by automating the neural network architecture for any kind of dataset.
- Auto tuning the hyperparameters
- Applying the same solution to detect different things like handwriting, objects , etc,. by making the model more generalized

Scan the QR code for more details