Sworn Declaration

I hereby declare under oath that the submitted Bachelor Thesis has been written solely by me without any third-party assistance, information other than provided sources or aids have not been used and those used have been fully documented. Sources for literal, paraphrased and cited quotes have been accurately credited.

The submitted document here present is identical to the electronically submitted text document.

Linz, August 21, 2023

Abstract

This paper focuses on the accuracy of interpretability methods for machine learning models. The main research problem is the lack of ground truth for evaluation method in interpretability methods. While there are inherent interpretative models, black-box networks perform better and have developed rapidly. While existing interpretability methods for neural networks such as [RoaR], [New Definitions and Evaluations for Saliency Methods: Staying Intrinsic, Complete and Sound] and [Sanity Checks for Saliency Maps] provide a numerical evaluation method, it is still doubtful that they really represent the ground truth.

This paper summarizes the different evaluation methods, compares them and adds a computational look at RoAR. The results show ...

Further research is needed on how to correctly evaluate interpretability methods.

Contents

List of Figures

1 Introduction

Artificial Intelligence has been developing rapidly in the last years. In today's modern era of mobile phones and computers, there has been a hugely increased volume of available data, which can be used through modern algorithm to improve the efficiency. Especially Deep Neural Networks have been gaining popularity due to the increasingly available computational power. The success of deep neural networks began in 2012, when the ImageNet classification challenge was won by a deep learning approach. In modern devices several AI-algorithms are already being used and in the future years AI-Algorithms will play an even more important role.

The biggest problem of Artificial Intelligence Methods is the lack of interpretation methods. While there exist many interpretation methods, there is no reliable truth. Especially in the domain of image, with technologies as automated driving and face recognition, evaluation methods are required. By now, we do not understand why a neural network makes a decision: It could be because of the background, a side object in the picture a high lighting in the image. Interpretability methods try to solve this problem by giving pixels an importance value, how much they influence the decision of the neural network.

1.1 Structure of the thesis

1. First of all, modern machine learning algorithms are summarized and categorized into inherent interpretability and Not inherent
2. Ad-Hoc interprability methods: An Overview
3. Evaluation of Post-Hoc Interpretation Methods
4. Applying ROAR to Food101: A Practical example

2 Machine Learning and their Interpretability

In supervised machine learning, there exist inherent interpretable models and so called black boxes. Inherent models as decision trees, decision rules, random forest and linear regression are interpretable by nature, as they learn rules which can be observed.

There also exist agnostic methods, which can be applied on any possible algorithm. There exist global methods, which try to explain the whole model and local methods, which try to explain the differences of 2 classes or just one area of the input.

2.1 Global Model-Agnostic Methods

Global model-agnostic methods describe expected outcomes based on the distribution of the data. They can show a correlation between singular or multiple features and an outcome.

An example offer partial dependency plots [4], Accumulated Local Effect (ALE) Plot [2], Feature Interaction [9], Functional Anova [6] &[7], Permutation Feature Importance [3], Global Surrogate and Prototype & Criticism [10].

2.1.1 Advantages, Disadvantages and Criticism

Advantages: For small models, few features, all described models offer a good solution.

Disadvantages: Limited scale for deep neural networks.

2.2 Local Model-Agnostic Methods

To explain individual predictions, Local model-agnostic methods are used. A single outcome is correlated to some features and explain the model.

Several methods exist: Individual conditional expectation curves Local surrogate models Scoped rules (anchors) Counterfactual explanations Shapley values SHAP

2.3 Neural Network Interpretation

In the domain of NLP and Computer Vision deep learning is very successful. By passing the data input through many layers of multiplication with learned weights and non-linear transformations a prediction is made. This can, depending on the task, include LSTMs and Convolution layers. Because there are millions of mathematical operations made in a single predictions, humans have no change to follow the exact mapping. To understand predictions, we would have to make sense of the hundreds different kernels and weights. To evaluate the behavior and predictions of Deep Neural networks, specific interpretation methods are needed, which take care of the pixel attribution.

Model-agnostic methods such as local models and global surrogate can be used to make sense of neural networks, but it makes sense to consider interpretation methods which were specifically designed for neural networks. The weights and kernels saved in the hidden layers can be used as additional information to evaluate the algorithm. Additionally, the gradients can be analyzed more effectively.

Because images and natural text are saved in a highly dimensional format, if converted to tabular form, most interpretation methods are not able to be used. Special neural network interpretation as described in the next chapter try to solve this problem.

2.3.1 Feature visualization and Network Dissection

Neural networks as big unit are not understandable for humans. Network dissection tries to abstract singular layers and link them to concepts.

The high-level features are linked to concepts, as visible in 2.1. The image is transformed every time it passes a constitutional layer. In each convolutional layer, the network learns new and increasingly complex features. Using fully connected layers, the transformed image information turns into a prediction.

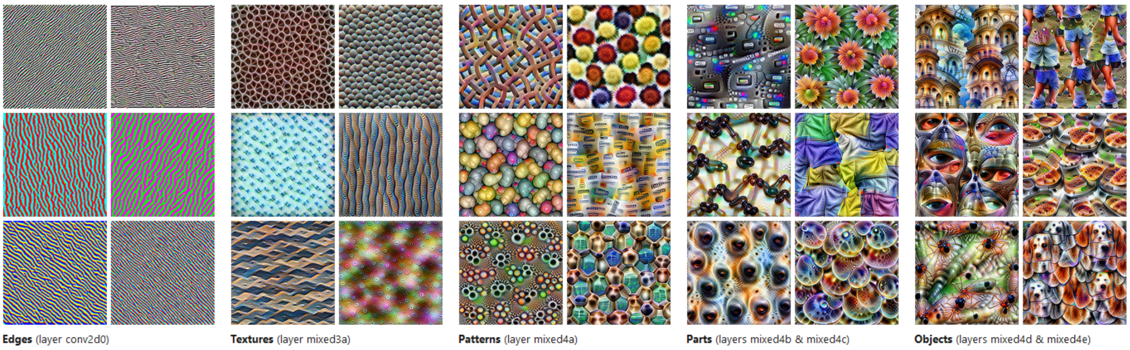

Figure 2.1: Feature Visualization https://distill.pub/2017/feature-visualization/

An example is visible in the image. Firstly, the convolutional layers learn simple features as edges and simple textures. then it learns textures and patterns. At the deepest convolutional layers, parts and objects are learned. Those object information are then passed to the hidden layers.

The feature visualization can be done through optimizing the activation of a singular unit. This is a single neuron. This can done in 2 methods: Either by finding the training image which maximizes the activation or by using prelabeled other images. To consider is also: Maximizing and minimizing here has the same effect. Using training data has the problem that if more than one object is visible, it is unclear which object is responsible for the maximization. Because of this, using prelabeled data is prefered.

2.3.2 Saliency Maps

2.3.3 Integrated gradients

2.4 Evaluation of post-hoc interpretability methods

Gradient-methods which generate saliency maps are hard to measure. While humans can evaluate the maps by giving a general statement, this is not a scientific statement and can also not be applied to thousands of images. Despite many significant recent contributions to saliency maps, the valuable effort of explaning machine learning models face this methodological challenge: the difficulty of assessing the scope and quality of model explanations [1].

Different methods have been proposed, which involve removing pixels [8] and [5].

2.4.1 RoAR

2.4.2 New Definitions and Evaluations for Saliency Methods: Staying intrinsic, Complete and Sound

2.4.3 Sanity checks for Saliency maps

2.4.4 Comparison of the Evaluation Methods & critique

3 Project work

3.1 Project Goal

3.1.1 Summary

3.1.2 Project Setup

3.1.3 Results and Plots

3.1.4 Discussion of results

JOHANNES KEPLER
UNIVERSITY LINZ

Bibliography

[1] Julius Adebayo et al. *Sanity Checks for Saliency Maps*. 2020. arXiv: `1810.03292 [cs.CV]`.

[2] Daniel W. Apley and Jingyu Zhu. *Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models*. 2019. arXiv: `1612.08468 [stat.ME]`.

[3] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. *All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously*. 2019. arXiv: `1801.01489 [stat.ME]`.

[4] Brandon M. Greenwell, Bradley C. Boehmke, and Andrew J. McCarthy. *A Simple and Effective Model-Based Variable Importance Measure*. 2018. arXiv: `1805.04755 [stat.ML]`.

[5] Arushi Gupta et al. *New Definitions and Evaluations for Saliency Methods: Staying Intrinsic, Complete and Sound*. 2022. arXiv: `2211.02912 [stat.ML]`.

[6] Giles Hooker. "Discovering additive structure in black box functions". In: Aug. 2004, pp. 575–580. DOI: `10.1145/1014052.1014122`.

[7] Giles Hooker. "Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables". In: *Journal of Computational and Graphical Statistics* 16 (2007), pp. 709–732. URL: `https://api.semanticscholar.org/CorpusID:10727333`.

[8] Sara Hooker et al. *A Benchmark for Interpretability Methods in Deep Neural Networks*. 2019. arXiv: `1806.10758 [cs.LG]`.

[9] Alan Inglis, Andrew Parnell, and Catherine Hurley. *Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models*. 2021. arXiv: `2108.04310 [stat.CO]`.

[10] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability". English (US). In: *Advances in Neural Information Processing Systems* (2016). 30th Annual Conference on Neural Information Processing Systems, NIPS 2016 ; Conference date: 05-12-2016 Through 10-12-2016, pp. 2288–2296. ISSN: 1049-5258.