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 25, 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. Existing interpretability methods for neural networks such as ROAR, KAR, BAM and Real Time Image Saliency for Black Box Classifiers provide a numerical evaluation method but they do still suffer from ambiguity. This paper summarizes the different evaluation methods, compares them and also calculate ROAR on 2 different saliency methods.

Contents

List of Figures

1 Introduction

Artificial Intelligence (AI) has undergone rapid development in the last years. In today's modern era of mobile phones and computers, algorithm's are used on a daily basis to have quick access to information and improve the efficiency of the daily life.

While various Algorithms (e.g.: Decision Trees, Linear Regression, Support Vector Machines, etc.), which are understandable by design, have been developed, the spotlight has turned to Deep Neural Networks(DNN). This shift is attributed to the increase in computational power and the exponential increase in available data. Despite their remarkable accuracy, DNN remain opaque black-boxes, which we are struggling to understand. Nevertheless, the immense improvement in performance and their ability to handle massive datasets have led to widespread adoption in contemporary devices. It is predicted, that algorithms which are based on Neural Networks will be becoming increasingly popular in the next years.

However, one of the primary difficulties with Neural Networks is the lack of reliable interpretation techniques. Numerous interpretation methods exist, yet a universally reliable method remains missing. Particularly in the domain of image analysis, encompassing critical applications like automated driving and facial recognition, no solution is present. The decision-making rationale of neural networks remain unclear, attributed to factors like background elements, peripheral objects or lighting conditions. Efforts to address this issue have given rise to saliency methods, aiming to assign significance values to pixels and represent their importance on neural network decisions. Another alternative option to mitigate the black-box nature of algorithms involves employing model-agnostic methods. These methods offer an computational linkage between inputs and outputs, irrespective which model is used. Although highly effective for smaller datasets, they begin to struggle as the data size and their complexity increases. Because of this, they do not offer a reliable way to quickly make Neural Networks interpretable.

In light of these prevalent problems, the object of this thesis is to recapitulate the existing model-agnostic methods and offer an overview of interpretation methods for Neural networks. Emphasis is placed on the evaluation of post-hoc Interpretability techniques, forecasting potential future developments and focusing on the strengths and weaknesses of distinct techniques. Concluding the theoretical segment, a practical demonstration showcasing the application of ROAR is shown.

1.1 Structure of the thesis

1. The first part presents a comprehensive overview of contemporary machine learning algorithms, categorizing them into two main groups: algorithms with inherent interpretability and those without.
2. Subsequent sections delve into interpretability, emphasizing global model-agnostic techniques. These methods offer insights into overall model behavior, regardless of algorithm specifics.
3. In parallel with the discussion on global methods, the focus shifts to local model-agnostic approaches. These strategies inspect individual predictions, enhancing the understanding of model decision-making at a granular level.
4. Additionally, in the domain of interpretability techniques for Neural Networks, the paper explores ad-hoc methods for Neural Networks.
5. After introducing existing interpretation methods, the paper's focus transitions to evaluating post-hoc interpretation methods. Various approaches to assess the effectiveness and dependability of these methods in offering meaningful insights into intricate models are introduced and discussed. Additionally, the advantages and disadvantages of these approaches are carefully examined to provide a comprehensive understanding of their applicability.
6. To exemplify the discussed concepts, the practical application of the ROAR framework using the Food101 dataset and MNIST dataset is presented. This real-world instance illustrates effective employment of interpretability techniques in image recognition.

2 Machine Learning and their Interpretability

This paper focuses on two types of machine learning methods: Unsupervised and Supervised. However, we'll only look at supervised methods here because interpreting unsupervised methods works differently at a basic level.

In supervised machine learning, there are several base methods of classification. They are shortly introduced.

2.1 Supervised Methods

This section analyzes the base functionality of each singular model. Furthermore, an analysis of the interpretability from a human perspective is made. Methods are developing rapidly, therefore only the most common methods are included.

2.1.1 Linear Models

When it comes to predicting outcomes, a simple method is to use a linear regression model. This model predicts by adding up different features, each multiplied by a weight. The linear nature of this model makes it easy to understand. Mathematically, the predictive output, denoted as $\hat{y}$, is captured in the equation:

$$\hat{y} = \alpha_0 + \alpha_1 x_0 + \alpha_2 x_1 + ... + \alpha_n x_{n-1} + \epsilon$$

The alphas α_i indicate the significance of each feature. The initial coefficient α_0 is known as the intercept, signifying the baseline. The noise ϵ encapsulates the inevitable errors stemming from inherent non-linearity in real-world dynamics or measurement inaccuracies.
To train model, the MSE-Loss or the absolute loss can be applied. When using regularization methods, the absolute loss is taken to be more resilient to outliers.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{ABS} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i$$

The interpretability of the model is very simple. The factors are given through the coefficient matrix. Each feature has a distinctive importance to the model and it can be seen easily how important each factor is, when the data is normalized.

$$\alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \ldots \\ \alpha_n \end{bmatrix}$$

To customize the model and avoid over-fitting, regularization methods as Lasso-Regularization (L1) and Ridge-Regularization (L2) can be applied. The parameter lambda can be fine tuned.

$$\text{Lasso Loss} = \text{Loss} + \lambda_1 \sum_{i=1}^{p} |\alpha_i|$$

Lasso regularization is particularly useful when you want to emphasize a subset of relevant features or when dealing with multicollinearity issues. By encouraging certain coefficients to become zero, Lasso can lead to a more interpretable and sparse model.

$$\text{Ridge Loss} = \text{Loss} + \lambda_2 \sum_{i=1}^{p} \alpha_i^2$$

Ridge regularization is particularly effective when dealing with multicollinearity (highly correlated features) and helps prevent over-fitting by keeping the coefficients from taking large values. It doesn't force coefficients to be exactly zero, but it pushes them towards zero, striking a balance between fitting the data and preventing over-fitting.

Although linear models possess comprehensibility and provide a straightforward method for prediction and are inherent understandable, their application is limited to linear relationships. Using an regularization term does not have a effect on the interpretability.

2.1.2 Distance-Based methods

K-Nearest Neighbors is used for classification and uses the nearest neighbars as classification. KNN is not interpretable by default, as there are no parameters to learn and analyze. One can argue, that KNN is interpretable by the fact, that it just describes if there are similar samples.

Support Vector Machines (SVM) aim to find a hyperplane that maximizes the margin between different classes of data points.

$$\text{Minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{Subject to } y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \text{ for } i = 1, \ldots, n$$

With using the kernel trick, SVM can be used for non-linear data. In higher dimensionality, SVM becomes non-interpretable, as displaying the weight matrix is not understandable.

2.1.3 Decision Tree-Based Methods

Decision Trees based on minimizing the gini-index are inherently interpretable. As the depth increases, the models become less understandable. One disadvantage is the lack of smoothness. If one boundary is reached, the model classification changes.

Random forests are an ensemble of multiple decision trees. Their advantage is a smoother predictive power. But it suffers from being less interpretable. Using special techniques like SHAP values or partial dependence trees can make them more understandable.

Gradient boosting like XGBoost, LightGBM and CatBoost are similar to random forest and decision trees, but with differential learned weights for each decision. They suffer from the same interpretability issues as random forests.

2.1.4 Probabilistic Methods

Logistic Regression, a classic classification technique, and Naive Bayes, rooted in Bayesian principles, offer unique perspectives on model interpretability. Much like Decision Trees based on minimizing the Gini-index, these methods hold inherent interpretability due to their fundamental concepts. However, they possess nuances that influence their ease of understanding.

Logistic Regression's interpretability springs from its linear nature and probabilistic foundation. Coefficients associated with features serve as clear indicators of influence. Positive coefficients denote a positive relationship, while negative coefficients signify the opposite. The magnitude of coefficients reveals the strength of the relationship. This inherent linearity promotes transparency but may falter when dealing with non-linear relationships.

Naive Bayes, grounded in probability theory, facilitates understandable reasoning. Conditional probabilities for features given the class illuminate the likelihood of feature occurrences within specific classes. Feature independence assumption simplifies computations, making probabilities intuitive. However, the "naive" assumption might not align with real-world relationships, presenting interpretability challenges.

Yet, as with Decision Trees, model complexity can hamper interpretability. Deeper Logistic Regression models may obscure transparency as intricate interactions emerge. Likewise, the simplicity of Naive Bayes might not capture complex data relationships.

2.1.5 Neural Networks

The rise of neural networks and their strong predictive power makes them a common choice for classification task. But as they increase in size, understandability by taking a look at the weights becomes impossible. Special interpretability methods considering the learned weights are looked in detail in 3.3.

2.1.6 Discriminant Analysis, LDA, QDA

Discriminant Analysis, encompassing Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), offers a distinctive approach to classification tasks. In this section, we delve into the interpretability of these methods, comparing their strengths and limitations.

Understanding Discriminant Analysis:

Discriminant Analysis is designed to find the optimal boundaries that separate classes. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) extend this concept with varying degrees of complexity.

Interpretability of Linear Discriminant Analysis (LDA):

LDA focuses on projecting data onto a lower-dimensional space while maximizing class separability. This inherently linear approach simplifies interpretation.

Projection and Decision Boundaries: The projection aims to maximize the distance between class means while minimizing the variance within each class. Decision boundaries emerge as linear lines or planes. This linearity enhances interpretability by offering clear visual separations.

Feature Importance: LDA computes linear coefficients for features that contribute to separating classes. These coefficients offer insights into feature significance.

Class Means: The class means’ positions and distances provide insights into how classes are distributed in the projected space.

Interpretability of Quadratic Discriminant Analysis (QDA):

QDA relaxes the assumption of equal covariances across classes, accommodating quadratic relationships. While more flexible, it introduces complexity that can affect interpretability.

Quadratic Decision Boundaries: QDA's quadratic decision boundaries allow capturing more intricate class relationships. However, these boundaries might not be as straightforward to interpret as LDA's linear counterparts.

Covariance Matrices: The separate covariance matrices for each class reflect their distinct data distributions. Analyzing these matrices can provide insights into how features contribute to class separability.

Trade-offs and Techniques:

While LDA promotes interpretability through its linear projections and decision boundaries, QDA's increased flexibility might challenge transparency. The choice between the two depends on the trade-off between interpretability and the model's ability to capture complex relationships.

Both LDA and QDA can benefit from techniques like feature importance analysis, visualization of decision boundaries, and examination of class means. However, QDA's additional complexity might necessitate more advanced visualization techniques.

In summary, Discriminant Analysis, LDA, and QDA offer varying degrees of interpretability. LDA's linear approach enhances understanding, while QDA's flexibility introduces nuanced insights at the cost of increased complexity. Techniques that aid interpretation, such as visualizations and feature analyses, empower users to extract meaningful insights from these classification methodologies.

3 Interpretation of models

3.1 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.

3.2 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.

3.3 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.

3.3.1 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 3.1. 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.

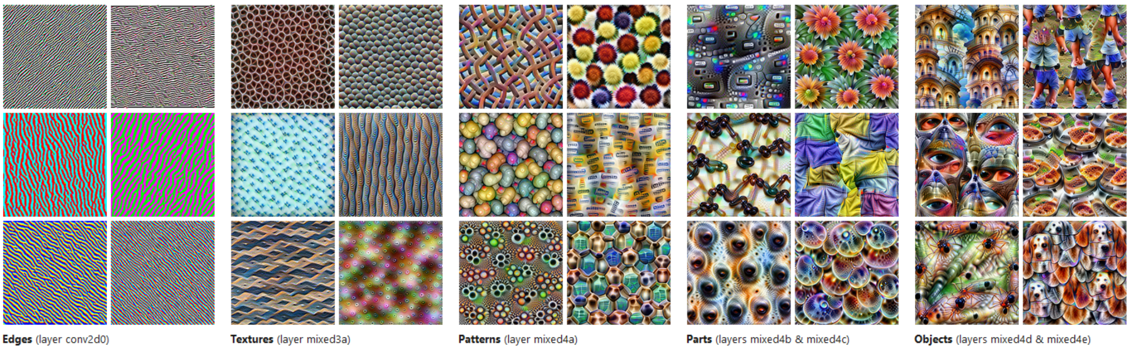

Figure 3.1: Feature Visualization https://distill.pub/2017/feature-visualization/

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.

3.3.2 Saliency Maps

Saliency maps are visualizations that highlight the regions of an input image that have the most significant impact on a model's output. By revealing the areas that strongly influence a prediction, saliency maps bridge the gap between the model's "black-box" nature and human understanding.

3.3.3 Integrated gradients and Grad-CAM

Gradient-based saliency methods tap into the gradients of the model's output with respect to the input data. The magnitudes of these gradients represent the importance of each input pixel in influencing the prediction. Common techniques, like Grad-CAM (Gradient-weighted Class Activation Mapping), amplify this understanding by overlaying saliency maps on the original image.

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].

4.1 A benchmark for interpretability methods in deep neural networks

RoaR is a algorithm which estimates the effectiveness of saliency maps. This is done, by removing supposedly informative features from the input and observing the reaction of the neural network.

As shown in the paper, KAR seems to be a bad method to evaluate the efficiency.

4.2 New Definitions and Evaluations for Saliency Methods: Staying intrinsic, complete and sound [5]

Explaining Completeness:

Logical reasoning: All correct statements are proveable. Example: The dog is responsible for the nets output as dog.

Soundness: Incorrect statements cannot be proved. Example: The sun is responsible for the nets output as dog.

4.3 Sanity checks for Saliency maps

4.4 Comparison of the Evaluation Methods & critique

While the described methods both offer a numerical evaluation method, they do still lack a clear evaluation structure. They show, that some evaluation methods are indeed correct, but they still suffer ambuigity from different datasets.

5 Project work

5.1 Project Goal

In RoaR[8] the paper does not list the standard deviation of the trained nets. We expect to validate the results by achieving similar results.

As training 25 image nets requires high computational power we do not have right now, we limited our research to evaluating food-101 using only 2 interpretation methods and comparing it to the baseline.

Additionally, we also add a mini evaluation using the MNIST dataset.

5.1.1 Project Setup

5.1.2 Results and Plots

5.1.3 Discussion of results

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.