Explaining the Black Box: A Study on the Application of Explainable Artificial Intelligence (XAI) Techniques to Machine Learning Models

Magdalena Pakuła* and Jakub Pawlak†

WFTIMS, Politechnika Łódzka
Email: *254220@edu.p.lodz.pl, †254222@edu.p.lodz.pl

***Abstract*—The increasing reliance on artificial neural networks in various applications has raised concerns about their *black box* nature, making it challenging to understand the decision-making processes behind their predictions. This report addresses this challenge by exploring the application of so-called Explainable Artificial Intelligence (XAI) techniques to machine learning models. Specifically, we employ local explanation approaches, including attributions and counterfactual examples, using the Captum library in the PyTorch such as LIME, saliency maps or Integrated Gradients techniques. Our study aims to shed light on the previously trained models, providing a deeper understanding of how they make predictions and present those findings in this report**

I. Introduction

ARTIFICIAL neural networks (ANNs) have revolutionized various industries and applications, from healthcare to finance, by enabling accurate predictions and decision-making. However, the increasing reliance on these models has also raised concerns about their *black box* nature, making it challenging to understand the decision-making processes behind their predictions. This opacity can lead to a lack of trust in the models, as well as difficulties in identifying biases and errors. To address this challenge, Explainable Artificial Intelligence (XAI) techniques have emerged as a crucial step towards building more transparent and accountable machine learning models. By applying XAI methods, we can gain insights into the internal workings of complex models, providing a deeper understanding of how they make predictions and enabling more informed decision-making. This project aims to explore the application of XAI techniques to machine learning models, specifically focusing on local explanation approaches such as attributions and counterfactual examples. By shedding light on the previously trained models, we aim to provide a better understanding of how they make predictions and contribute to the development of more transparent and trustworthy AI systems.

II. Related Work

During this study, we explored various methods used to enhance the interpretability and transparency of ANNs through XAI techniques. The first paper in this domain we examined was [**?**] by Ribeiro et al. This paper was the first to publicly introduce the Local Interpretable Model-Agnostic Explanations (LIME) framework, which generates local, interpretable explanations for the predictions of any machine learning model. The key idea behind LIME is to approximate the original complex model with an interpretable model that is locally faithful to the prediction of interest. This paper was an important step in the development of XAI techniques, as it demonstrated the feasibility and utility of interpretable explanations. Since then, the field of XAI has continued to evolve, with many other techniques (e.g. SHAP, TCAV, GradCAM) being proposed to enhance the transparency and interpretability of AI systems. However, LIME provides local explanations that are specific to individual predictions and while useful, these local explanations may not offer a comprehensive understanding of the model's overall behavior [**?**].

Different approach for understanding model predictions are saliency maps thoroughly explained by Simonyan et al. [**?**]. This paper presents the method, which computes the gradient of the score of the predicted class with respect to the input image. The gradient essentially measures how small changes in each pixel value of the image will affect the output score. The absolute values of these gradients are taken to represent the importance of each pixel and pixels with larger gradient magnitudes are considered more important for the classification decision. Therefore, saliency maps effectively highlight discriminative image regions and objects, which are easily interpretable by a human eye. Such examples are shown and prove the usefulness of this method.

To have a broader view and better understanding using another method we reviewed another influential paper by Sundararajan et al. [**?**], which provided an alternative surrogate-based methods like LIME or sometimes misleading saliency maps [**?**]. The authors introduce and demonstrate the usefulness of Integrated Gradients through various case studies, including image classification. They show that this method can provide meaningful and intuitive explanations of the model's predictions, which can be valuable for tasks like feature importance analysis. Additionally, the authors compare Integrated Gradients method to other gradient-based attribution methods, such as Saliency Maps and Deconvolution, and highlight the advantages of it in terms of satisfying desirable properties for explanations.

Furthermore, counterfactuals are an essential aspect of XAI, as they provide a way to evaluate the robustness of a model's predictions by considering what would have happened if

certain conditions had been different [**?**]. This is particularly important in high-stakes applications, such as healthcare or finance, where the consequences of incorrect predictions can be severe. For example, Adebayo et al. [**?**] evaluated the reliability of saliency maps on the MNIST and CIFAR-10 image classification datasets by applying counterfactual perturbations like rotation and occlusion to the input images. They found that many popular saliency map methods failed to provide faithful explanations, underscoring the importance of using counterfactuals to assess the robustness of model predictions.

Those papers provide a solid ground for experimenting on different models with different methods in order to fully understand models predictions.

III. Methods

This experiment will be conducted on various types of data and ANN models for testing the versatility of different XAI methods. Models used in this experiment are:

- Three MLP models for classifying iris flowers, distinguishing wines and diagnosing breast cancer
- Three MLP models for classifying handwritten numbers from the MNIST dataset
- CNN model for the MNIST dataset
- CNN model for distinguishing objects in images from the CIFAR10 dataset

After thoroughly reviewing research paper, we decided to firstly understand what was important in making a decision for a specific input by finding attributions and counterfacts using the following methods:

- **LIME:** is a model-agnostic technique that explains the predictions of any classifier by learning an interpretable model locally around the prediction. It approximates the decision boundary around a given input $\mathbf{x}$ using a linear regression model:

  $$f(\mathbf{x}') = \boldsymbol{\theta} \cdot \mathbf{x}' + b$$

  where $\boldsymbol{\theta}$ are the coefficients of the linear model and b is the bias term. By analyzing $\boldsymbol{\theta}$, LIME identifies the important features that influenced the prediction $f(\mathbf{x}')$.
- **Saliency Maps:** highlight input features that are most influential for a model's prediction. They are computed by taking the absolute gradients of the output $f(\mathbf{x})$ with respect to the input $\mathbf{x}$:

  $$S(\mathbf{x}) = \left| \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right|$$

  These maps provide insights into how sensitive the model's output is to changes in each input feature $\mathbf{x}_i$.
- **Integrated Gradients:** attribute an output prediction to input features by integrating the gradients of the output $f(\mathbf{x})$ with respect to the input $\mathbf{x}$ along a linear path from a baseline input $\mathbf{x}'$ to the actual input $\mathbf{x}$:

  $$\phi_i(\mathbf{x}) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}_i} \, d\alpha$$

  This method provides a comprehensive attribution of the model's prediction to each feature $\mathbf{x}_i$, reflecting how each feature contributed to the final prediction.
- **Minimal Parameter Perturbation:** is a technique that identifies the smallest possible changes to the model parameters that can significantly change the model's output prediction. By finding these minimal perturbations, we can gain insights into the model's decision-making process and the features that are most important for its predictions.

Additionally, we will use SLIC (Simple Linear Iterative Clustering), which is a superpixel segmentation algorithm that partitions an image into compact, nearly uniform superpixels $\{S_k\}$. It achieves this by iteratively clustering pixels based on their color similarity and spatial proximity, producing segments that enhance the interpretability of XAI methods:

$$S(\mathbf{x}) = \arg\min_{S_k} \left(\|\mathbf{x} - \mathbf{m}_k\|^2 + \frac{\beta}{N_k} \sum_{\mathbf{x}' \in S_k} \|\mathbf{x} - \mathbf{x}'\|^2 \right)$$

where $\mathbf{m}_k$ is the mean color of superpixel S_k, N_k is the number of pixels in S_k, and β balances color proximity and spatial proximity.

After presenting the result of this experiment, we will try to discover the general principles that guide the models in order to understand the overall behavior and assumptions of the trained models.

IV. Results

Present the results of the local and global explanations. Interpret the findings, discuss their significance, and potential implications.

A. Iris MLP classifier

B. Wine MLP classifier

C. Breast cancer MLP classifier

D. MNIST MLP classifier

1) Flatten:

2) LBP:

3) HOG:

E. MNIST CNN classifier

The CNN for recognizing handwritten digits from the MNIST dataset was analyzed using the saliency methods. The resulting saliency maps are visualized on fig. 1. The blue pixels show the areas that most influence the model's predictions.

Unexpectedly, the obtained saliency maps predominantly highlight background pixels adjacent to the digits rather than the digits themselves. One plausible explanation of this phenomenon is that the CNN has learned the absence of pixels in a specific region as impotant information for its classification decisions — for example, the digits 4 and 5 are characterized by the lack of the closed regions. If one was to add pixels to these areas 4 would turn into a 9, and 5 would turn into a 6. Therefore, these areas are highlited as important to the classification results, because it is essential that no pixels exist in those areas. Such behaviour might suggest that the model is not recognizing the digits by their shape directly, but rather

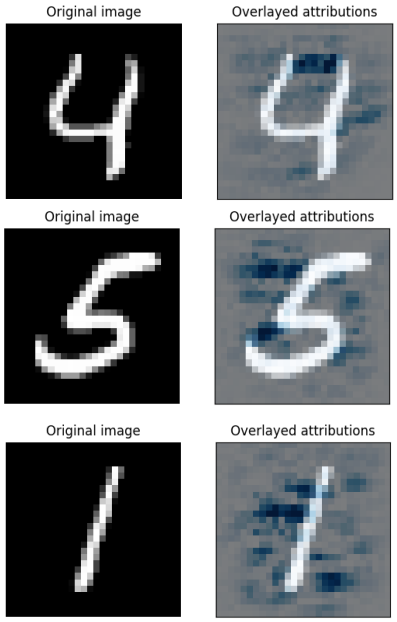

Fig. 1. Saliency maps of chosen MNIST digits

by ellimination — i.e. an important characteristic of a digit 4 is that it is not a 9.

The background pixels might also provide contextual information about the area around the digit. This is especially impornant in shape recognition tasks, such as digit recognition. For example, the top bar at the digit 5 is oriented to the right-hand side, however, there are pixels highlited to the left-hand side. That is because for a line to be considered as oriented to the right, there need to be both a presence of pixels to the right, and a lack of them to the left. Therefore, the background also needs to be considered in order to correctly determine the shape characteristics and boundaries.

Another unfortunate possibility of such behaviour may be some issues with the model. Overfitting could cause the model to focus on some background pixels, that are not really representative of the digit features. However, it should be noted that during the model evaluation, we did not observe a drop in accuracy on the test set during training — a typical sign of overfitting. Additionally, the performance of the model on the testing data was satisfactory, which speaks against the potential issues with the model.

F. CIFAR10 CNN classifier

V. Summary

Summarize the main conclusions and suggest future work.

VI. References