An XAI-based masking approach to improve classification systems

Andrea Apicella[1,2,3,*,†], Salvatore Giugliano[1,2,3,†], Francesco Isgrò[1,2,3,†], Andrea Pollastro[1,2,3,4,†] and Roberto Prevete[1,2,3,†]

[1]*Laboratory of Augmented Reality for Health Monitoring (ARHeMLab)*
[2]*Laboratory of Artificial Intelligence, Privacy & Applications (AIPA Lab)*
[3]*Department of Electrical Engineering and Information Technology, University of Naples Federico II*
[4]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

Abstract

Explainable Artificial Intelligence (XAI) seeks to elucidate the decision-making mechanisms of AI models, enabling users to glean insights beyond the results they produce. While a key objective of XAI is to enhance the performance of AI models through explanatory processes, a notable portion of XAI literature predominantly addresses the explanation of AI systems, with limited focus on leveraging XAI methods for performance improvement. This study introduces a novel approach utilizing Integrated Gradients explanations to enhance a classification system, which is subsequently evaluated on three datasets: Fashion-MNIST, CIFAR10, and STL10. Empirical findings indicate that Integrated Gradients explanations effectively contribute to enhancing classification performance.

Keywords

XAI, Machine Learning, DNN, Integrated Gradients, attributions

1. Introduction

Explainable Artificial Intelligence (XAI) plays a crucial role in understanding the decision-making processes of AI models, especially as they become integral to critical applications in healthcare, finance, and everyday life. While existing XAI literature primarily focuses on providing explanations for AI systems, there's a notable gap in leveraging these explanations to enhance the performance of the models. This paper addresses this gap by examining established an XAI method commonly employed in Machine Learning (ML) classification tasks. The goal is to utilize explanations for model improvement. The core concept hinges on the idea that explanations about model outputs offer insights to fine-tune the ML system parameters effectively. However, interpreting Deep Neural Networks (DNNs) can be challenging due to their

2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-23, co-located with AIxIA 2023, Roma Tre University, Roma, Italy, 2023

[*] Corresponding author.

[†] These authors contributed equally.

andrea.apicella@unina.it (A. Apicella); salvatore.giugliano2@unina.it (S. Giugliano); francesco.isgro@unina.it (F. Isgrò); andrea.pollastro@unina.it (A. Pollastro); rprevete@unina.it (R. Prevete)

0000-0002-5391-168X (A. Apicella); 0000-0002-1791-6416 (S. Giugliano); 0000-0001-9342-5291 (F. Isgrò); 0000-0003-4075-0757 (A. Pollastro); 0000-0002-3804-1719 (R. Prevete)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3615/short2.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

inherent complexity, demanding explanations that are human-readable. This work operates on the premise that explanation-derived knowledge can be harnessed to comprehend the model's strengths and weaknesses, thereby enhancing its adaptability to various inputs. In this context, explanations are constructed based on the behavior of the ML system, shedding light on its input-output relationships. Consequently, they enable the identification of input characteristics influencing outputs, thereby empowering adjustments to the ML system itself. This paper specifically delves into the exploration of Integrated Gradient [1] XAI method to assess whether the relevant features it highlights can be used in conjunction with input data to augment the classification performance of an ML system. The results of this approach have been more extensively treated in [2].

2. Related works

The internal mechanisms of modern ML approaches, particularly in the realm of Deep Learning, often remain opaque, making it challenging for AI scientists to fully grasp the underlying processes guiding their behaviors. The utilization of XAI methods has gained prominence in providing explanations for various classification systems across domains like images [3, 4, 5, 6, 7], natural language processing [8, 9], clinical decision support systems [10], and more. In particular, in [1] Integrated gradient was proposed, an XAI method that involves calculating the average of gradients between an input $\mathbf{x}$ and a reference $\mathbf{x}^{ref}$, where $C(\mathbf{x}^{ref})$ yields a given model to a neutral prediction. This approach, termed Integrated Gradient (IG), considers the magnitude of gradients of features of inputs closer to the baseline. The significance of each feature x_i is determined by aggregating the gradients along the intermediate inputs on the straight-line path connecting the baseline and the input. However, the application of XAI methods to enhance the performance of ML models in classification tasks is a relatively underexplored area in current research. A survey in [11] provides an overview of works leveraging XAI methods to improve classification systems. Furthermore [12, 13, 2] conduct an empirical analysis of several well-known XAI methods on an ML system trained on EEG data, showing that many components identified as relevant by XAI methods can potentially be employed to build a system with improved generalization capabilities. In contrast, the primary focus of the current study is to assess the effectiveness of selected XAI methods in enhancing the performance of a machine learning system for image classification tasks. Additionally, the study delves into various strategies for integrating input data and explanations to optimize the ML system's performance. The detailed results have been further elaborated in [2], where they are also compared with an alternative strategy.

3. Method

This study endeavors to propose a viable method for leveraging an XAI explanation to enhance the performance of a classifier. However, it is essential to note that our approach begins with the premise that, for a specific input, an explanation of the model's output for the correct target class is accessible. While this assumption may not hold in real-world scenarios where the correct class for new input is unknown, it is a starting point for effectively investigating the potential

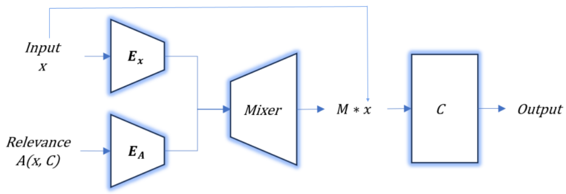

Figure 1: Architecture of the soft-masking schema.

improvement in classification performance through the utilization of explanations. We suggest a potential approach for integrating IG explanations into the classification process through a *soft-masking* scheme. In essence, we make a model able to combine the relevance $A(\mathbf{x}, C)$ with the input $\mathbf{x}$. To accomplish this, we introduce an additional mixer network, denoted as the *Mixer*, which is connected to the classifier C, as illustrated in Fig. 1. We employ two additional networks, $E_{\mathbf{x}}$ and E_A, to reduce the dimensionality of $\mathbf{x}$ and $A(\mathbf{x}, C)$ respectively. The outputs of $E_{\mathbf{x}}$ and E_A are then concatenated and fed into the Mixer. The resulting output of the Mixer can be interpreted as an input mask M, which is used to weight the input $\mathbf{x}$ for classifier C. The parameters of Mixer, $E_{\mathbf{x}}$, and E_A can be learned while keeping the parameters of C fixed. This involves employing standard training procedures on the non-fixed parameters, effectively searching for the optimal set of parameters for Mixer, $E_{\mathbf{x}}$, and E_A that effectively reduce and integrate $A(\mathbf{x}, C)$ and $\mathbf{x}$ for a given classifier C.

4. Experimental assessment

Fashion-MNIST [14], CIFAR10, and STL10 datasets were used as benchmark datasets, while ResNet18 [15] pre-trained on ImageNet dataset was adopted as classifier C for the CIFAR10 and STL10 dataset, and a two fully-connected layers Neural Network equipped with ReLU activation function for Fashion-MNIST dataset. Baselines was computed fine tuning C with the training set provided in each adopted dataset. Then, for each input and baseline the Integrated Gradient explanation have been built. The architectures adopted for $E_{\mathbf{x}}$ and E_A are reported in Tab.

STL10		***CIFAR10***			***Fashion-Mnist***	
$E_{\mathbf{x}}, E_A$	**Mixer**	$E_{\mathbf{x}}, E_A$	**Mixer**	$E_{\mathbf{x}}, E_A$	**Mixer**	C
FC 4096	FC 512	FC 2048	FC 512	FC 512	FC 512	FC 128
batch norm.+ReLU	batch norm.+ReLU	batch norm.+ReLU	batch norm.+ReLU	batch norm.+ReLU	batch norm.+ReLU	ReLU
FC 2048	FC 1024	FC 1024	FC 1024	FC 256	FC 784	FC 64
batch norm.+ReLU	batch norm.+ReLU	batch norm.+ReLU		batch norm.+ReLU		ReLU
FC 1024	FC 4096	FC 512		FC 128		FC 10
batch norm.+ReLU	batch norm.+ReLU	batch norm.+ReLU				
FC 512	FC 9216	FC 256				
batch norm.+ReLU		batch norm.+ReLU				
FC 256		FC 128				
batch norm.+ReLU						
FC 128						

Table 1
Architectures adopted. For each Fully-Connected (FC) layer, the numbers indicate how many neurons are employed. The C module adopted for CIFAR10 and STL10 was a ResNet18 pretrained on ImageNet.

1. The training consisted in training the Mixer network, $E_{\mathbf{X}}$, and E_A while freezing the C

Model	*CIFAR10*	*STL10*	*Fashion-MNIST*
baseline	85.7 %	66.3 %	87.3 %
proposed	**87.6** %	**68.6** %	**99.9** %

Table 2
Accuracy scores on test set using the soft masking scheme.

parameters. The training was made with the Adam algorithm and a validation set of 30% of the training data to stop the iterative learning process. Best batch size and learning rate were found with a grid-search approach, with batch sizes $\{64, 128, 256\}$, learning rates in range $[0.001, 0.01]$ with step of 0.02.

5. Results & conclusions

In Tab. 2 the results of the proposed schema are reported. It is highlighted that the proposed strategies lead to an improvement in accuracy in all the investigated datasets. The proposed approach offers a strategy to effectively integrate explanations with input data, leading to enhanced model classification performance. This is achieved by allowing the model to autonomously determine the optimal mixing strategy through a learning process. The results demonstrate promise in the experimental scenario for all the investigated datasets. It's important to note, however, that all results are derived under the assumption that accurate explanations for the correct classes are available for the test data. This assumption, while useful for this study, is unrealistic in practice since the true class of test data is typically unknown. Therefore, the findings of this research can pave the way for the development of a system that can provide reliable approximations of explanations even in the testing phase. We intend to further explore and expand upon this avenue in our future research endeavors.

Acknowledgments

This work is supported by the European Union - FSE-REACT-EU, PON Research and Innovation 2014-2020 DM1062/2021 contract number 18-I-15350-2, and was partially supported by the Ministry of University and Research, PRIN research project "BRIO – BIAS, RISK, OPACITY in AI: design, verification and development of Trustworthy AI.", Project no. 2020SSKZ7R, by the Ministry of Economic Development, "INtegrated Technologies and ENhanced SEnsing for cognition and rehabilitation" (INTENSE) project, and by Centro Nazionale HPC, Big Data e Quantum Computing (PNRR CN1 spoke 9 Digital Society & Smart Cities, CUP: E63C22000980007). Furthermore, we acknowledge financial support from the PNRR MUR project PE0000013-FAIR (CUP: E63C22002150007).

References

[1] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[2] A. Apicella, L. Di Lorenzo, F. Isgrò, A. Pollastro, R. Prevete, Strategies to exploit xai to improve classification systems, in: Explainable Artificial Intelligence. xAI 2023, Springer Nature Switzerland, 2023, pp. 147–159.
[3] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[4] A. Apicella, F. Isgrò, R. Prevete, A. Sorrentino, G. Tamburrini, Explaining classification systems using sparse dictionaries, ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2019).
[5] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 193–209.
[6] A. Apicella, S. Giugliano, F. Isgrò, R. Prevete, A general approach to compute the relevance of middle-level input features, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Springer, 2021, pp. 189–203.
[7] A. Apicella, S. Giugliano, F. Isgro, R. Prevete, et al., Explanations in terms of hierarchically organised middle level features, in: CEUR WORKSHOP PROCEEDINGS, volume 3014, CEUR-WS, 2021, pp. 44–57.
[8] K. Qian, M. Danilevsky, Y. Katsis, B. Kawas, E. Oduor, L. Popa, Y. Li, Xnlp: A living survey for xai research in natural language processing, in: 26th International Conference on Intelligent User Interfaces-Companion, 2021, pp. 78–80.
[9] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, arXiv preprint arXiv:1606.04155 (2016).
[10] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, K. Van Den Bosch, Human-centered xai: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies 154 (2021) 102684.
[11] L. Weber, S. Lapuschkin, A. Binder, W. Samek, Beyond explaining: Opportunities and challenges of xai-based model improvement, Information Fusion (2022).
[12] A. Apicella, F. Isgrò, A. Pollastro, R. Prevete, Toward the application of XAI methods in eeg-based systems, in: Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence(AIxIA 2022), Udine, Italy, November 28 - December 3, 2022, volume 3277 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 1–15.
[13] A. Apicella, F. Isgrò, R. Prevete, XAI approach for addressing the dataset shift problem: BCI as a case study (short paper), in: Proceedings of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2022) co-located with the 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, Italy, December 2, 2022, volume 3319 of *CEUR Workshop Proceedings*, 2022, pp. 83–88.
[14] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, arXiv preprint arXiv:1708.07747 (2017).
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.