Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230416

A Study on the Reliability of Visual XAI Methods for X-Ray Images

Jan STODT[a,1] and Manav MADAN[a] and Christoph REICH[a], and Luka FILIPOVIC[b] and Tomi ILIJAS[c]
[a] *Institute for Data Science, Cloud Computing, and IT Security; Furtwangen University; Furtwangen; Germany*
[b] *University Donja Gorica, 81000 Podgorica, Montenegro*
[c] *Arctur d.o.o., SI-5000 Nova Gorica, Slovenia*

Abstract. The YOLO series of object detection algorithms, including YOLOv4 and YOLOv5, have shown superior performance in various medical diagnostic tasks, surpassing human ability in some cases. However, their black-box nature has limited their adoption in medical applications that require trust and explainability of model decisions. To address this issue, visual explanations for AI models, known as visual XAI, have been proposed in the form of heatmaps that highlight regions in the input that contributed most to a particular decision. Gradient-based approaches, such as Grad-CAM [1], and non-gradient-based approaches, such as Eigen-CAM [2], are applicable to YOLO models and do not require new layer implementation. This paper evaluates the performance of Grad-CAM and Eigen-CAM on the VinDrCXR Chest X-ray Abnormalities Detection dataset [3] and discusses the limitations of these methods for explaining model decisions to data scientists.

Keywords. Visual xai, yolo, unreliability

1. Introduction

The transparency and interpretability of artificial intelligence (AI) systems have become increasingly important as AI is becoming more prevalent in various domains, such as healthcare. To ensure the trust in AI decisions, visual explanation for AI models (XAI) methods have emerged as a valuable tool. Despite their potential benefits, visual XAI methods still face several challenges regarding their reliability and accuracy. Overfitting and subjectivity are two of the main issues, which can lead to inaccurate explanations and reduce the confidence in the methods. This paper aims to demonstrate the lack of reliability of visual XAI methods by Grad-CAM (Gradient-weighted Class Activation Mapping) and Eigen-CAM (Eigen-based Class Activation Mapping) on a YOLOv5x model trained on the VinDr-CXR Chest X-ray Abnormalities Detection dataset [3]. The results of the comparison will shed light on the limitations of the visual XAI methods and highlight the need for a comprehensive evaluation of the understandability of the explanations to improve the practical application of AI.

[1] Corresponding Author: Jan Stodt, Jan.Stodt@hs-furtwangen.de.

2. Related Work

Grad-CAM (Gradient-weighted Class Activation Mapping) [1] and Eigen-CAM (Eigenbased Class Activation Mapping) [2] are visualization methods to highlight the regions of an image that contribute the most to a CNN (convolutional neural network) decision. Grad-CAM produces a heatmap that highlights the image regions that have the highest gradient (which requires backpropagation) in context to the final layer of CNN. EigenCAM visualizes the components of learned features by a CNN from the final convolutional layer. This translates to being computationally efficient with no modification, retraining, and backpropagation of gradients required. Two studies have compared EigenCAM and Grad-CAM. Muhammad et al. [4] compared Eigen-CAM to Grad-CAM. They found that Eigen-CAM is more consistent, able to differentiate between classes more effectively, and less affected by errors in the dense layers of a model. They showed an improvement of up to 15% compared to Grad-CAM. Rahman et al. [5] compared the effectiveness of Grad-CAM and Eigen-CAM by applying them to the YOLOv5 (You Only Look Once) detection model. It is concluded that Eigen-CAM performs well, but Grad-CAM is better for low-light conditions. Muhammad et al. provide results based on empirical methods and demonstrates a clear improvement. For Rahman et al. it is unclear which goal the comparison is intended to achieve as they do not provide ground truth.

3. Lack of Reliability of Visual XAI Methods

The visual explanation for AI Models decisions (XAI) is rapidly nowadays for transparency, interpretability, and for strengthening trust in AI systems. However, despite their high popularity and their potential benefits, visual XAI faces several challenges regarding the reliability and accuracy of the explanation. One of the main issues is overfitting [6], which occurs when the method is too much optimized for a single model, thus leading to inaccurate explanations for other models. This problem is intensified when the model is trained on biased data, as visual explanations might amplify these biases and lead to incorrect interpretations of the model behavior [7]. Additionally, XAI methods can be subjective, as their explanations may be influenced by the perspective and background of the person interpreting the explanation. This can lead to contradictory results and thus reduce the confidence in the explanation and lead to the fact that these methods are no longer used [8]. Complex models, such as DNN (deep neural network), lead to limited interpretability, and therefore again limiting the reliability of the explanation of the model's behavior [9].

4. Comparison of Visual XAI Methods

To demonstrate the lack of reliability of visual XAI methods, the Eigen-CAM and GradCAM were applied on a YOLOv5x model, which is a single-stage detector. The model was trained on the VinDr-CXR Chest X-ray Abnormalities Detection dataset [3] using 4392 images to localize and classify 14 kinds of thoracic abnormalities. The developed model available as a PyTorch model has been fine-tuned on a pretrained YOLOv5X model [10], which was trained on the COCO dataset [11]. The best tuned model resulted in the mAP@0.5 value of 61% and the precision of 72% on the test set

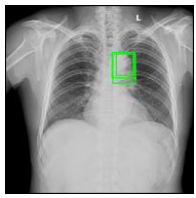

Figure 1. Ground Truth.

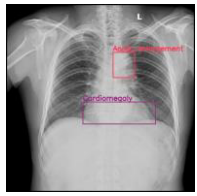

Figure 2. Detection.

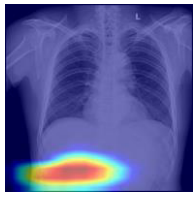

Figure 3. Eigen-CAM.

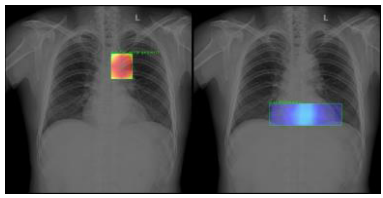

Figure 4. Grad-CAM.

containing 2966 images. The ground truth (Figure 1) marks three bounding boxes indicating the presence of aortic enlargement. The model detects (Figure 2) one aortic enlargement correct (true positive), overlooks two aortic enlargements (false negative), as well as one being cardiomegaly (false positive). We show the lack of reliability of visual XAI methods by comparing the results from Eigen-CAM (Figure 3) and Grad-CAM (Figure 4) to the ground truth and the model's detection. When comparing the detection with the GradCAM output, it can be seen that the high activation is accurately shown for the detection of aortic enlargement (true positive). For the cardiomegaly (false positive), GradCAM correctly shows the low activation by the model, indicating the wrong detection. In contrast, the Eigen-CAM output fails to explain the detection of the aortic enlargement, since it does not highlight the correct area for the aortic enlargement. Additionally, since Eigen-CAM returns the first principal component of the activations, and thus the dominant object, there is no direct link between the observed XAI output and the model detection. As demonstrated, it is important to be cautious when relying solely on the output generated by visual XAI. It is recommended to verify their accuracy by testing with some sample images and conducting a manual evaluation of the result. An alternative approach could be to use an automated evaluation method that compares the highlighted areas with the ground truth and the detection. This can help determine which XAI tool is most suitable for a specific domain and model. The comparison of XAI approaches revealed that while there may be more precise methods, they may be difficult or impossible to apply to newer YOLO models because of skip connections and newer architectural designs such as CSP (cross stage partial) connections. To use gradient-based approaches, these newer structural changes must be reimplemented to allow for the backpropagation of the error signal to determine the most activated pixels in the input.

5. Conclusion

In conclusion, the popular visual XAI methods face several challenges. Challenges such as overfitting, subjectivity, and limited interpretability of complex models can lead to inaccurate and contradictory results. To address these challenges, our comparison of the Eigen-CAM and Grad-CAM methods on a YOLOv5x model demonstrates the

importance of being cautious when relying solely on the output generated by visual XAI tools. Verifying the accuracy through sample images and manual evaluations, or using an automated evaluation method, is recommended to determine the most suitable XAI tool for a specific domain and model. While there might be more precise XAI approaches, adapting them to newer YOLO models can be challenging, and further reimplementation may be required.

6. Acknowledgment

The paper is funded by the DigNest project https://dignest.me for which the results of this paper will be disseminated at seminars and workshops. The authors also would like to acknowledge the support from the team of Project Q-AMeLiA (funded by MWK BW, reference: 7547.223-6/12/4).

References

[1] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 618-26.

[2] Muhammad MB, Yeasin M. Eigen-Cam: Class Activation Map Using Principal Components. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. p. 1-7.

[3] Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. VinDr-CXR: An Open Dataset of Chest X-rays with Radiologist's Annotations. Scientific Data. 2022;9(1):1-7.

[4] Bany Muhammad M, Yeasin M. Eigen-CAM: Visual Explanations for Deep Convolutional Neural Networks. SN Computer Science. 2021;2(1):1-14. Available from: https://doi.org/10.1007/s42979-021-00449-3.

[5] Rahman AN, Andriana D, Machbub C. Comparison between Grad-CAM and EigenCAM on YOLOv5 Detection Model. In: 2022 International Symposium on Electronics and Smart Devices (ISESD); 2022. p. 1-5.

[6] Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning – A Brief History, State-of-theArt and Challenges. In: Koprinska I, Kamp M, Appice A, Loglisci C, Antonie L, Zimmermann A, et al., editors. ECML PKDD 2020 Workshops. Communications in Computer and Information Science. Springer International Publishing; 2020. p. 417-31.

[7] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In: Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc.; 2016. p. 1-9. Available from: https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[8] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv; 2017. Available from: http://arxiv.org/abs/1702.08608.

[9] Montavon G, Samek W, Muller KR. Methods for Interpreting and Understanding Deep Neural Net-¨works. Digital Signal Processing. 2018;73:1-15. Available from: https://www.sciencedirect.com/science/article/pii/S1051200417302385.

[10] Ibrahim M. Yolov5-VinBigData. Kaggle; 2021. Available from: https://kaggle.com/code/mostafaibrahim17/yolov5-vinbigdata.

[11] Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, et al.. Microsoft COCO: Common Objects in Context; 2014. Cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list. Available from: http://arxiv.org/abs/1405.0312.