Experimental Analysis of Four CAM Techniques

Izabella Kamila Ożóg
Student Number: 220997300
ec22400@qmul.ac.uk

***Abstract*—In this project, I performed an experimental analysis of four important Class Activation Mapping (CAM) techniques. The objectives were to conduct a quantitative analysis of the techniques, visualize heatmaps on different convolutional neural network (CNN) layers, and explain a texture classifier. Each of the objectives yielded valuable insights into the performance of the techniques.**

I. Introduction

As the applications of artificial intelligence (AI) became more and more widespread, it also became crucial to understand how AI makes decisions. This is especially relevant for fields such as medicine or law, where the incorrect decisions can be severe. This is the reason for the rapid development of hundreds new eXplainable AI (XAI) methods in recent years [7].

Class Activation Mapping (CAM) techniques are a popular post-hoc explanation method for visualising the reasoning of CNNs. In this project, I analysed four important CAM techniques: Grad-CAM, Grad-CAM++, Score-CAM and Relevance-CAM.

II. Problem Definition

Obj. 1: Quantitative analysis

One of the biggest issues with XAI methods is the difficulty of evaluating their correctness due to the lack of ground truth. As a result, XAI methods are often evaluated only with anecdotal evidence, lacking quantitative, objective analysis [7]. It is important to verify XAI methods in-depth to ensure they are not producing misleading results.

Obj. 2: Visualisation of heatmaps on different CNN layers

Additionally, CAM is mostly used to produce heatmaps for the final convolutional layer, without analysing other layers. Analysing more layers has the potential to provide more insight into AI reasoning.

Obj.3: Provide explanations for a texture classifier

To my knowledge, there has been limited research on applying XAI techniques to texture images. Such images containing repetitive patterns can provide insights into XAI method performance.

III. Key Works

The focus of this project is on four CAM methods introduced in the following works: Grad-CAM [9], Grad-CAM++ [2], Score-CAM [11], and Relevance-CAM [5]. To evaluate the performance of these methods quantitatively, I employed the Quantus library [4] and applied metrics from five categories: region perturbation [9] (from the "faithfulness" category), continuity [6] ("robustness" category), complexity [1] ("complexity"), non-sensitivity [6] ("axiomatic"), and random logit [10] ("randomisation").

IV. Evaluation Criteria

The CNN model was trained on the Describable Textures Dataset DTD [3] using ten randomly selected texture categories out of 47: “bubbly”, “crosshatched”, “fibrous”, “honeycombed”, “knitted”, “lined”, “polka-dotted”, “veined”, “waffled” and “zigzagged”. Transfer learning with a pretrained ResNet50 model was used for training. For XAI evaluation, one representative of each class was picked (for total of 10 test images). Each of the CAM methods was applied twice, once for the intermediate layer and once for the last convolutional layer. At the end, using Quantus library, the five selected metrics were calculated.

The evaluation for this project included a visual comparison of the heatmaps produced by the CAM techniques and ranking the methods based on the metrics calculated.

V. Discussion

The discussion is divided into two sections. In the first one, the heatmaps produced are compared and in the second one calculated metrics are presented.

A. Qualitative analysis

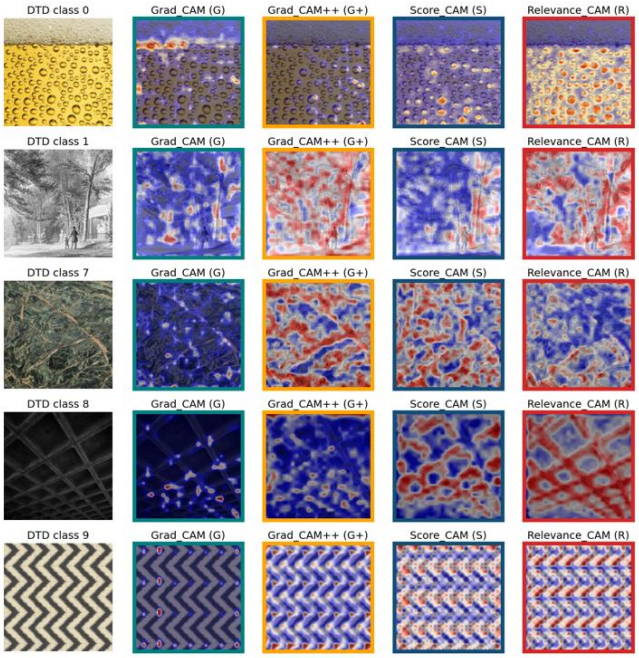

Fig. 1. Explanations produced by each method at the intermediate layer.

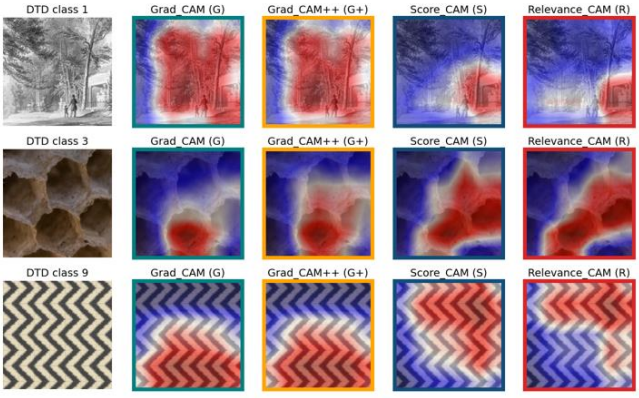

Fig. 2. Explanations produced by each method at the last layer.

The heatmaps shown in Fig. 1 are more detailed than those in Fig. 2. Activations in intermediate layers capture more information about the input image.

Fig. 1 demonstrates that Relevance-CAM generally performs better at deeper layers. This is due to Layer-wise Relevance Propagation (LRP), which addresses the problem of shattered gradients. Other methods often suffer from noisy gradients at deeper layers due to this problem. Interestingly, Relevance-CAM performs worse on classes 1 and 7, where there are no simple objects like bubbles present, as in class 0.

Fig. 2 presents heatmaps for the last layer. It is difficult to identify any method that clearly outperforms the others. However, it is noticeable that Grad-CAM and Grad-CAM++ highlight similar regions, and Score-CAM and Relevance-CAM also highlight similar regions. This is because they have similar methods for generating these highlights. Grad-CAM and Grad-CAM++ use gradients of the output class with respect to activations in the last layer, while Score-CAM and Relevance-CAM both base their explanations on the target class. Score-CAM focuses on activations that most contribute to the prediction of the target class, while Relevance-CAM focuses on activations that are most relevant to the target class. The similarity regions were captured clearly due to the use of texture images.

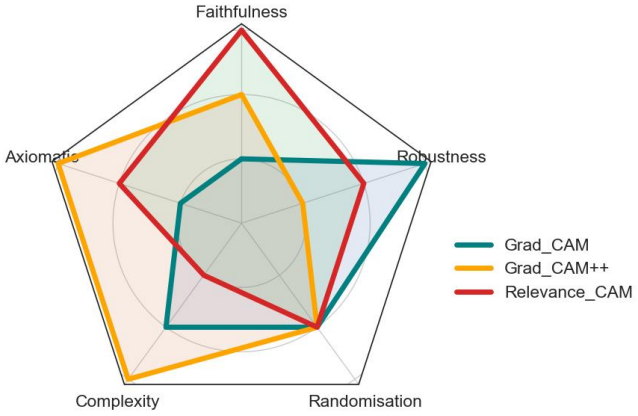

Fig. 3. Spyder graph visualising the ranking of the CAM techniques based on all five evaluation metrics calculated. The further a score is from the central part of the graph, the better its rank.

TABLE I. METRICS SCORES RANKED

Methods	**Metrics**				
	Faith.	***Axiom.***	***Robust.***	***Complex.***	***Random.***
Grad_CAM	1	1	3	2	2[a]
Grad_CAM++	2	3	1	3	2[a]
Relevance_CAM	3	2	2	1	2[a]

[a] All the methods had the same results in *Randomisation*.

Unfortunately, due to high computational requirements for calculating metrics of Score-CAM performance, I was not able to perform it.

The "Faithfulness" metric measures whether the explanations follow the predictions correctly [4]. The Region Perturbation [9] measures how the class disappears from the image as more and more information disappears. Relevance-CAM scored best result and Grad-CAM worst.

The "Axiomatic" category contains metrics measuring certain axiomatic properties [4]. I used Non-Sensitivity [6] metric that measures sensitivity to changes in input data. Similarly to Faithfulness, Grad-CAM received lowest rank.

"Robustness" measures stability of explanations under input perturbations, assuming the output stayed the same [4]. The Continuity [6] metric calculated the strongest variation of explanation with formula:

$$R(x) \square R(x') \square\square\square\square\square \quad \square\square\square$$

"Complexity" describe and compare. "Randomisation"

VI. CONCLUSION

Drawbacks: no comparison to other methods in section B. Hard to evaluate to what extend results are similar using ranking. Also, Score-CAM not included. Interesting results in section A, especially Relevance-CAM on Fig.1 class 1 and 7. Thanks to using texture observed similarities on Fig.2.

REFERENCES

[1] Bhatt, U., Weller, A. and Moura, J.M. (2020) "Evaluating and aggregating feature-based model explanations," Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence [Preprint]. Available at: https://doi.org/10.24963/ijcai.2020/417.

[2] Chattopadhay, A. et al. (2018) "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) [Preprint]. Available at: https://doi.org/10.1109/wacv.2018.00097.

[3] Cimpoi, M. et al. (2014) "Describing textures in the wild," 2014 IEEE Conference on Computer Vision and Pattern Recognition [Preprint]. Available at: https://doi.org/10.1109/cvpr.2014.461.

[4] Hedström, A. et al. (2022) "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations," arXiv [Preprint]. Available at: https://doi.org/https://doi.org/10.48550/arXiv.2202.06861.

[5] Lee, J.R. et al. (2021) "Relevance-cam: Your model already knows where to look," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Preprint]. Available at: https://doi.org/10.1109/cvpr46437.2021.01470.

[6] Montavon, G., Samek, W. and Müller, K.-R. (2018) "Methods for interpreting and Understanding Deep Neural Networks," Digital Signal Processing, 73, pp. 1–15. Available at: https://doi.org/10.1016/j.dsp.2017.10.011.

[7] Nauta , M. et al. (2022) "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," arXiv [Preprint]. Available at: https://doi.org/10.48550/arXiv.2201.08164.

[8] Samek, W. et al. (2017) "Evaluating the visualization of what a deep neural network has learned," IEEE Transactions on Neural Networks and Learning Systems, 28(11), pp. 2660–2673. Available at: https://doi.org/10.1109/tnnls.2016.2599820.

[9] Selvaraju, R.R. et al. (2017) "Grad-cam: Visual explanations from deep networks via gradient-based localization," 2017 IEEE International Conference on Computer Vision (ICCV) [Preprint]. Available at: https://doi.org/10.1109/iccv.2017.74.

[10] Sixt, L., Maximilian , G. and Tim , L. (2019) "When Explanations Lie: Why Many Modified BP Attributions Fail," International Conference on Machine Learning (ICML) [Preprint]. Available at: https://proceedings.mlr.press/v119/sixt20a.html

[11] Wang, H. et al. (2020) "Score-cam: Score-weighted visual explanations for Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) [Preprint]. Available at: https://doi.org/10.1109/cvprw50498.2020.00020