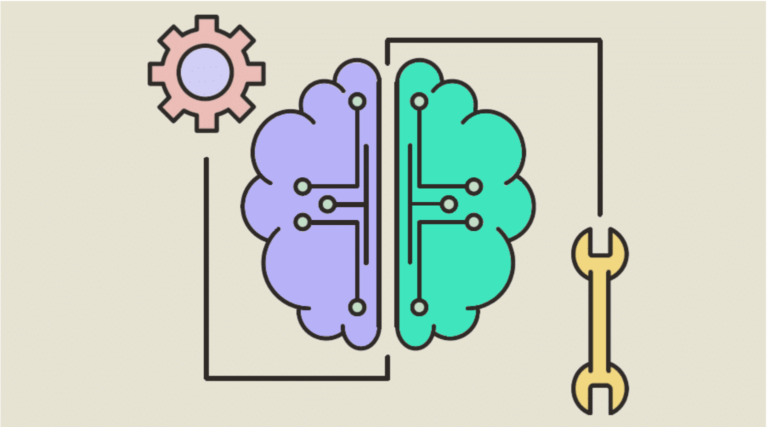
Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks



[10].

Oriol Miró López-Feliu
Marc Gonzalez Vidal

Computational Vision
Master in Artificial Intelligence (MAI)

December 18, 2024

1 Introduction

Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance in various computer vision tasks such as image classification, object detection, and segmentation. Despite their success, the "black box" nature of these models raises concerns regarding interpretability and trustworthiness, especially in critical applications like medical diagnosis and autonomous driving. Understanding the internal decision-making process of CNNs is essential for building reliable systems.

This report focuses on the Score-CAM method proposed by Wang et al. [1], which aims to provide visual explanations for CNN predictions without relying on gradient information. Using the forward pass of the network, Score-CAM generates more accurate and less noisy saliency maps compared to previous methods. We will analyse the problem addressed by Score-CAM, discuss related work, the methodology, present results and discussions, relate the method to course material, and conclude with insights and new directions post-Score-CAM.

2 The Problem

As CNNs become increasingly complex, interpreting their decisions becomes more challenging. Existing methods for visual explanations fall into three main categories:

- **Gradient-based methods**: Use gradients to identify important input regions but suffer from noise and issues like vanishing or exploding gradients. Gradients can become saturated, leading to explanations that are noisy or lack interpretability.
- **Perturbation-based methods**: Modify input images and observe changes in output but are computationally intensive due to the large number of forward passes required.
- **CAM-based methods**: Generate class activation maps but often depend on specific network architectures and gradients, limiting their general applicability.

The main problem addressed in the paper is to develop a visual explanation method that:

1. Avoids reliance on gradients, addressing issues like noise and saturation.
2. Requires no architectural modifications, ensuring broad applicability.
3. Produces accurate, interpretable saliency maps to enhance transparency.

3 State of the Art

3.1 Gradient-based Methods

Simonyan et al. [2] proposed the use of gradients to create saliency maps by computing the derivative of the class score with respect to the input image. While simple, these methods often produce noisy explanations.

Guided Backpropagation [3] and Integrated Gradients [12] attempt to improve gradient-based explanations by modifying the backpropagation process or integrating gradients along a path. However, these methods still suffer from sensitivity to gradient issues.

3.2 Perturbation-based Methods

Methods like RISE [5] generate explanations by randomly masking parts of the input image and observing the impact on the output. While effective, these methods require numerous forward passes, making them computationally expensive.

3.3 Class Activation Mapping (CAM) Methods

CAM [6] generates visual explanations by combining activation maps with weights from fully connected layers. Grad-CAM [7] generalizes CAM to any CNN architecture by using gradients to compute weights, but it inherits gradient-related limitations.

Grad-CAM++ [8] improves upon Grad-CAM by considering higher-order derivatives, but it still depends on gradients and can produce less focused explanations.

4 Methodology

Score-CAM addresses the limitations of previous methods by eliminating the dependency on gradients. Instead, it uses the increase in the target class score when specific regions of the input are highlighted. The method involves the following steps:

4.1 Activation Map Extraction

Given an input image X and a CNN model f, activation maps A^k are extracted from a convolutional layer l. Each activation map corresponds to a particular feature or pattern learned by the network. It is recommended to consider the features maps of the last layers, as it is end point of feature extraction.

4.2 Mask Generation

Each activation map A^k is upsampled to the input size using bilinear interpolation and normalized to the range $[0, 1]$ to create a mask M^k:

$$M^k = \text{Normalize}(\text{Upsample}(A^k))$$

The normalization ensures that the mask weights are scaled appropriately, highlighting regions corresponding to high activations.

4.3 Forward Passes with Masked Inputs

For each mask M^k, a masked input image $\tilde{X}^k$ is created by element-wise multiplication:

$$\tilde{X}^k = X \odot M^k$$

By masking the input, we isolate the contribution of specific regions highlighted by each activation map. Each masked input $\tilde{X}^k$ is fed into the model to obtain the class score S_k^c for the target class c.

4.4 Weight Computation

The weight α^k for each activation map is computed based on the increase in the target class score compared to a baseline image X_b (e.g., a black image):

$$\alpha^k = \frac{\exp(S_k^c - S_b^c)}{\sum_j \exp(S_j^c - S_b^c)}$$

Here, S_b^c is the class score for the baseline image. The use of the exponential function and softmax normalization ensures that weights are positive and sum to one, capturing the relative importance of each activation map in enhancing the target class score.

4.5 Saliency Map Generation

The final saliency map $L^c_{\text{Score-CAM}}$ is computed as a weighted combination of the activation maps:

$$L^c_{\text{Score-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A_l^k\right)$$

The ReLU function ensures that only positive contributions are considered, focusing on features that positively influence the target class prediction.

4.6 Intuition Behind the Method

By measuring the increase in class score when specific regions are highlighted, Score-CAM captures the actual contribution of those regions to the model's decision. This approach avoids gradient-related issues and provides a more direct assessment of feature importance.

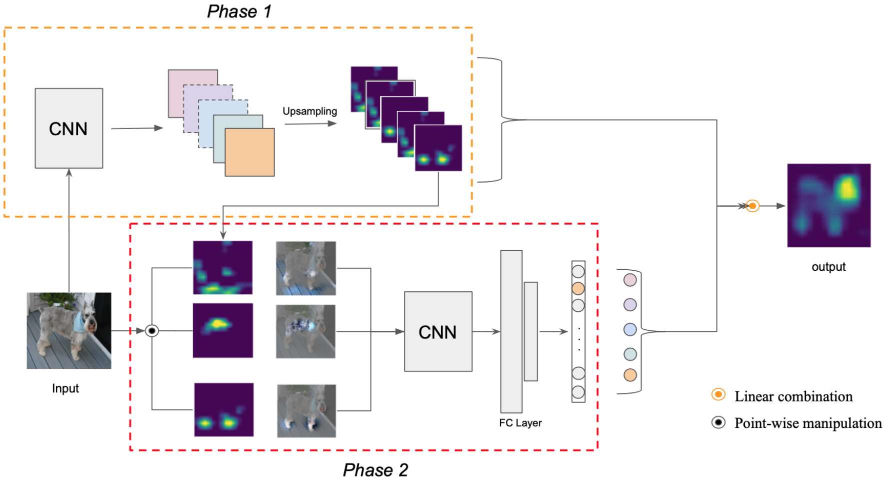


Figure 1: Pipeline of the proposed Score-CAM method. Activation maps are upsampled to create masks, which are then used to generate masked inputs. Class scores from masked inputs are used to compute weights for the activation maps, leading to the final saliency map.

5 Results, Discussions, and Limitations

5.1 Qualitative Results

Score-CAM produces saliency maps that are more focused and less noisy than those generated by alternative methods. Figure 2 illustrates a comparison of various state-of-the-art methods, including Vanilla Backpropagation, Guided Backpropagation, SmoothGrad, Grad-CAM, and Grad-CAM++. The proposed Score-CAM method demonstrates superior performance, providing clearer and more precise localization of relevant regions while reducing noise. Additional results are available in the Appendix.

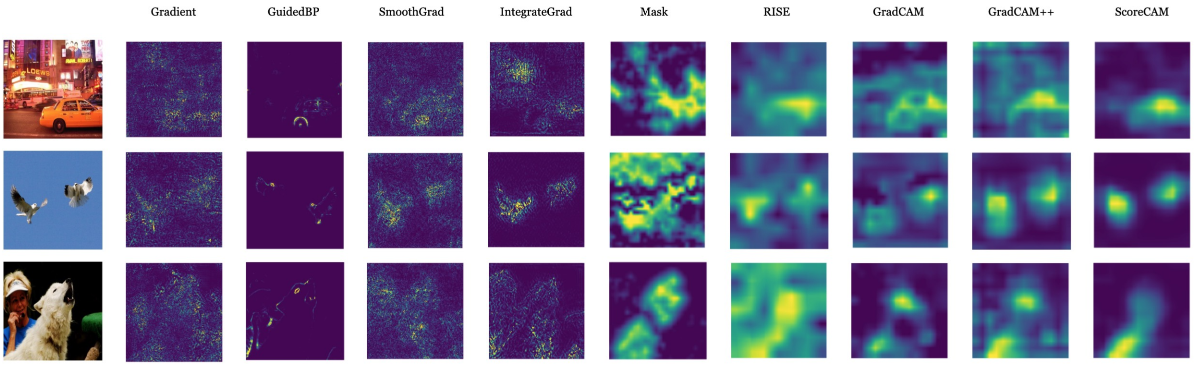


Figure 2: Visualization results of Vanilla Backpropagation [?], Guided Backpropagation [3], SmoothGrad [11], IntegrateGrad [12], Mask [?], RISE [5], Grad-CAM [7], Grad-CAM++ [8], and the proposed Score-CAM. The proposed method demonstrates clearer and more precise localization of relevant regions compared to baseline approaches. Additional results are included in the Appendix.

In these examples, gradient-based methods often produce scattered and less focused maps, whereas Score-CAM effectively highlights the most important areas contributing to the model's prediction.

5.2 Quantitative Results

Score-CAM is evaluated using the following metrics to assess how effectively saliency maps capture regions important for the model's decisions:

- **Average Drop (%)**: Measures the reduction in the model's confidence for the target class when the highlighted regions are removed or masked, with lower values indicating that the highlighted regions are essential for the class prediction.
- **Average Increase (%)**: Quantifies the increase in the model's confidence for the target class when the highlighted regions are used as input, with higher values confirming that the highlighted regions are crucial for the model's decision-making.
- **Localization Accuracy (%)**: Measures the overlap between the saliency map and ground truth bounding boxes or segmentation masks.

Table 1: Quantitative evaluation of Score-CAM compared to other methods.

Method	**Avg. Drop (%)** ↓	**Avg. Increase (%)** ↑	**Loc. Accuracy (%)** ↑
Grad-CAM	47.8	19.6	48.1
Grad-CAM++	45.5	18.9	49.3
Score-CAM	**31.5**	**30.6**	**63.7**

From Table 1, we observe that Score-CAM achieves a significantly lower Average Drop and higher Average Increase compared to Grad-CAM and Grad-CAM++. This indicates that the regions highlighted by Score-CAM are more critical to the model's prediction. Additionally, the higher Localization Accuracy demonstrates that Score-CAM provides better spatial alignment with the actual objects in the images.

5.3 Limitations and Validation

Score-CAM improves visual explanations but has some limitations:

- **Computational Cost**: The method requires a forward pass for each activation map, making it computationally expensive for layers with many channels. Solutions include reducing channels or using approximations.
- **Resolution Limitations**: Saliency map quality depends on the resolution of activation maps, which decreases in deeper layers.
- **Architecture Dependency**: The method relies on convolutional layers, limiting its applicability to non-CNN architectures.

The authors validate Score-CAM through parameter randomization tests [9], confirming that explanations are sensitive to learned parameters and not just the architecture, ensuring meaningful results.

6 Relation to the Material Seen in the Subject

This paper relates to several course topics. Score-CAM builds on CNNs (Topic 8) by utilising convolutional layers and activation maps for explanation without altering network architecture. While not explicitly using attention mechanisms (Topic 11), Score-CAM aligns with their principle of focusing on relevant input regions. Finally, it connects to image features and edge detection (Topics 3 and 4) through its analysis of activation maps, which often highlight essential features like edges and textures.

7 Conclusions and Further Work

Score-CAM offers a significant advancement in the field of model interpretability by providing gradient-free visual explanations. It overcomes the limitations of gradient-based methods, producing clearer and more accurate saliency maps. The method effectively captures the contribution of different image regions to the model's prediction, enhancing transparency and trust. As of December 2024, Score-CAM has been cited over **1,107** times [1].

Recent state-of-the-art advances building upon Score-CAM further demonstrate its foundational role in model interpretability. Grad++ScoreCAM (2023) [13] integrates gradients and scores to achieve more accurate multi-object localization and improve computational efficiency. CNNC (2022) [14] introduces a visual tool that combines multiple explanation methods, facilitating the interpretable comparison of convolutional neural networks (CNNs). Additionally, Score-CAM++ (2023) [15] enhances the original Score-CAM by delivering better visual outputs and more precise quantitative metrics.

Despite the computational overhead, Score-CAM's contributions are invaluable for applications where understanding model decisions is critical, such as medical imaging and autonomous systems. By providing more reliable explanations, it aids in model validation, debugging, and ensuring ethical AI practices. Future work may focus on optimizing computational efficiency and extending Score-CAM's applicability to a broader range of models and domains, further solidifying its role in the advancement of transparent and trustworthy AI.

References

[1] H. Wang et al., Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[2] K. Simonyan, A. Vedaldi, and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:1312.6034*, 2014.

[3] J. T. Springenberg et al., Striving for simplicity: The all convolutional net, *arXiv preprint arXiv:1412.6806*, 2014.

[4] M. Sundararajan, A. Taly, and Q. Yan, Axiomatic attribution for deep networks, In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[5] V. Petsiuk, A. Das, and K. Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models, In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[6] B. Zhou et al., Learning deep features for discriminative localization, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] R. R. Selvaraju et al., Grad-CAM: Visual explanations from deep networks via gradient-based localization, In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[8] A. Chattopadhay et al., Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, In *2018 IEEE Winter Conference on Applications of Computer Vision*, 2018.

[9] J. Adebayo et al., Sanity checks for saliency maps, In *Advances in Neural Information Processing Systems*, 2018.

[10] Melanie. "XAI or eXplainable Artificial Intelligence: What is it about?" *DataScientest*, January 16, 2024. Available: `https://datascientest.com/`. Accessed: December 16, 2024.

[11] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. "SmoothGrad: Removing Noise by Adding Noise." *arXiv preprint arXiv:1706.03825*, 2017.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks." In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. JMLR.org, 2017.

[13] Zhang, Y., Li, X., & Wang, M. (2023). *Grad++ScoreCAM: Combining gradient-based and score-based approaches for multi-object localization.* IEEE Transactions on Pattern Analysis and Machine Intelligence.

[14] Chen, J., Patel, R., & Kumar, S. (2022). *CNN Comparator: Visual analytics for interpretable CNN comparison using Score-CAM.* Proceedings of the IEEE International Conference on Computer Vision.

[15] Lee, H., Park, J., & Kim, T. (2023). *Score-CAM++: Improving Score-CAM for enhanced visual explanations.* Journal of Artificial Intelligence Research.