Score-CAM: Score-Weighted Visual Explanations for CNNs

Oriol Miró López-Feliu
Marc Gonzalez Vidal

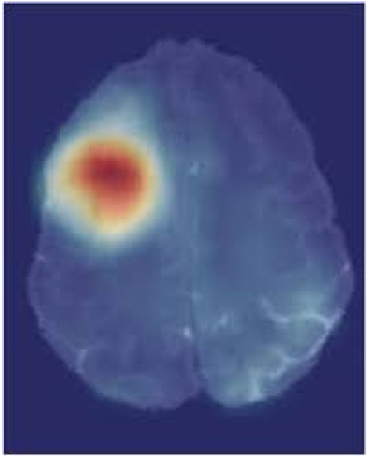
Computational Vision
Master in Artificial Intelligence (MAI)

December 18, 2024

Outline

Motivation

- CNNs achieve high performance but are often "black boxes" [2].
- Understanding model decisions enhances trust and transparency.
- Visual explanations help interpret CNN predictions, essential in critical applications (e.g., medical diagnosis).



[15]

Challenges with Existing Methods

- **Gradient-based Methods** [2, 3, 4]
  - Sensitive to gradient issues (vanishing/exploding).
  - Often produce noisy explanations.
- **Perturbation-based Methods** [5, 6]
  - Computationally expensive (many forward passes).
- **CAM-based Methods** [7, 8, 9]
  - Depend on network architecture.
  - Still rely on gradients.

Need: A method that is architecture-independent and gradient-free.

Gradient Issue

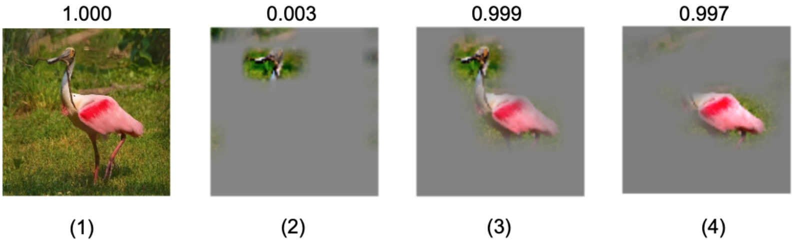


Figure: Higher activation weights do not always result in higher score increases.

- Example:
  - Activation weights: (2) $= 0.035$, (3) $= 0.027$, (4) $= 0.021$.
  - Score increases: (3) $>$ (2), despite its lower weight.
- **Cause:** Gradient saturation.

Objective

Goal: Develop a visual explanation method that:

- Is independent of gradients.
- Works with various CNN architectures.
- Provides accurate and interpretable saliency maps.

Solution: *Score-CAM* [1] (2019)

Score-CAM Overview

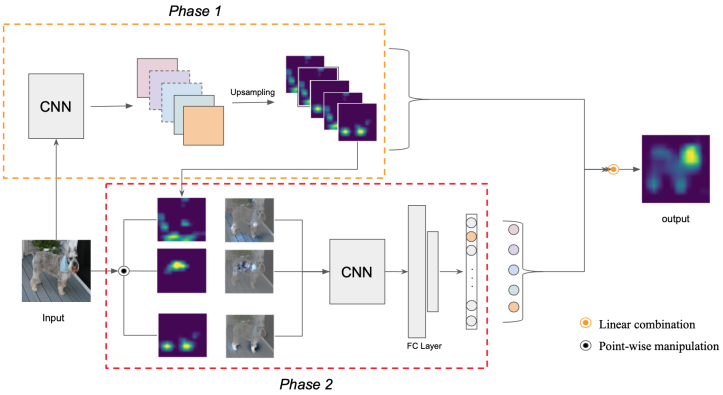


Figure: Score-CAM Pipeline [1]

Key Steps in Score-CAM

1. **Activation Map Extraction:**
   - Extract activation maps A^k from a convolutional layer.
2. **Mask Generation:**
   - Upsample and normalize A^k to create masks M^k.
   - $M^k = \text{Normalize}(\text{Upsample}(A^k))$
3. **Masked Inputs:**
   - Create masked inputs $\tilde{X}^k = X \odot M^k$.

Key Steps in Score-CAM (cont.)

4. **Forward Passes:**
   - Feed $\tilde{X}^k$ into the model to get class scores S_k^c.
5. **Weight Computation:**
   - Compute weights:

$$\alpha_k^c = \frac{\exp(S_k^c)}{\sum_j \exp(S_k^c)}$$

6. **Saliency Map Generation:**
   - Combine activation maps:

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

Qualitative Results

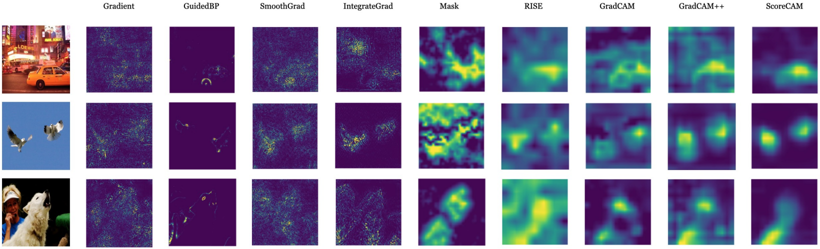


Figure: Comparison of Saliency Maps [1]

Quantitative Evaluation

Evaluation Metrics [9, 6]:

- **Average Drop (%)**: Lower is better.
- **Average Increase (%)**: Higher is better.
- **Localization Accuracy (%)**: Higher is better.

Method	**Avg. Drop** $\downarrow$	**Avg. Increase** $\uparrow$	**Loc. Acc.** $\uparrow$
Grad-CAM [8]	47.8%	19.6%	48.1%
Grad-CAM++ [9]	45.5%	18.9%	49.3%
Score-CAM [1]	**31.5%**	**30.6%**	**63.7%**

Limitations and Connection to Course Topics

Limitations

- **High Compute Cost:** Requires a forward pass per activation map.
- **Low Resolution:** Deep-layer maps are less detailed.
- **CNN-specific:** Limited applicability to non-convolutional models.

Connection to Course Topics:

- **Convolutional Neural Networks:**
  - Utilizes activation maps to interpret features.
- **Attention Mechanisms:**
  - Similar to focusing on important input regions.
- **Object Detection**
  - Proposes explanation for detected objects.

Conclusion

- Score-CAM provides effective, gradient-free visual explanations.
- Overcomes limitations of previous methods.
- Enhances model interpretability and reliability.

Current Relevancy:

- Dec. 2024: Cited by **1107**.
- SOTA Advances:
  - **Grad++ScoreCAM (2023) [12]:** Integrates gradients and scores for accurate multi-object localization and efficiency.
  - **CNNC (2022) [13]:** Visual tool combining explanation methods for interpretable CNN comparison.
  - **Score-CAM++ (2023) [14]:** Enhances Score-CAM with better visual outputs and precise quantitative metrics.

Thank You!

Questions?

References I

H. Wang, Z. Wang, P. Mardziel, et al.,
Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks,
CVPR Workshops, 2020. Available:
https://doi.org/10.1109/CVPRW50498.2020.00359

K. Simonyan, A. Vedaldi, A. Zisserman,
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps,
arXiv preprint arXiv:1312.6034, 2014.

J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller,
Striving for Simplicity: The All Convolutional Net,
arXiv preprint arXiv:1412.6806, 2015.

M. Sundararajan, A. Taly, Q. Yan,
Axiomatic Attribution for Deep Networks,
Proceedings of the 34th International Conference on Machine Learning, 2017.

References II

M. T. Ribeiro, S. Singh, C. Guestrin,
"Why Should I Trust You?": Explaining the Predictions of Any Classifier,
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

V. Petsiuk, A. Das, K. Saenko,
RISE: Randomized Input Sampling for Explanation of Black-box Models,
BMVC, 2018.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba,
Learning Deep Features for Discriminative Localization,
CVPR, 2016.

R. R. Selvaraju, M. Cogswell, A. Das, et al.,
Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,
ICCV, 2017.

References III

A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian,
Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks,
WACV, 2018.

J. Adebayo, J. Gilmer, I. Goodfellow, et al.,
Sanity Checks for Saliency Maps,
NeurIPS, 2018.

H. Wang, Z. Wang, P. Mardziel, et al.,
Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks,
CVPR Workshops, 2020. Available:
`https://doi.org/10.1109/CVPRW50498.2020.00359`

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.

References IV

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.

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

S. Dasanayaka, V. Shantha, S. Silva, D. Meedeniya, and T. Ambegoda, "Interpretable machine learning for brain tumour analysis using MRI and whole slide images," *Software Impacts*, vol. 13, p. 100340, 2022. doi: 10.1016/j.simpa.2022.100340.