Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37150

Visual Explanation Methods for CNNs: Grad-CAM, Score-CAM and LIME

Shiming Song

School of Mathematics, Sichuan University, Chengdu, China
S2236211500@outlook.com

Abstract. Convolutional Neural Networks (CNN) are not widely used in high-risk decision-making because they have black-box nature. This paper investigates three visualization methods, Gradient-weighted class activation mapping (Grad-CAM), Score-weighted visual explanations for convolutional neural networks (Score-CAM), and Local interpretable model-agnostic explanations (LIME), across classification, detection, and medical imaging tasks. Theoretically, Grad-CAM relies on gradient backpropagation to weight feature maps. This offers computational efficiency but suffers from gradient instability and coarse spatial resolution. Score-CAM replaces gradients with forward pass confidence scores. This enhances robustness and spatial alignment at the cost of linearly increasing inference overhead. LIME is a model agnostic surrogate. It provides flexible local explanations but is inherently sensitive to perturbation sampling and hyperparameter choices, leading to poor reproducibility. Grad-CAM performs best in dermatologist agreement and localization accuracy. Score-CAM show superior spatial alignment in pneumonia detection with gradient free robustness. LIME shows flexibility but suffers from instability and hyperparameter sensitivity. The differences are obvious: Grad-CAM is good at interpretability, Score-CAM is good at robustness with computational cost, and LIME is good at model agnostic applicability. Based on these differences, these three visualization tools can be selected for real-world tasks.

Keywords: Grad-CAM, Score-CAM, LIME, visualization

1. Introduction

CNNs have become one of the most widely adopted and best-performing models in the field of deep learning, achieving remarkable success in numerous tasks such as computer vision and natural language processing [1]. Particularly in image recognition, CNNs are highly favored for their powerful feature extraction capabilities. From the breakthrough results in the ImageNet large scale visual recognition challenge to the continuous emergence of advanced architectures, their recognition accuracy has approached or even surpassed human-level performance on specific benchmarks [2]. Although CNNs have superior performance, their decision-making process is difficult to interpret intuitively. This is because their internal structure contains an enormous number of parameters and undergoes complex nonlinear transformations. These factors create a typical black-box problem. In high-risk decision-making scenarios such as healthcare, finance, and law, the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37150

reasons behind model decisions are often as important as the decisions themselves. Non-transparent CNNs are not sufficiently trustworthy for practical deployment.

To probe the decision logic, researchers have developed various interpretability tools. Grad-CAM is proposed by Selvaraju et al. [3]. It uses gradient information to visually explain the decision basis of CNNs and generates heatmaps. It is currently one of the most widely cited visual explanation methods. Score-CAM proposed by Wang et al. [4] eliminates the dependence on gradients by using forward-pass confidence scores to weight activation maps, achieving improvements in both visual quality and fairness. LIME proposed by Ribeiro et al. [5] is a model-agnostic surrogate method. It explains individual predictions of any classifier by fitting an interpretable local model around the instance to be explained. These tools provide strong support for understanding the internal decisions of CNNs from different perspectives.

They each have distinct characteristics in terms of explanation fidelity, stability, computational efficiency, and applicable scenarios. In practice, there is still no systematic consensus on which method is superior. Differences in their performance regarding visualization granularity, sensitivity to adversarial perturbations, and class discriminability directly affect users' trust in the explanation results. The motivation of this paper is to systematically examine the strengths and weaknesses of these three methods, aiming to provide a reference for researchers and engineer in method selection. The structure of this paper is arranged as follows. Section 2 shows the algorithmic principles, mathematical foundations, and implementation details of Grad-CAM, Score-CAM, and LIME. Section 3 demonstrates their application performance on different typical tasks such as image classification, object detection, and weakly supervised localization. This includes heatmap visualization quality, localization accuracy, and quantitative evaluation metrics. Section 4 provides useful advice for the selection of explainability methods and concludes the paper.

2. Methods

This section will present the principles of three visualization methods, along with their relevant formulas.

2.1. Grad-CAM

Grad-CAM is an explanation method that generates visual heatmaps based on gradient information. Its core idea is as follows: for a target class c, compute the gradient of each channel k's feature map A^k in the last convolutional layer with respect to the class score y^c, and then apply global average pooling to obtain the channel weight $\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$, where Z is the total number of pixels in the feature map. This weight reflects the contribution of that channel's feature map to the prediction of the target class. Subsequently, the feature maps of all channels are linearly weighted by their weights. The result is passed through a ReLU activation to generate the final heatmap $L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$. The ReLU operation ensures that only positively contributing regions are highlighted. Grad-CAM is broadly applicable, requires no modification to the model architecture, and can be adapted to various CNN architectures. It produces discriminative heatmaps for different classes, facilitating model debugging and failure analysis. But its heatmap spatial resolution is limited by the size of the last convolutional layer. This causes relatively coarse localization. The weight estimation heavily depends on gradient quality; vanishing or exploding gradients may introduce bias. The weight estimation heavily depends on gradient quality. Vanishing

DOI: 10.54254/2755-2721/2026.37150

or exploding gradients may introduce bias. This method mainly captures spatial features. It is hard to reflect high-level semantic information.

2.2. Score-CAM

Score-CAM aims to eliminate the dependency on gradients and instead uses forward-propagation confidence scores to compute channel importance. Specifically, each channel feature ma A^k from the last convolutional layer is upsampled to the input image size $(H\times W)$ and multiplied element-wise with the original input I, serving as an attention mask that is fed into the network to obtain the softmax or logit score $f_c(I\odot A^k)$ for target class c. The difference between this score and the score obtained from a baseline input I_b (e.g., all-zero or blurred image) gives the channel weight $\alpha_k=f_c(I\odot A^k)-f_c(I_b)$. The final heatmap is again obtained by weighted summation followed by ReLU: $(L^c_{\text{Score-CAM}}=\text{ReLU}\left(\sum_k \alpha_k\cdot A^k\right))$. Since it does not rely on backpropagation, Score-CAM avoids gradient instability, yields more robust results, and exhibits strong resistance to adversarial examples and noise. In certain scenarios, its localization accuracy is also finer. However, its main drawbacks are the high computational cost, each channel requires an independent forward pass, and the overhead grows linearly with the number of channels, and the sensitivity of the weights to the choice of baseline input (zero, blurred, or random). The final heatmap is still constrained by the spatial size of the last convolutional feature maps.

2.3. LIME

LIME adopts a completely gradient-free surrogate-model approach and is applicable to any black-box model. For a given sample x to be explained, LIME generates numerous perturbed samples in its vicinity (e.g., zeroing out super-pixels in images or deleting words in text) and obtains the black-box model f's predictions on these perturbations. Then, using the similarity π_x between each perturbed sample and the original sample as weights, it fits a simple and interpretable surrogate model g (from a family of models G). The optimization objective is $(\xi(x)=\underset{g\in G}{\text{argmin}}\,[L(f,g,\pi_x)+\Omega(g)])$, where (L) is the fidelity loss between the surrogate model's predictions and the black-box outputs, and $\Omega(g)$ is a complexity penalty on the surrogate model (e.g., L1 regularization to encourage sparsity). After fitting, the coefficients of the surrogate model directly reflect the contribution of each feature (e.g., image super-pixels or text words) to the prediction, thereby generating importance rankings or visual heatmaps. The greatest advantage of LIME is its model-agnostic nature, it can explain any classifier, including CNNs, RNNs, tree-based models, and its intuitive output, high flexibility, ability to produce instance-specific explanations, and compatibility with various data modalities. However, it suffers from poor stability; the randomness in perturbation sampling and similarity definition often leads to fluctuating explanations. Its computational efficiency is low due to the large number of perturbations and prediction calls, making it difficult to meet real-time requirements. Hyperparameters (perturbation range, number of samples, similarity kernel, etc.) significantly affect the explanation quality. LIME is only valid in local linear regions and cannot reflect the model's global decision logic.

DOI: 10.54254/2755-2721/2026.37150

3. Application

This section will demonstrate the performance of the three visualization methods in different tasks and provide a comparative analysis.

3.1. Performance in classification tasks

Grad-CAM demonstrates the most robust performance in image classification tasks. In a comparative study on skin lesion (melanoma) classification, Grad-CAM achieved a total agreement rate of 40% with dermatologist judgments and a no-agreement rate of only 6%, significantly better than other methods [6]. In the overall analysis, Grad-CAM's high overall partial agreement rate ranks first among the five explanation techniques evaluated. In acute lymphoblastic leukemia (ALL) diagnosis, Grad-CAM precisely highlighted non-functional leukocyte regions. It closely matched ground truth annotations [7]. In chest X-ray pneumonia detection, Grad-CAM consistently highlighted pathology-relevant lung regions, providing radiologists with interpretable and trustworthy predictive evidence [8, 9]. User studies indicate that trust and satisfaction scores did not differ significantly across methods, Grad-CAM performed best in helping users predict model performance and correct classification errors [10].

Performance of Score-CAM in classification tasks is mixed. In skin lesion classification, Score-CAM had the poorest quantitative evaluation results but was consistently ranked by dermatologists as the third most preferred method, behind Grad-CAM and Grad-CAM++ [9]. In pigmentation abnormality assessments, Score-CAM showed low rate of high agreement and relatively high rate of no agreement rates [6]. In the overall analysis, Score-CAM's high overall partial agreement rate ranks last [6]. However, in chest X-ray pneumonia detection, DenseNet121 combined with Score-CAM achieved the highest mean IoU, indicating the best spatial alignment with actual lesion regions [9].

LIME shows significant performance variability in classification tasks. In ALL diagnosis, LIME struggled to maintain consistency with ground truth annotations, both omitting key features and including irrelevant regions [7]. In skin lesion classification, LIME ranked second quantitatively but was ranked second-to-last by dermatologists [6]. In pigmentation abnormality and boundary assessments, LIME performed poorly, with agreement rates that are extremely low. In user studies, LIME's stability issues were prominent, the same super-pixel could be marked positive in one run and negative in another for the same image [10].

3.2. Performance in detection and segmentation tasks

Grad-CAM has good applicability in object detection tasks. In YOLOv5-based anterior segment eye disease diagnosis, Grad-CAM++ analysis revealed a layer-wise attention mechanism. It progresses from shallower to deeper layers [11]. Grad-CAM enables unsupervised segmentation. It can visualize the association of each pixel with specific objects without relying on pixel-level labels. In brain tumor detection, Grad-CAM is used to generate class-discriminative localization maps, highlighting regions most influential for classification.

Score-CAM primarily visualizes model attention via heatmaps in segmentation tasks. But this incurs high computational costs [9]. In diabetic retinopathy analysis, Score-CAM resulted in excessive computational overhead [8].

LIME can provide local explanations in segmentation tasks by using super-pixel segmentation, identifying key regions most influential for predictions [10]. But LIME's explanations are based on

DOI: 10.54254/2755-2721/2026.37150

image segmentation rather than leveraging the network's internal knowledge. Its most significant regions do not match those identified by Grad-CAM methods [12].

3.3. Performance on natural images

A user study has compared Grad-CAM, LIME, and GBP. It found that while trust and satisfaction scores did not differ significantly when using different methods, Grad-CAM better helped users anticipate model performance and correct classification errors [12]. Grad-CAM localized regions of interest more precisely than LIME. In a clinical evaluation of chest radiographs, most participants preferred Grad-CAM heatmaps, while only a few preferred LIME visualizations [8]. Grad-CAM outperformed LIME in coherence and trustworthiness. LIME's explanation quality on natural images is significantly affected by super-pixel segmentation and random perturbations, resulting in suboptimal stability [10]. But some clinicians raised concerns regarding Grad-CAM's practical usability.

3.4. Performance on medical images

A systematic evaluation across four modalities (CT, chest X-ray, MRI, and ultrasound) found Grad-CAM performed best on CT, CXR, and MRI, achieving the lowest average drop percentage (AD%) and highest confidence increase percentage (IC%) [11]. In skin lesion diagnosis, Grad-CAM most accurately delineated lesion boundaries, showing the best match with clinically relevant regions [6].

Score-CAM, combined with DenseNet121, achieved the highest mean IoU in lung X-ray pneumonia detection, indicating superior spatial localization accuracy [9]. In musculoskeletal disease assessments, Score-CAM performed well on ADCC metrics with DenseNet-169 and Inception-v3 [13]. However, Score-CAM also obtained the worst localization score in some evaluations [13].

A systematic evaluation focus on four modalities (CT, chest X-ray, MRI, and ultrasound) found Grad-CAM performed best on CT, CXR, and MRI. Its average drop percentage (AD%) is the lowest and its confidence increase percentage (IC%) is the highest. In skin lesion diagnosis, Grad-CAM most accurately delineated lesion boundaries, showing the best match with clinically relevant regions [6].

When combined with DenseNet121, Score-CAM's mean IoU is the highest in lung X-ray pneumonia detection, indicating superior spatial localization accuracy [9]. In musculoskeletal disease assessments, Score-CAM performed well on ADCC metrics with DenseNet-169 and Inception-v3 [13]. But Score-CAM also obtained the worst localization score in some evaluations.

LIME performs best on ultrasound images, maintaining the highest level of confidence of the original model, and is most preferred by clinicians [9]. In skin lesion boundary assessments, LIME generated difficult-to-interpret geographic regions [6]. Grad-CAM ranked first in dermatologist preference, followed by Grad-CAM++ in second, Score-CAM in third, with LIME ranking lower.

3.5. General observations

The three methods each have strengths and weaknesses across the five dimensions of understandability, usefulness, trustworthiness, informativeness, and satisfaction. Grad-CAM shows clear advantages in localization accuracy and expert agreement [6, 9]. Score-CAM, sometimes lagging in quantitative evaluations, receives recognition from experts for its visual quality [9, 13]. LIME offers the flexibility of being model-agnostic, but exhibits notable shortcomings in stability

DOI: 10.54254/2755-2721/2026.37150

and fidelity. Method selection should consider the specific task type, computational resources, modality characteristics, and user requirements [12].

4. Conclusion

This study investigates three mainstream CNN visualization methods, including Grad-CAM, Score-CAM, and LIME, across tasks including classification, detection, and medical imaging. The results reveal their distinct complementarities and limitations. Grad-CAM proves most robust in classification tasks, achieving a high overall partial agreement rate with dermatologists' judgments, and yields the lowest average drop percentage and highest confidence increase on CT, X ray, and MRI data. Its heatmaps rank first in both localization accuracy and expert preference, making it the preferred tool for interpreting classification decisions. Although Score-CAM ranks lower in quantitative assessment of skin lesions, its gradient-free design endows it with strong robustness against adversarial examples; moreover, it achieves the highest mean IoU in pneumonia detection, indicating that its spatial localization capability can surpass that of Grad-CAM in specific scenarios. However, its linearly increasing inference cost severely limits real-time deployment in large-channel networks. As a model-agnostic method, LIME performs best on ultrasound images and gains clinical preference, yet its instability is a notable weakness. Repeated explanations for the same super pixel can yield contradictory results, and it is highly sensitive to hyperparameters, with issues in both omitting critical features and including irrelevant regions, which undermines its trustworthiness in high-stakes decision-making. These findings provide clear guidance for researchers and practitioners. If localization accuracy and expert interpretability are prioritized, Grad-CAM serves as a robust baseline. If gradient instability or adversarial environments are a concern, Score-CAM can be adopted with a trade-off in cost. For non-CNN black-box models where real-time constraints are less stringent, LIME can serve as a flexible supplement. Future directions urgently needed, including hybrid explanation frameworks that integrate the strengths of multiple methods, adaptive sampling strategies to improve LIME stability, efficient visualization solutions for 3D medical images, and the establishment of unified quantitative evaluation criteria aligned with human cognition.

References

[1] Alzubaidi, L., Zhang, J., Humaidi, A. J., et al. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*(1), 1–74.

[2] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986.

[3] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

[4] Wang, H., Wang, Z., Du, M., et al. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 24–25.

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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*, 1135–1144.

[6] Giavina-Bianchi, M., Vitor, W. G., Fornasiero de Paiva, V., Okita, A. L., Sousa, R. M., & Machado, B. (2023). Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification. *Frontiers in Medicine, 10*, 1241484.

[7] Muhammad, D., Salman, M., Keles, A. et al. (2025). ALL diagnosis: Can efficiency and transparency coexist? An explainble deep learning approach. *Scitific Reports, 15*, 12812.

DOI: 10.54254/2755-2721/2026.37150

[8] Ihongbe, E. I., Fouad, S., Mahmoud, T. F., Rajasekaran, A., & Bhatia, B. (2024). Evaluating explainable artificial intelligence (XAI) techniques in chest radiology imaging through a human-centered lens. *PLOS ONE, 19*(10), e0308758.

[9] Mahtabi, B., & Nasr-Esfahani, E. (2026). Parameter-efficient deep learning for pneumonia detection on chest X-rays: A comparative evaluation of explainable AI methods. *medRxiv*. https://doi.org/10.64898/2026.07.14.26358065

[10] Bora, R. P., Terhörst, P., Veldhuis, R., Ramachandra, R., & Raja, K. (2024). SLICE: Stabilized LIME for consistent explanations for image classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10988–10996.

[11] Senaratne, S., Senaweera, O., & Dharmarathne, H. S. A. G. (2025). XAI techniques used in medical image classification: A modality-specific analysis. *2025 7th International Conference on Advancements in Computing (ICAC)*, 1–6.

[12] Kazmierczak, R., Azzolin, S., Berthier, E., Hedström, A., Delhomme, P., Frehse, G., Mancini, M., & Franchi, G. (2025). Benchmarking XAI explanations with human-aligned evaluations. *arXiv*, arXiv:2411.02470. https://doi.org/10.1609/aaai.v40i44.41082

[13] Thota, G., Nagaraju, K., & Korra, S. B. (2026). Quantitative assessment of class activation maps: An empirical study on musculoskeletal disorders. *Communications in Computer and Information Science, 2473*, 338–352.