Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.041419
*Editorial*

# Grad-CAM: Understanding AI Models

**Shuihua Wang**[1,2] **and Yudong Zhang**[2,*]

[1]School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, UK
[2]Department of Information Systems, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
*Corresponding Author: Yudong Zhang. Email: yudongzhang@ieee.org
Received: 21 April 2023; Accepted: 17 May 2023; Published: 30 August 2023

**Keywords:** Artificial intelligence; Grad-CAM; deep learning; convolutional neural networks; classification; location; explainable

Artificial intelligence (AI) [1,2] allows computers to think and behave like humans, so it is now becoming more and more influential in almost every field [3]. Hence, users in businesses, industries, hospitals [4], etc., need to understand how these AI models work [5] and the potential impact of using them.

Visualization [6,7] is an important tool for understanding AI models. The feature maps of various convolutional neural networks (CNNs) [8] can be easily extracted from convolution layers and then be visualized [9]. However, classical CNNs [10] are composed of huge numbers of various layers [11], producing an exceedingly large number of these features [12] to be visualized because of the variation of the combination of different types of layers [13].

Those large number of features benefit the effectiveness of CNNs in terms of prediction accuracy [14]. Still, they cause confusion in explaining the AI models' mechanisms and in understanding AI models [15]. This impairs the widespread applications of AI models in fields where black-box AI models are not welcome, such as medical and biological fields.

Several methods have been developed to offer possible explanations for CNNs. The most promising type of method attempts to deliver some visual heatmaps [16] with different color maps within the input image regions containing essential features [17] that the CNN managed in prediction. Some existing methods include saliency mapping [18], class-activation mapping (CAM) [19], gradient-weighted CAM (Grad-CAM) [20], Grad-CAM++ [21], etc.

Among different methods, Grad-CAM exhibits the best performance in terms of localization [22], which is an expected characteristic of heatmaps. A well-localized heatmap can display the borders of the regions encompassing the features that contributed the most toward the classification outcomes [23] and may offer better discrimination [24] and conceivable explanations for the choices crafted by the cognate CNNs [25].

Grad-CAM is helpful in three ways: interpretability, debugging, and trustworthiness. First, AI models are often regarded as black boxes since it is hard to understand how they arrive at their decisions. Grad-CAM can provide insights into the decision-making process [26] of AI models by highlighting the regions of an image that are most important in making a particular prediction [27].

Second, if an AI model does not perform well, Grad-CAM can help identify areas of the image with which the model is struggling. This information can be used to debug the model [28]. Finally, in critical applications, such as medical diagnosis [29], it is essential to know why a particular decision was made. Grad-CAM can explain the model's decision, which can help build trust in the model's deployment [30].

On 16/Feb/2023, Tech Science Press invited Dr. Ramprasaath R. Selvaraju, who was the first author of the Grad-CAM [31] and is now a Senior Machine Learning Scientist at Salesforce Research, to give a talk entitled "Empowering Human Decision-Making in AI Models: The Path to Trust and Transparency" to more than 300 attendees from all over the world.

In this talk, Dr. Ramprasaath presented his recent work toward making deep networks more interpretable, trustworthy, and unbiased. He discussed algorithms that provide explanations for the decisions made by deep networks, which will help: (1) understand why the model made the decisions it did, (2) correct unwanted biases learned by AI models, and (3) encourage human-like reasoning in AI. This talk provided a comprehensive overview of his research in Explainable AI (XAI) and how it could improve the transparency and accountability of deep networks, making them more trustworthy and usable in real-world applications.

After the talk, Nanjing University of Information Science and Technology reported the seminar on its official webpage. It said, "The teachers and students of the college responded strongly to Ramprasaath's lecture, saying that they had a more comprehensive understanding of the field of AI and benefited from the talk. They would take this lecture as an opportunity to further extensively study related fields of knowledge to gain a thorough understanding and mastery."

In the future, there will be new potential applications of Grad-CAM: (i) Object localization. The Grad-CAM can localize objects [32] by highlighting the regions of the objects that contribute to the predictions of objects' presence. (ii) Image classification. Grad-CAM is able to understand the features [33] used by AI models to classify images [34]. The developers can use those features to refine the architecture of AI models. (iii) Autonomous driving. Grad-CAM can help autonomous driving systems [35] better understand the environment, thus making better decisions about navigation. (iv) Video analysis. Grad-CAM can highlight the keyframes [36], which help researchers understand how the AI models analyze video data and improve the performance of AI models.

**Acknowledgement:** The authors thank Lei Zhou for organizing the webinar event "Empowering Human Decision Making in AI Models: The Path to Trust and Transparency".

**Funding Statement:** This paper is partially supported by: Sino-UK Education Fund (OP202006); Royal Society (RP202G0230); MRC (MC_PC_17171); BHF (AA/18/3/34220); Hope Foundation for Cancer Research (RM60G0680); GCRF (P202PF11); BBSRC (RM32G0178B8); Sino-UK Industrial Fund (RP202G0289); Data Science Enhancement Fund (P202RE237); LIAS (P202ED10 & P202RE969); Fight for Sight (24NN201).

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References


[1] M. Sako, "Technology strategy and management contracting for artificial intelligence considering the promises and perils of contracting for the use of artificial intelligence tools and data," *Communications of the ACM*, vol. 66, no. 4, pp. 20–23, 2023.

[2] M. Soliman, T. Fatnassi, I. Elgammal and R. Figueiredo, "Exploring the major trends and emerging themes of artificial intelligence in the scientific leading journals amidst the COVID-19 era," *Big Data and Cognitive Computing*, vol. 7, no. 1, 12, 2023.

[3] D. Monroe, "Artificial intelligence for materials discovery," *Communications of the ACM*, vol. 66, no. 4, pp. 9–10, 2023.

[4] B. P. Cabral, L. A. M. Braga, S. Syed-Abdul and F. B. Mota, "Future of artificial intelligence applications in cancer care: A global cross-sectional survey of researchers," *Current Oncology*, vol. 30, no. 3, pp. 3432–3446, 2023.

[5] D. V. Meenakshi-Siddharthan, C. Livia, T. E. Peterson, P. Stalboerger, Z. I. Attia *et al.,* "Artificial intelligence-derived electrocardiogram assessment of cardiac age and molecular markers of senescence in heart failure," *Mayo Clinic Proceedings*, vol. 98, no. 3, pp. 372–385, 2023.

[6] D. O. Silva, P. N. de Souza, M. L. D. Sousa, C. C. A. Morais, J. C. Ferreira *et al.,* "Impact on the ability of healthcare professionals to correctly identify patient-ventilator asynchronies of the simultaneous visualization of estimated muscle pressure curves on the ventilator display: A randomized study ($p_{\text{mus}}$ study)," *Critical Care*, vol. 27, no. 1, 128, 2023.

[7] R. Duan, J. Y. Tong, A. J. Sutton, D. A. Asch, H. T. Chu *et al.,* "Origami plot: A novel multivariate data visualization tool that improves radar chart," *Journal of Clinical Epidemiology*, vol. 156, no. 6, pp. 85–94, 2023.

[8] A. Amziane, O. Losson, B. Mathon and L. Macaire, "MSfA-Net: A convolutional neural network based on multispectral filter arrays for texture feature extraction," *Pattern Recognition Letters*, vol. 168, no. 5, pp. 93–99, 2023.

[9] A. S. A. Al-Ghamdi and M. Ragab, "Tunicate swarm algorithm with deep convolutional neural network-driven colorectal cancer classification from histopathological imaging data," *Electronic Research Archive*, vol. 31, no. 5, pp. 2793–2812, 2023.

[10] M. Alhajlah, "A novel efficient patient monitoring fer system using optimal dl-features," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 6161–6175, 2023.

[11] S. Bourbia, A. Karine, A. Chetouani, M. El Hassouni and M. Jridi, "No-reference 3D point cloud quality assessment using multi-view projection and deep convolutional neural network," *IEEE Access*, vol. 11, pp. 26759–26772, 2023.

[12] A. Bhasin and A. Mistry, "Convolutional neural networks for problems in transport phenomena: A theoretical minimum," *Journal of Flow Visualization and Image Processing*, vol. 30, no. 3, pp. 1–38, 2023.

[13] A. Salehi and M. Balasubramanian, "DDCNet: Deep dilated convolutional neural network for dense prediction," *Neurocomputing*, vol. 523, no. 4, pp. 116–129, 2023.

[14] P. T. Lee, A. Tahmasebi, J. K. Dave, M. R. Parekh, M. Kumaran *et al.,* "Comparison of gray-scale inversion to improve detection of pulmonary nodules on chest x-rays between radiologists and a deep convolutional neural network," *Current Problems in Diagnostic Radiology*, vol. 52, no. 3, pp. 180–186, 2023.

[15] L. K. Pour and A. Farrokhi, "Language recognition by convolutional neural networks," *Scientia Iranica*, vol. 30, no. 1, pp. 116–123, 2023.

[16] I. Kourbane and Y. Genc, "Skeleton-aware multi-scale heatmap regression for 2D hand pose estimation," *Informatica*, vol. 45, no. 4, pp. 593–604, 2021.

[17] F. J. M. Shamrat, S. Azam, A. Karim, K. Ahmed, F. M. Bui *et al.,* "High-precision multiclass classification of lung disease through customized MobileNetV2 from chest x-ray images," *Computers in Biology and Medicine*, vol. 155, 106646, 2023.

[18] A. Radman, R. Shah-Hosseini and S. Homayouni, "An unsupervised saliency-guided deep convolutional neural network for accurate burn mapping from sentinel-1 sar data," *Remote Sensing*, vol. 15, no. 5, 1184, 2023.

[19] B. J. Kim, G. Koo, H. Choi and S. W. Kim, "Extending class activation mapping using gaussian receptive field," *Computer Vision and Image Understanding*, vol. 231, 103663, 2023.
[20] A. Abhishek, R. K. Jha, R. Sinha and K. Jha, "Automated detection and classification of leukemia on a subject-independent test dataset using deep transfer learning supported by grad-cam visualization," *Biomedical Signal Processing and Control*, vol. 83, 104722, 2023.
[21] I. D. Apostolopoulos, I. Athanasoula, M. Tzani and P. P. Groumpos, "An explainable deep learning framework for detecting and localising smoke and fire incidents: Evaluation of grad-cam plus plus and lime," *Machine Learning and Knowledge Extraction*, vol. 4, no. 4, pp. 1124–1135, 2022.
[22] J. C. Chien, J. D. Lee, C. S. Hu and C. T. Wu, "The usefulness of gradient-weighted cam in assisting medical diagnoses," *Applied Sciences*, vol. 12, no. 15, 7748, 2022.
[23] J. P. Li, Z. X. Qiu, K. Y. Cao, L. Deng, W. J. Zhang *et al.,* "Predicting muscle invasion in bladder cancer based on MRI: A comparison of radiomics, and single-task and multi-task deep learning," *Computer Methods and Programs in Biomedicine*, vol. 233, 107466, 2023.
[24] B. S. Rajeshwari, M. Patra, A. Sinha, A. Sengupta and N. Ghosh, "Detection of phonocardiogram event patterns in mitral valve prolapse: An automated clinically relevant explainable diagnostic framework," *Ieee Transactions on Instrumentation and Measurement*, vol. 72, 4001709, 2023.
[25] S. H. Wang, K. M. Attique and G. Vishnuvarthanan, "Deep rank-based average pooling network for covid-19 recognition," *Computers, Materials & Continua*, vol. 70, no. 2, pp. 2797–2813, 2022.
[26] J. P. Kucklick and O. Muller, "Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, 6, 2023.
[27] J. W. Baek and K. Y. Y. Chung, "Explainable anomaly detection using vision transformer based SVDD," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 6573–6586, 2023.
[28] J. M. Sun, S. Lapuschkin, W. Samek and A. Binder, "Explain and improve: Lrp-inference fine-tuning for image captioning models," *Information Fusion*, vol. 77, no. 12, pp. 233–246, 2022.
[29] F. Zulfiqar, U. I. Bajwa and Y. Mehmood, "Multi-class classification of brain tumor types from mr images using efficientnets," *Biomedical Signal Processing and Control*, vol. 84, 104777, 2023.
[30] P. Singh and A. Sharma, "Interpretation and classification of arrhythmia using deep convolutional network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2518512, 2022.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh *et al.,* "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.
[32] A. I. Xavier, C. Villavicencio, J. J. Macrohon, J. H. Jeng and J. G. Hsieh, "Object detection via gradient-based mask r-cnn using machine learning algorithms," *Machines*, vol. 10, no. 5, Article ID: 340, 2022.
[33] E. Kim, J. Kim, J. Park, H. Ko and Y. Kyung, "Tinyml-based classification in an ecg monitoring embedded system," *Computers, Materials & Continua*, vol. 75, no. 1, pp. 1751–1764, 2023.
[34] J. M. An, Y. Du, P. Hong, L. Zhang and X. G. Weng, "Insect recognition based on complementary features from multiple views," *Scientific Reports*, vol. 13, no. 1, pp. 2966, 2023.
[35] D. Dworak and J. Baranowski, "Adaptation of Grad-CAM method to neural network architecture for lidar pointcloud object detection," *Energies*, vol. 15, no. 13, 4681, 2022.
[36] S. Pericherla and E. Ilavarasan, "Cyberbullying detection on multi-modal data using pre-trained deep learning architectures," *Ingenieria Solidaria*, vol. 17, no. 3, pp. 1–20, 2021.