Interpreting the Incomprehensible: Benchmarking Visual Explanation Methods for Deep Convolutional Networks

Wei Hong Lim[*1], Marwa M. Eid[2]

[1] Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur 56000, Malaysia
[2] Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 11152, Egypt
Emails limwh@ucsiuniverisity.edu.my; mmm@ieee.org

Abstract

Deep Convolutional Networks (CNNs) have revolutionized various fields, including computer vision, but their decision-making process remains largely opaque. To address this interpretability challenge, numerous visual explanation methods have been proposed. However, a comprehensive evaluation and benchmarking of these methods are essential to understand their strengths, limitations, and comparative performance. In this paper, we present a systematic study that benchmarks and compares various visual explanation techniques for deep CNNs. We propose a standardized evaluation framework consisting of benchmark explain ability methods. Through extensive experiments, we analyze the effectiveness, and interpretability of popular visual explanation methods, including gradient-based methods, activation maximization, and attention mechanisms. Our results reveal nuanced differences between the methods, highlighting their trade-offs and potential applications. We conduce a comprehensive evaluation of visual explanation methods on different deep CNNs, the results demonstrate the ability to achieve informed selection and adoption of appropriate techniques for interpretability in real-world applications.

Keywords: Convolutional Neural Networks (CNNs); Benchmarking, Interpretability; Class Activation Maps (CAM); Deep learning, Image classification; Explainable AI.

1. Introduction

Deep Convolutional Networks (CNNs) have demonstrated remarkable success in various computer vision tasks, such as image classification, object detection, and semantic segmentation. However, one of the major challenges in leveraging the power of CNNs lies in their inherent black box nature. The complex hierarchical layers of convolution and pooling operations make it difficult to understand how these networks arrive at their decisions, hindering their interpretability [1]. This lack of interpretability raises concerns in critical domains where transparency and trust are crucial, such as healthcare, autonomous driving, and security. To address this challenge, researchers have proposed visual explanation methods that aim to shed light on the decision-making process of CNNs by generating human-interpretable explanations [2]. While a plethora of visual explanation techniques have been proposed, there is a need for a systematic evaluation and benchmarking of these methods. Such an evaluation is vital to assess their effectiveness, understand their limitations, and compare their performance across different domains and tasks [3]. Additionally, a standardized evaluation framework can help researchers and practitioners in selecting appropriate visual explanation methods for their specific needs. In this paper, we present a comprehensive study that benchmarks and compares various visual explanation methods for deep CNNs [4].

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

Our objective is to provide a thorough analysis and evaluation of popular visual explanation techniques, including gradient-based methods, activation maximization, and attention mechanisms. We propose a standardized evaluation framework consisting of diverse benchmark datasets, representative of different computer vision tasks and domains. Furthermore, we define evaluation metrics that capture the quality, interpretability, and computational efficiency of the generated explanations [5]. Through extensive experiments and comparisons, we aim to uncover the strengths and weaknesses of each method, enabling researchers and practitioners to make informed decisions when choosing a visual explanation technique for their specific application [6]. By conducting this benchmarking study, we also aim to highlight the advancements made in the field of visual explanation for deep CNNs. We investigate the progress made in terms of interpretability, robustness, and generalizability of the methods. Additionally, we explore the trade-offs between interpretability and performance, providing insights into the practical considerations for deploying visual explanation techniques in real-world scenarios.

2. Background

Visual explanation methods have emerged as a promising approach to address the interpretability challenge of deep Convolutional Neural Networks (CNNs). These methods aim to generate visual representations or heatmaps that highlight the regions of an input image that are most influential in the network's decision-making process. In this section, we provide an overview of some popular visual explanation methods, including Class Activation Maps (CAM), Gradient-weighted Class Activation Mapping (Grad-CAM), and Grad-CAM++ [7]. CAM is a widely adopted visual explanation method that provides localization information by leveraging the global average pooling layer in CNNs. CAM generates a heatmap by computing the weighted sum of the activation maps of the last convolutional layer based on the importance of each feature map in predicting a specific class [8]. The resulting heatmap highlights the discriminative regions in the input image that contribute most to the network's classification decision.

Gradient-weighted Class Activation Mapping (Grad-CAM) extends the idea of CAM by utilizing the gradients flowing into the last convolutional layer. Instead of relying solely on the global average pooling layer, Grad-CAM computes the importance weights of the feature maps based on the gradients of the target class score with respect to the feature maps. This allows Grad-CAM to generate more precise and fine-grained heatmaps that highlight the relevant regions for a particular class. Building upon Grad-CAM, Grad-CAM++ further enhances the localization accuracy by incorporating higher-order gradients. By considering both the positive and negative gradients, Grad-CAM++ provides a more comprehensive understanding of the regions that contribute positively or negatively to the target class. This results in sharper and more informative heatmaps that help in interpreting the network's decision with greater clarity. In addition to CAM, Grad-CAM, and Grad-CAM++, several other visual explanation methods have been proposed in the literature [8-10]. These include Integrated Gradients, which computes the integral of gradients along a straight path between a baseline image and the input image to determine the importance of each pixel. Smooth Grad applies noise to the input image and computes the average gradients across multiple noisy samples, reducing the effect of local perturbations. Guided Backpropagation highlights the input pixels that have the most impact on the target class by computing the gradients propagated back from the output layer [11]. These visual explanation methods provide valuable insights into the decision-making process of deep CNNs by highlighting the regions of importance in the input images. However, they vary in terms of their interpretability, computational efficiency, and robustness to adversarial attacks. In the following sections, we will evaluate and compare these methods, along with others, in order to gain a comprehensive understanding of their strengths and limitations for visual explanations in deep CNNs [12].

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

3. Visual Explanation Methods

The section on Visual Explanation Methods aims to provide an in-depth exploration and evaluation of various techniques that shed light on the decision-making process of CNNs. Visual explanation methods for deep CNNs is taxonomized into distinct categories based on their underlying principles and techniques, which can provide a structured framework for understanding and comparing the various methods, allowing researchers and practitioners to gain insights into their strengths and limitations for interpretability purposes (See Figure 1).

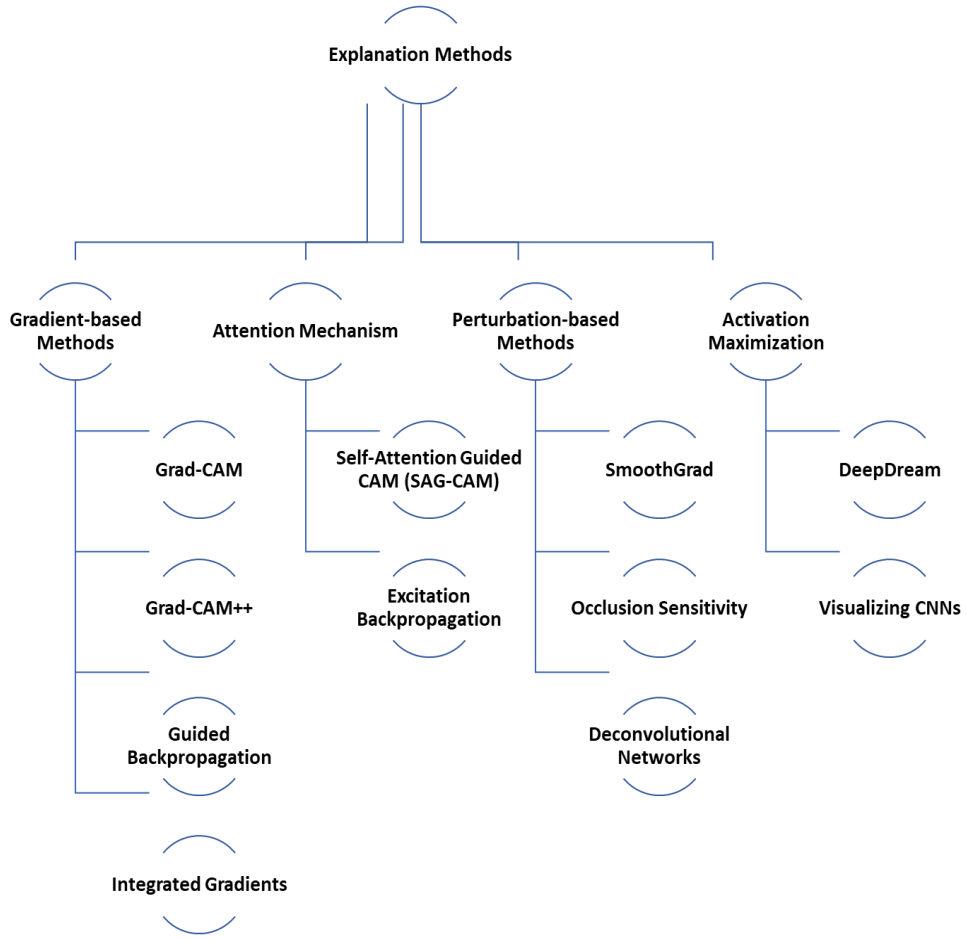

Figure1: Taxonomy of visual explanation mechanisms

Gradient-based methods utilize the gradients of the network's output with respect to the input image to determine the importance of each pixel or region. These methods calculate the gradient information to highlight the areas of the image that have the most influence on the network's decision. Techniques such as Grad-CAM and Grad-CAM++ use these gradients to generate heatmaps that visualize the regions that contribute significantly to the classification decision. Guided Backpropagation takes this approach a step further by backpropagating gradients only through the positive activations, resulting in a more focused visualization (see figure 2 and figure 3). Integrated Gradients compute the integral of gradients along a straight path from a baseline image to the input image, providing a pixel-wise attribution of importance [13-15].

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

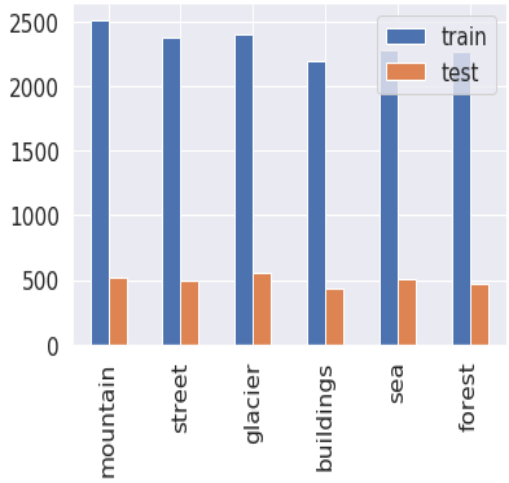

Figure 2: Distribution of train and test samples

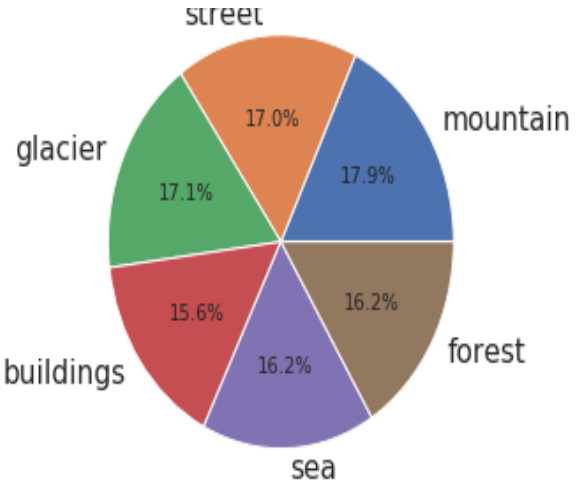

Figure 1: class distribution in intel image classification dataset.

Perturbation-based methods introduce small perturbations or modifications to the input image and observe the resulting changes in the network's output. By analyzing the sensitivity of the network to these perturbations, these methods highlight the regions that are crucial for the network's decision-making process. Techniques like Smooth Grad apply noise to the input image and compute the average gradients across multiple noisy samples, reducing the effect of local perturbations. Occlusion Sensitivity systematically occludes different parts of the image to evaluate their impact on the network's output. Deconvolutional Networks reverse the forward process of convolutional layers to generate visual explanations [10-13].

Activation maximization techniques aim to generate input images that maximize the activation of a specific neuron or target class in the network. By iteratively modifying an input image to amplify the response of the desired feature or class, these methods provide insights into the learned representations of the network. Deep Dream is a prominent example of this approach, producing dream-like images that strongly activate certain features. Visualizing CNNs involves optimizing an input image to maximize the activation of a specific class, enabling the visualization of what the network has learned to associate with that class [7-9].

Attention-based methods focus on identifying the regions in the input image that receive the highest attention or importance during the network's decision-making process. Inspired by human visual attention, these methods often involve the use of attention maps or masks. Self-Attention Guided CAM (SAG-CAM) combines the attention mechanism of self-attention networks with the Grad-CAM framework, resulting in improved localization performance. Excitation Backpropagation utilizes the concept of attention to propagate excitation values backward through the network, highlighting the regions that contribute most to the final decision [11-16].

By considering these categories of visual explanation methods, researchers and practitioners can explore a diverse range of techniques and choose the most suitable approach based on their specific interpretability needs. The taxonomy facilitates a comprehensive understanding of the strengths, limitations, and trade-offs associated with different methods, ultimately advancing the field of visual explanation for deep CNNs.

In Figure 4, we illustrate the learning curves of our CNN. The learning curves provide valuable insights into the training process and model performance over epochs. It clearly depicts the changes in both training and validation loss as training progresses. The training loss represents the error or discrepancy between the predicted outputs and the ground truth labels for the training dataset. On the other hand, the validation loss measures the model's performance on a separate validation set that was

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

not used for training. This allows us to assess the convergence and generalization capabilities of our CNN model and observe that our model can effectively learn from the training data and avoid overfitting.

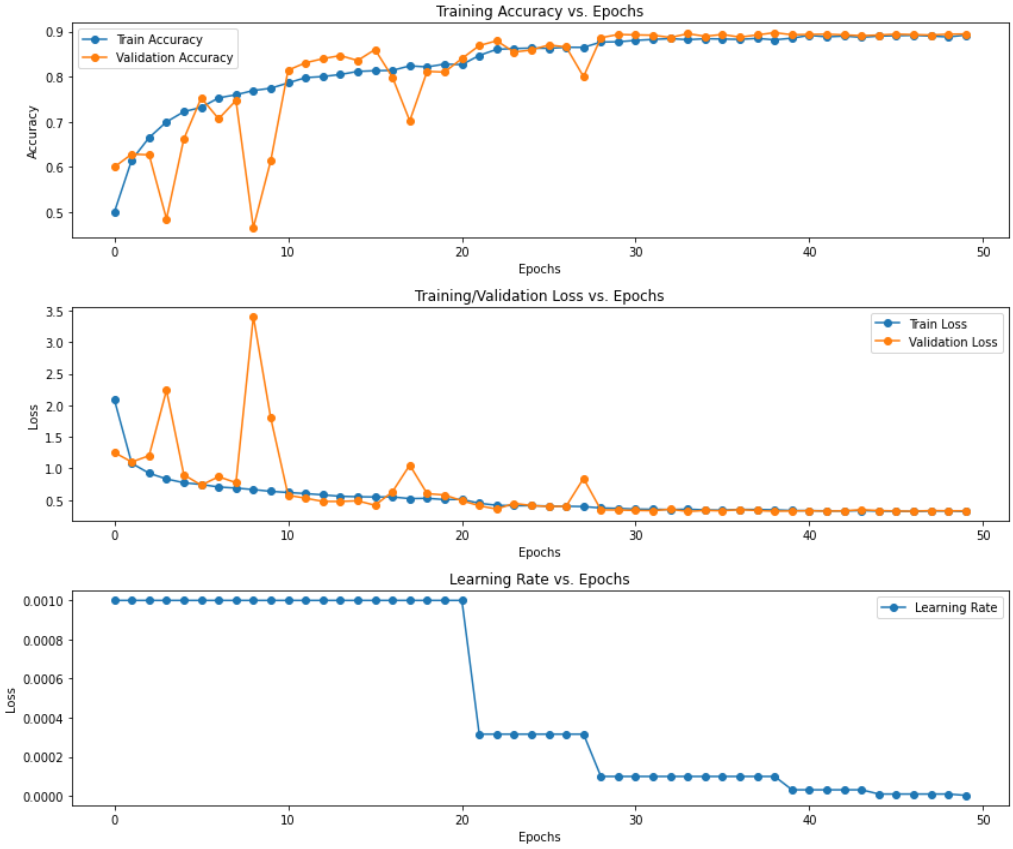

Figure 4: Visualization of learning behavior in our convolutional model

4. Results and Discussion

In our work, we conduct experiments on the Intel Image Classification dataset, which comprises approximately 25,000 images of size 150x150. The dataset encompasses images distributed across six distinct categories: buildings, forest, glacier, mountain, sea, and street. These categories are represented by numerical labels ranging from 0 to 5. The Train subset contains approximately 14,000 images, the Test subset comprises around 3,000 images, and the Prediction subset consists of approximately 7,000 images. The dataset was originally released by Intel and made available on the website datahack.analyticsvidhya.com, serving as the foundation for an Image Classification Challenge hosted by Intel. Figure 2 shows the proportion of training and test samples on each of the above classes. Figure 3 pie chart to indicate the class distribution in our dataset. Figure 4 display the learning curves of our model to ensure that it is correctly trained without any overfitting.

In our experiment, we visualize the explanation heatmaps and model predictions on different classes, as demonstrated in Figure 5 and Figure 6. These visualizations offer valuable insights into how the visual explanation methods perform and provide interpretability for the CNN. Figure 5 showcases the explanation heatmaps generated by various visual explanation methods for building and forest classes. Each heatmap highlights the regions of the input image that contribute most significantly to the model's decision for a specific class. In Figure 6, we present the explanation of model predictions on different mountain and sea classes alongside the corresponding input images. By comparing the predicted classes with the ground truth labels, we can determine the model's ability to correctly classify diverse images from the dataset. The combination of explanation heatmaps and model predictions in Figure 5 and Figure 6 provides a comprehensive analysis of the visual explanation

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

methods and the performance of the CNN model. These visualizations aid in interpreting the decision-making process of the model and provide a better understanding of how CNN perceives and categorizes different classes.

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

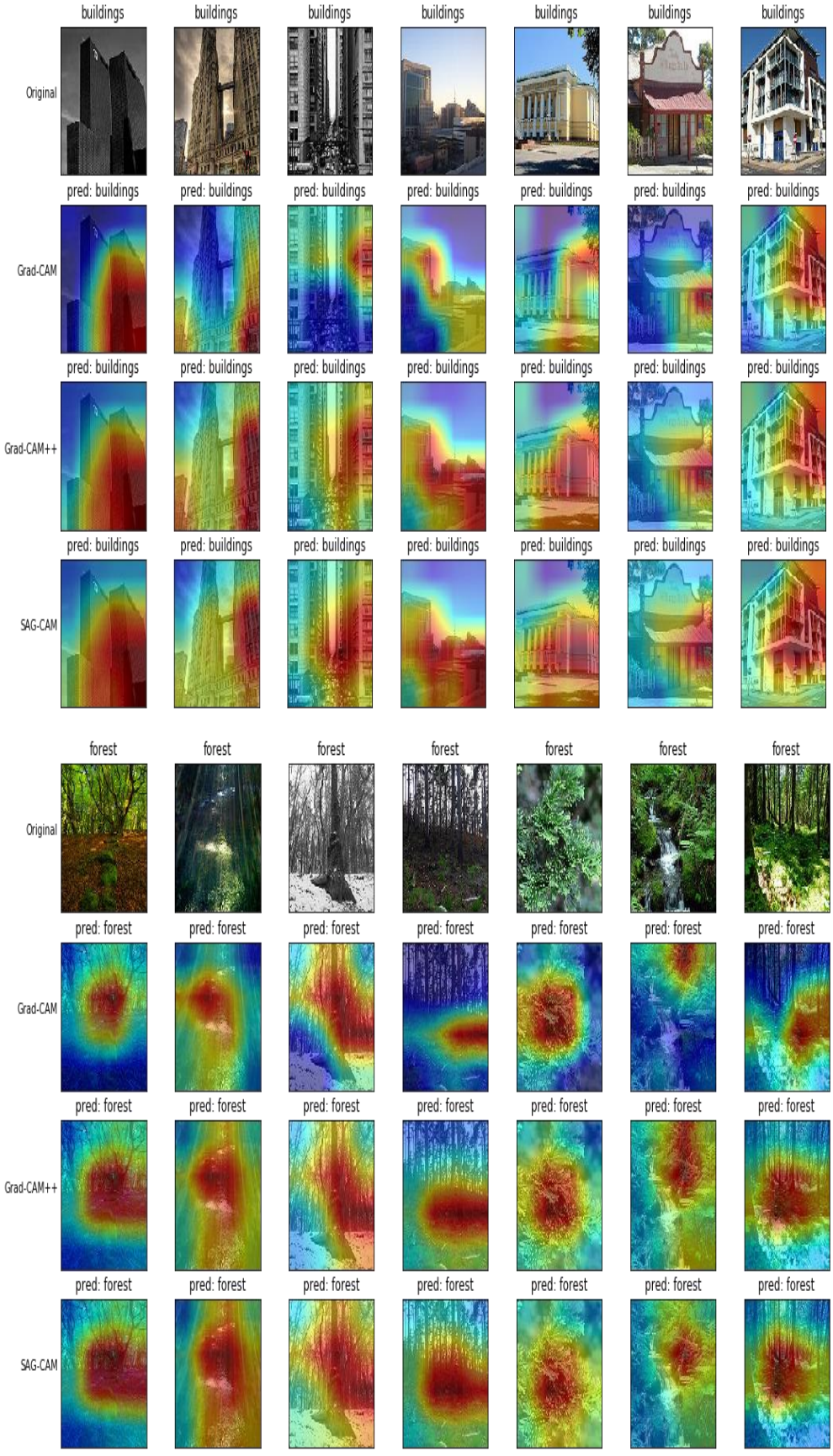

Figure 5: Visual comparison between explanation heatmaps
or model predictions on building and forest classes.

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

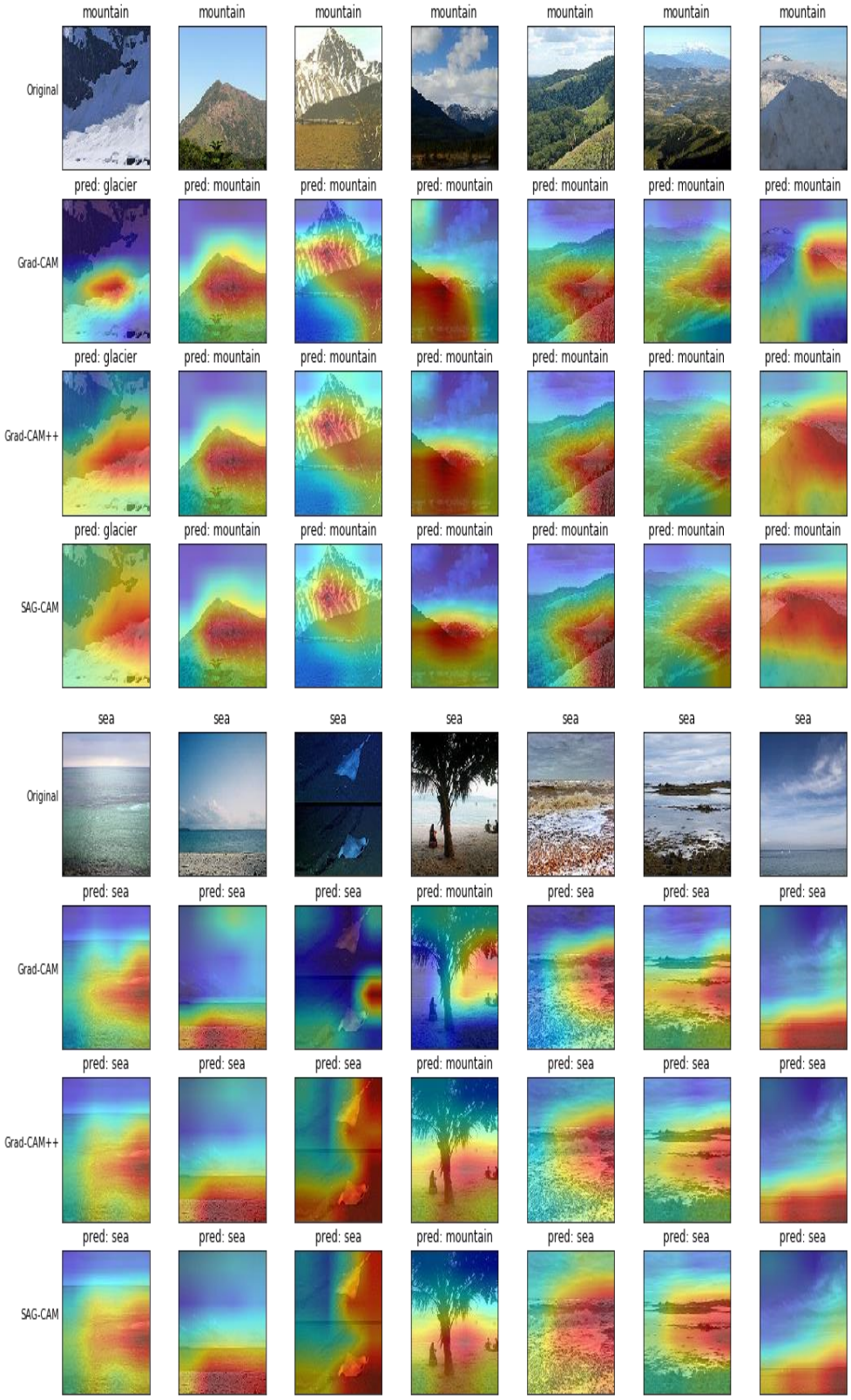

Figure 6: Visual comparison between explanation heatmaps or model predictions on mountain and sea classes.

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

5. Conclusion

In conclusion, our work benchmarked and evaluated various visual explanation methods for deep CNNs on the Intel Image Classification dataset. Through extensive experimentation, we gained valuable insights into the interpretability, performance, and robustness of these methods. Our findings demonstrate the effectiveness of certain techniques, such as SAC-CAM, Grad-CAM, and Grad-CAM++, in generating informative heatmaps that highlight important regions in the input images. Additionally, we visualized the learning curves of our CNN model, providing insights into its convergence and generalization capabilities. By combining explanation heatmaps and model predictions, we obtained a comprehensive understanding of the decision-making process of the model and its ability to classify different image classes.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Ramaswamy, Harish Guruprasad, Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020.
[2] Fei Z. et al., Deep convolution networkbased emotion analysis towards mental health care. Neurocomputing, 388, 212-227, 2020.
[3] Hägele M et al., Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Scientific reports, 10(1), 1-12, 2020.
[4] 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. In Proceedings of the IEEE international conference on computer vision, 618-626, 2017.
[5] Nayak R., Pati U. C., Das S. K., A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106, 104078, 2021.
[6] Mohamed Saber, Efficient phase recovery system, IJEECS, 5(1), 2017.
[7] Zhong B., Pan X., Love P. E., Ding L., Fang W. , Deep learning and network analysis: Classifying and visualizing accident narratives in construction. Automation in Construction, 113, 103089, 2020
[8] Lee J. H., Han S. S., Kim Y. H., Lee, C., Kim I, Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. Oral surgery, oral medicine, oral pathology and oral radiology, 129(6), 635-642, 2020.
[9] Mohamed Saber, A novel design and implementation of FBMC transceiver for low power applications. IJEEI, 8(1), 83-93, 2020.
[10] Kitaguchi D., Takeshita N., Matsuzaki H., Takano H., Owada Y., Enomoto T., Ito M., Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. Surgical endoscopy, 34, 4924-4931, 2020.
[11] Van der Velden, B. H. Kuijf, H. J. Gilhuijs, K. G., Viergever, M. A., Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 102470, 2022.
[12] Roy S., Menapace W., Oei S., Luijten B., Fini E., Saltori C., Demi L., Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. IEEE transactions on medical imaging, 39(8), 2676-2687, 2020.
[13] Wang H., Wang Z., Du M., Yang F., Zhang Z., Ding S.,Hu X., Score-CAM: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 24-25, 2020.
[14] Naeem H., Ullah F., Naeem M. R., Khalid S., Vasan D., Jabbar S., Saeed S., Malware detection in industrial internet of things based on hybrid image visualization and deep learning model. Ad Hoc Networks, 105, 102154, 2020.

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023

[15] Arshad H., Khan M. A., Sharif M. I., Yasmin M., Tavares J. M. R., Zhang Y. D., Satapathy S. C., A multilevel paradigm for deep convolutional neural network features selection with an application to human gait recognition. Expert Systems, 39(7), e12541, 2022.

[16] Abouelatta, Mohamed A., Sayed A. Ward, Ahmad M. Sayed, Karar Mahmoud, Matti Lehtonen, and Mohamed MF Darwish, Measurement and assessment of corona current density for HVDC bundle conductors by FDM integrated with full multigrid technique. Electric Power Systems Research, 199, 107370, 2021.

Doi: https://doi.org/10.54216/JAIM.040103
Received: October 12, 2022 Revised: January 23, 2023 Accepted: June 11, 2023