Visualizing the Unseen: Exploring GRAD-CAM for Interpreting Convolutional Image Classifiers

Sunil Kumar[*1], Abdelaziz A. Abdelhamid[2], Zahraa Tarek[3]

[1] School of Computer Science, University of Petroleum and Energy Studies, Dehradun, 248001, India
[2] Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[3] Computer Science Department, Faculty of Computers and Information, Mansoura University
Emails: skumar@ddn.upes.ac.in; abdelaziz@cis.asu.edu.eg; zahraatarek@mans.edu.eg

Abstract

Mathematical programming can express competency concepts in a well-defined mathematical model for a particular. Convolutional Neural Networks (CNNs) and other deep learning models have shown exceptional performance in image categorization tasks. However, questions about their interpretability and reliability are raised by their intrinsic complexity and black-box nature. In this study, we explore the visualization method of Gradient-Weighted Class Activation Mapping (GRAD-CAM) and its application to understanding how CNNs make decisions. We start by explaining why tools like GRAD-CAM are necessary for deep learning and why interpretability is so important. In this article, we provide a high-level introduction to CNN architecture, focusing on the significance of convolutional layers, pooling layers, and fully connected layers in the context of image categorization. Using the Xception model as an illustration, we describe how to generate GRAD-CAM heatmaps to highlight key areas in a picture. We highlight the benefits of GRAD-CAM in terms of localization accuracy and interpretability by comparing it to other visualization techniques like Class Activation Mapping (CAM) and Guided Backpropagation. We also investigate GRAD-CAM's potential uses in other areas of image classification, such as medical imaging, object recognition, and fine-grained classification. We also highlight the disadvantages of GRAD-CAM, such as its vulnerability to adversarial examples and occlusions, along with its advantages. We conclude by highlighting extensions and changes planned to address these shortcomings and strengthen the credibility of GRAD-CAM justifications. As a result of the work presented in this research, we can now analyze and improve Convolutional Image Classifiers with greater accuracy and transparency.

Keywords: Grad-CAM; Convolutional Neural Networks; Interpretability; Visualization; Explainable Ai, Deep learning.

1. Introduction

When it comes to essential decision-making, interpretability of deep learning models is crucial to their general adoption and acceptance. Convolutional Neural Networks (CNNs) and other deep learning models have shown amazing success in a wide range of applications, from image classification and object detection to NLP and speech recognition. In many cases, however, their inner workings are not made obvious, making them seem like black boxes that make reliable predictions but offer nothing in the way of explanation. Concerns regarding trust, reliability, and accountability are raised due to the

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

lack of transparency, which is especially problematic in high-stakes contexts such as healthcare, banking, and autonomous systems. Therefore, the capacity to explain how a model gets at its predictions (known as interpretability) has become an increasingly important part of deep learning study and application [1-2].

Deep learning models benefit greatly from interpretability. First, it increases confidence and approval from customers, stakeholders, and authorities. Interpretability increases confidence and decreases reliance on blind trust by offering clear explanations for the model's decision-making process, giving consumers insight into why a certain prediction was produced. Secondly, model biases and flaws can be found and corrected if the model is interpretable [3]. Identifying and fixing potential biases or flaws in the training data or model architecture can be accomplished by digging into the nuts and bolts of how predictions are made. As a result, judgements are more consistent and fairer because they are not influenced by unspoken biases. Thirdly, interpretability makes it easier to fix and improve models. Researchers can identify problems like overfitting, sensitivity to adversarial assaults, or mislabeled training data by gaining insights into the characteristics and patterns the model focuses on during its decision-making process. In the end, researchers, practitioners, and consumers are given the tools they need to fully appreciate, enhance, and ethically deploy deep learning models in the real world [4-5].

CNNs are a type of deep learning model specifically designed for processing and analyzing visual data, making them particularly effective in image classification tasks. CNNs leverage the concept of convolution, which involves sliding a small filter or kernel over an input image to extract meaningful features [6]. These features are learned through the model's training process, where the network adjusts its weights based on labeled training data. CNNs consist of multiple layers, typically including convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply filters to capture different visual patterns, such as edges, textures, and shapes, across the image. Pooling layers down sample the feature maps, reducing their spatial dimensions while retaining the most salient information. Fully connected layers connect the extracted features to the final output layer, which represents the predicted classes or probabilities [7].

In image classification tasks, CNNs have demonstrated exceptional performance and achieved state-of-the-art results. They can accurately classify images into various categories, such as recognizing objects, distinguishing between different species, identifying handwritten digits, or detecting diseases in medical images [8]. CNNs excel at capturing hierarchical representations, progressively learning abstract features from low-level edges to high-level object concepts. This hierarchical feature extraction enables CNNs to effectively model complex visual relationships, leading to robust and accurate image classification capabilities. As a result, CNNs have found applications in a wide range of domains, including autonomous driving, image-based search engines, medical diagnostics, quality control, and many more, revolutionizing the field of computer vision [9].

The need for explain ability techniques like GRAD-CAM arises from the increasing complexity and opacity of CNNs. As deep learning models continue to evolve and achieve remarkable accuracy in various tasks, their decision-making processes become less interpretable and more akin to black boxes. This lack of transparency poses significant challenges, particularly in critical domains where the justification for predictions is crucial. Explain ability techniques like GRAD-CAM provide insights into the inner workings of CNNs by highlighting the important regions and features in an input image that contribute to a particular prediction [10]. By visualizing these regions, researchers, practitioners, and end-users can understand the reasoning behind the model's decisions, identify potential biases or errors, and gain trust and confidence in the model's outputs. Explain ability techniques not only enhance accountability and transparency but also facilitate model improvement, debugging, and validation. They allow us to validate the model's reliance on meaningful features and identify any potential shortcomings or limitations. Ultimately, explain ability techniques like GRAD-CAM bridge the gap between the remarkable performance of CNNs and the need for understandable and trustworthy decision-making processes, enabling their responsible deployment in critical applications [10-12].

2. Background

CNNs are a type of deep learning architecture widely used for processing visual data, including images. The basic components of a CNN include convolutional layers, pooling layers, and fully connected layers. Let's take the Xception model as an example to illustrate these components. The

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

Xception model is a deep CNN architecture that achieved state-of-the-art performance on the ImageNet dataset. It is known for its extreme inception-like module design as shown in figure 1.

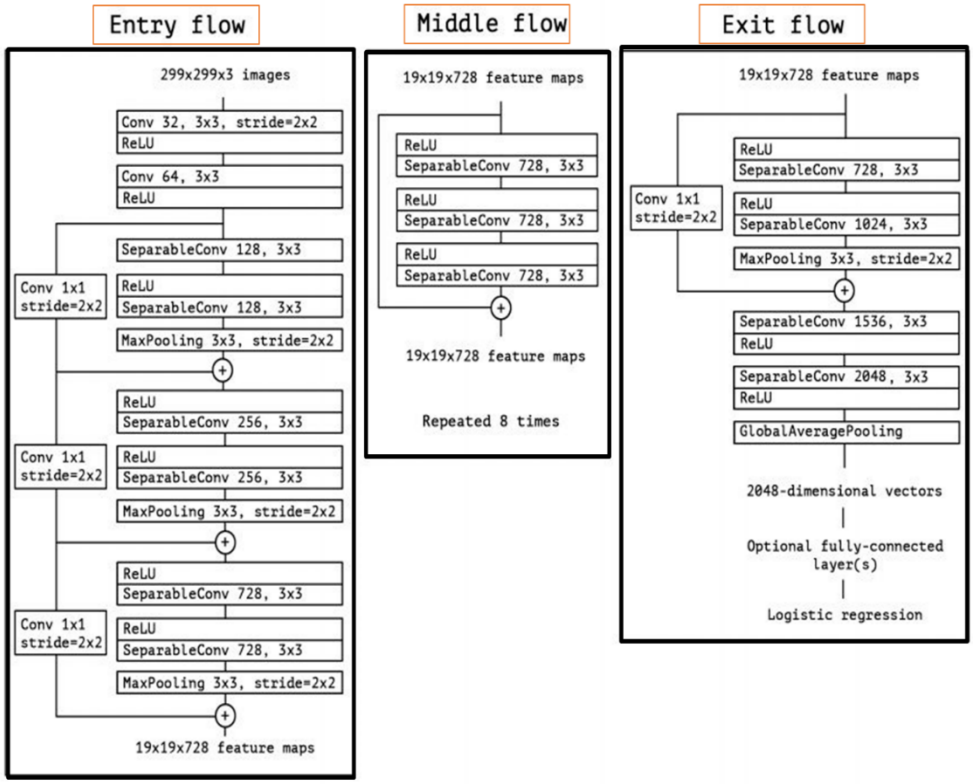


Figure 1: Visualization of architecture of Xception Network

It is the job of the convolutional layers to pick up specific details and patterns in the incoming data. By stacking convolutional layers, the Xception model can learn more sophisticated images. As part of the feature extraction process, each convolutional layer convolves the input data with a set of learnable filters or kernels. A feature map is what is produced by a convolutional layer.

$$feature\_surface_{out} = f(\sum_{i=3}^{3} M_i * W_i + B) \quad (1)$$

Pooling layers are used to downsample the spatial dimensions of the feature maps while retaining the most salient information. This helps reduce the computational complexity and makes the network more robust to spatial variations. Common pooling operations include max pooling and average pooling. In the Xception model, max pooling operation (See Figure 2) is applied within the inception modules to reduce the spatial dimensions and capture relevant information.

$$o = [\frac{(i-k)}{s} + 1] \quad (2)$$

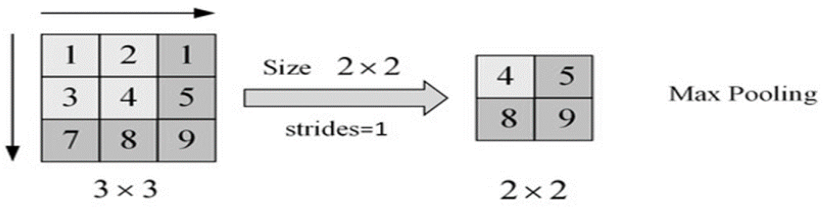


Figure 2: Illustration of max pooling operation

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

Linear layers are responsible for connecting the extracted features to the final output layer, which produces the predictions. In the Xception model, after the convolutional and pooling layers, there is a transition from spatial features to a traditional fully connected architecture. These Linear layers incorporate dense connections between neurons, where each neuron is connected to every neuron in the previous layer. The final Linear layer of the Xception model typically consists of a softmax activation function, which generates the class probabilities for image classification tasks. Feature maps play a crucial role in CNNs as they capture different levels of abstractions in an image. These feature maps are obtained by applying convolutional filters to the input image. At earlier layers of the network, the feature maps represent low-level features such as edges, corners, and textures.

As the network progresses deeper, the feature maps become increasingly abstract and capture higher-level concepts or semantic information. This hierarchical representation allows CNNs to learn complex patterns and relationships in the data. By capturing both local and global structures, feature maps enable the network to discern essential discriminative features for classification or other tasks. The ability of feature maps to capture varying levels of abstractions empowers CNNs to recognize not only low-level visual attributes but also higher-level concepts, leading to their remarkable performance in image understanding and analysis tasks [13].

3. GRAD-CAM Explanation

To better understand how various parts of an image contribute to a CNN's predictions, the GRAD-CAM (Gradient-weighted Class Activation Mapping) method can be used. GRAD-CAM delivers a more comprehensive perspective than previous visualization approaches by emphasizing the complete regions in the input image that are critical to the CNN's decision-making process. The gradient information from the final prediction layer of the network is used by GRAD-CAM to create a class activation map that pinpoints the specific areas that are important for the predicted class. By superimposing this heatmap over the original image, GRAD-CAM effectively shows which regions of the image had the greatest impact on the network's final verdict. Transparency, trust, and insights into the logic behind the model's predictions are all bolstered by this interpretability technique, which not only improves our comprehension of CNNs' decision-making process, but also helps discover crucial visual cues and discriminative characteristics used by the network.

The underlying principle of GRAD-CAM lies in leveraging the gradient information flowing back from the final prediction layer of a CNN to identify and highlight the relevant image regions, $L^c_{GradCAM}\mathbb{R}^{u\times v}$, for a specific class, c. GRAD-CAM achieves this by computing the gradients of the target class with respect to the feature maps, A^k, in the last convolutional layer of the CNN.

$$y_c = \frac{\partial y^c}{\partial A^k} \tag{3}$$

To generate the class activation map, GRAD-CAM multiplies the gradients with the corresponding feature map values. This multiplication emphasizes the importance of each spatial location in the feature map, indicating the relevance of that region to the target class. By summing up these weighted feature map activations, a single heat map is obtained, representing the importance of different regions in the input image for the predicted class.

$$Weights_{NeuronImportance} = \alpha_k^c = \frac{1}{Z}\sum_{i=1}^{u}\ \sum_{i=1}^{v}\ y_c \tag{4}$$

To highlight the relevant image regions, the generated heat map is overlaid onto the original image, where intense colors or high activation regions indicate the areas that contributed the most to the CNN's prediction.

$$\alpha_k^c = \frac{1}{Z}\sum_{i=1}^{u}\ \sum_{i=1}^{v}\ \frac{\partial y^c}{\partial A^k} \tag{5}$$

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

This visualization allows for a clear understanding of which regions the network attended to during the decision-making process and provides insights into the features or patterns the CNN found important for classifying the input image. The GRAD-CAM technique's effectiveness lies in its ability to provide interpretable visual explanations by linking the network's internal activations to specific image regions. By highlighting the relevant regions, GRAD-CAM helps users and researchers gain insights into the decision-making process of CNNs and enhances the transparency and interpretability of these deep learning models. The pipeline of GRAD-CAM calculation is depicted in Algorithm 1 in figure 3.

Algorithm 1. GRAD-CAM Explain ability method

Input: Pre-trained CNN model, Input image, Target class

Step 1: Pass the input image through the CNN model to obtain the final prediction probabilities or logits. $L^{c}_{GradCAM} = ReLU(\sum_{i=1}^{k} \alpha_k^c)$

Step 2. Calculate the gradients of the target class score with respect to the feature maps in the last convolutional layer.

Step 3. Multiply the gradients with the corresponding feature map values to obtain the weighted feature map activations.

Step 4. Sum up the weighted feature map activations across all channels to obtain a single heatmap.

Step 5. Normalize the heatmap to ensure its values lie within a specific range, typically between 0 and 1.

Step 6. Overlay the normalized heatmap onto the original input image.

Step 7. Display the overlaid image with the heatmap, where intense colors or high activations correspond to the regions that strongly contribute to the target class prediction.

Return: Heatmap overlaid on the input image highlighting the relevant regions for the target class prediction.

Figure 3: The pipeline of GRAD-CAM calculation

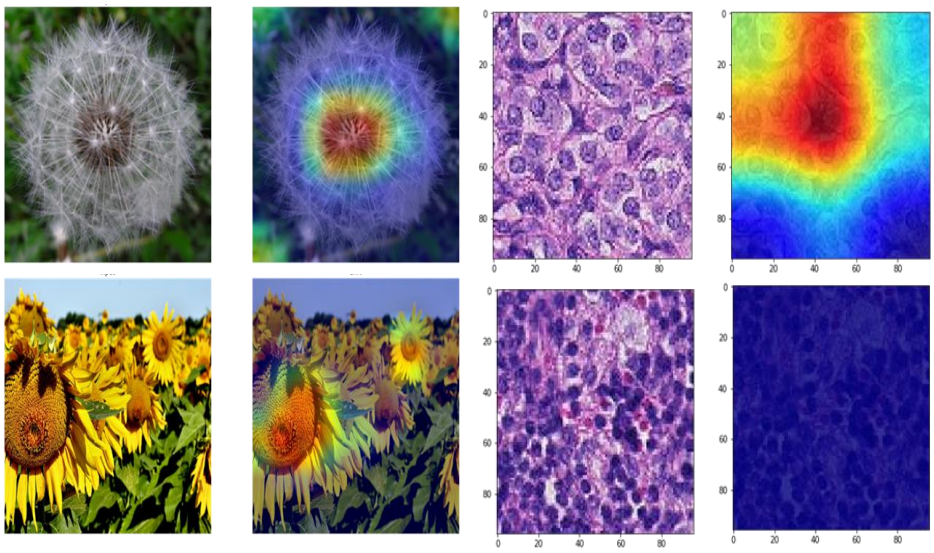
GRAD-CAM offers several advantages compared to other visualization techniques like Class Activation Mapping (CAM) and Guided Backpropagation. Unlike CAM, which is limited to global average pooling and cannot capture fine-grained details, GRAD-CAM provides more precise localization of important regions within the image by leveraging gradients. GRAD-CAM also overcomes the limitations of Guided Backpropagation, which only visualizes positive contributions without considering the importance of features in the context of the target class. In contrast, GRAD-CAM generates heatmaps that highlight both positive and negative influences, providing a more comprehensive understanding of CNN's decision- making process. Furthermore, GRAD-CAM is a gradient-based technique, making it compatible with a wide range of CNN architectures, while Guided Backpropagation requires specific modifications to the network.

4. Experimental Analysis

In this section, our analysis involved visualizing GRAD-CAM maps on the predictions of exceptions across different datasets. Figure 4 illustrates these visualizations, providing valuable insights into the underlying mechanisms of our model's exception detection capabilities. input data, we can identify the specific features or areas that drive the model's decision-making. The GRAD-CAM maps highlight the regions of input data that significantly influence the model's prediction of exceptions. By

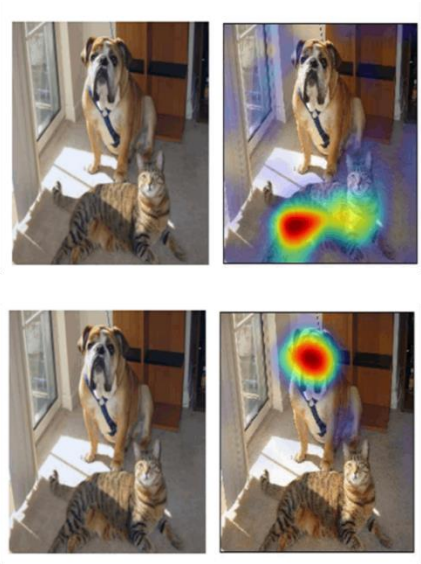
Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

overlaying the heatmaps generated by GRAD-CAM onto the original process. On flower dataset, the GRAD-CAM maps consistently highlighted the regions associated.



a) Flower data

b) Histopathologic cancer dataset



c) Dog-Cat dataset

Figure 4: Visualization of the GRAD-CAM explanation on different dataset

with known exceptions. For instance, in figure 4 (a), the map emphasized a particular area in an image where an exception was present, corroborating the model's correct prediction. This demonstrates the model's ability to effectively identify exceptions in flower dataset and aligns with our expectations. Surprisingly, when applying the same technique to histopathologic cancer dataset, the GRAD-CAM maps revealed different patterns. Figure 4 (b) illustrates an instance where the map focused on a different region of the image, which was not directly related to the known exceptions. Further analysis of the GRAD-CAM maps on dog-Cat dataset showcased mixed results. In figure 4(c), the map exhibited a strong activation on a region that was associated with an exception, indicating accurate prediction.

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

5. Application of GRAD-CAM to Convolutional Image Classification

GRAD-CAM can be used to decipher how a Convolutional Image Classifier arrives at its verdicts by revealing the image areas and characteristics on which the classifier depends. GRAD-CAM generates a heatmap overlay on the input image to emphasize regions that significantly influence the classifier's verdict. In this way, users can see exactly where the model is putting its attention in order to make accurate predictions. The classifier's strengths, shortcomings, and possible biases can be better comprehended because of this feature's ability to help users zero in on the discriminative features or patterns that influence the classifier's conclusions. GRAD-CAM aids in establishing confidence in the classifier's predictions by verifying the reasoning behind the model and pointing out any problems or limitations it may have.

There are many advantages to using GRAD-CAM to learn why a certain class was predicted for a given image. As a first step, GRAD-CAM offers visual justifications by emphasizing in the image those parts that have the greatest impact on the anticipated class. This aids users in comprehending how the model arrived at its categorization and pinpointing the precise elements or patterns that had a role in the choice. GRAD-CAM bridges the gap between the model's internal computations and human interpretability by visually representing the relevant regions, making it easier for users to understand the model's reasoning. Second, GRAD-CAM is useful for pinpointing the specific visual indicators and distinguishing characteristics that contributed to the final class prediction. GRAD-CAM helps users better understand the unique characteristics of a picture that led to a positive classification result by isolating those areas. This data can be used to validate the model's dependence on meaningful image regions, comprehend the model's focus on certain objects or features, or uncover potential biases or inaccuracies. In addition, GRAD-CAM makes it easier to fix and enhance models. By depicting the areas that the model priorities, users can spot instances of possible misclassification or over-reliance on irrelevant or deceptive features. This knowledge can be used to improve the model's performance and accuracy by altering its architecture, dataset, or training procedure.

The applications of GRAD-CAM span multiple eras including:

- Medical Imaging: GRAD-CAM has been employed in medical imaging tasks to understand the decision-making process of CNNs in disease diagnosis. For example, in a study focused on chest X-ray classification, GRAD-CAM was used to highlight the regions of the image that contributed most to the classification of specific diseases like pneumonia or tuberculosis. This helped clinicians and researchers gain insights into the important visual cues and areas of interest for accurate disease identification.

- Object Recognition: GRAD-CAM has been utilized in object recognition tasks to interpret the predictions of CNNs. For instance, in a study on image-based food recognition, GRAD-CAM was employed to visualize the regions in food images that influenced the classification results. This provided users with explanations regarding the specific parts of the food items that contributed most to their recognition, helping to validate the model's decisions and potentially aiding in quality control or dietary analysis.

- Fine-Grained Classification: GRAD-CAM has been valuable for understanding fine-grained classification tasks, where distinguishing between closely related subcategories is crucial. In a case study focused on bird species classification, GRAD-CAM was used to identify the regions of bird images that played a pivotal role in discriminating between different species. This facilitated the interpretation of the model's predictions and helped bird enthusiasts and ornithologists understand the distinguishing characteristics utilized by CNN for accurate species identification.

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

- Adversarial Attack Detection: GRAD-CAM has also been utilized in the detection of adversarial attacks on image classifiers. By visualizing the regions of an image that contribute to the model's predictions, GRAD-CAM can highlight the areas that are susceptible to adversarial perturbations. This insight aids in identifying vulnerabilities and developing robust defense mechanisms against adversarial attacks.

6. Conclusion

This paper provided a comprehensive outlook on the GRAD-CAM explanation technique for Convolutional Image Classifiers. The main findings of the paper include the importance of interpretability in deep learning models and the need for techniques like GRAD-CAM to gain insights into their decision-making process. We discussed the basics of CNN architecture, including convolutional layers, pooling layers, and fully connected layers, and provided an example with the Xception model. The steps involved in generating a GRAD-CAM heatmap were outlined, highlighting its ability to visualize the importance of different regions in an image. Furthermore, we emphasized the advantages of GRAD-CAM over other visualization techniques and discussed its applications in various image classification tasks. The limitations and potential drawbacks of GRAD-CAM, such as its sensitivity to adversarial examples or occlusions, were also addressed. Lastly, proposed extensions and improvements were mentioned to mitigate these limitations and enhance the reliability of GRAD-CAM explanations.

Funding: "This research received no external funding"

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

References

[1] Wang, Haofan, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu., 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.

[2] Vinogradova, Kira, Alexandr Dibrov, and Gene Myers, Towards interpretable semantic segmentation via gradient-weighted class activation mapping. In Proceedings of the AAAI conference on artificial intelligence, 34(10), 13943-13944, 2020.

[3] Selvaraju et al., Choose your neuron: Incorporating domain knowledge through neuron-importance. In Proceedings of the European conference on computer vision (ECCV), 526-541. 2018.

[4] Mohamed Saber, Efficient phase recovery system, IJEECS, 5(1), 2017.

[5] Sequeira Ana F., Wilson Silva, Joao Ribeiro Pinto, Tiago Gonçalves, and Jaime S. Cardoso, Interpretable biometrics: Should we rethink how presentation attack detection is evaluated?. In 2020 8th International Workshop on Biometrics and Forensics (IWBF), 1-6, 2020.

[6] Alshazly Hammam, Christoph Linse, Erhardt Barth, and Thomas Martinetz, Ensembles of deep learning models and transfer learning for ear recognition. Sensors 19(19), 2019.

[7] Chen L., Chen J., Hajimirsadeghi H., & Mori, G., Adapting grad-cam for embedding networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2794-2803, 2020.

[8] Panwar Harsh, P. K. Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, and Vaishnavi Singh, A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. Chaos, Solitons & Fractals, 140, 2020.

[9] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, Grad-cam: Visual explanations from deep networks via gradient-based

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023

localization. In Proceedings of the IEEE international conference on computer vision, 618-626. 2017.

[10] Mohamed Saber, A novel design and implementation of FBMC transceiver for low power applications. IJEEI, 8(1), 83-93, 2020.

[11] Omeiza, Daniel, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint arXiv:1908.01224, 2019.

[12] Wang, Haofan, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu, 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.

[13] Jonas Stefan, Andrea O. Rossetti, Mauro Oddo, Simon Jenni, Paolo Favaro, and Frederic Zubler. "EEG-based outcome prediction after cardiac arrest with convolutional neural networks: Performance and visualization of discriminative features. Human brain mapping, 40(16), 4606-4617, 2019.

Doi: https://doi.org/10.54216/JAIM.040104
Received: October 18, 2022 Revised: January 28, 2023 Accepted: June 14, 2023