Defence Science Journal, Vol. 75, No. 2, March 2025, pp. 188-198, DOI : 10.14429/dsj.20444

Pixel Ablation-CAM: A New Paradigm in CNN Interpretability for Feature Map Visual Explanations

Akash Samanta[#], Debasis Chaudhuri[#], Aniket Kumar Singh[#] and Manish Pratap Singh[$]

[#]*Techno India University, Salt Lake, Kolkata – 700091, India*
[$]*DRDO-Young Scientist Laboratory – CT, Chennai – 600113, India*
[*]*E-mail: debasis.chaudhury@technoindiaeducation.com*

ABSTRACT

Many cutting-edge computer vision systems now rely heavily on convolutional neural networks, or CNNs. However, conventional interpretation techniques frequently concentrate on 2D feature maps, ignoring the intricate contributions of individual pixels. This work aims to produce "visual explanations" that improve the explainability and transparency of decisions made by various CNN-based algorithms. We provide Pixel Ablation-CAM, a new method that builds on the ideas of Ablation-CAM by using pixel-wise ablation, which enables a finer-grained comprehension of model choices. With this method, activation maps are reinterpreted as arrays of one-dimensional vectors that represent channel-specific pixel activations. We show that, as compared to other approaches such as Grad-CAM, Pixel Ablation-CAM offers better resolution and accuracy in class-discriminative localisation maps by methodically ablating these vectors and monitoring changes in class activation scores. Our extensive testing demonstrates that Pixel Ablation-CAM improves model trust and interpretability, providing fresh perspectives on CNN behavior and propelling the field of explainable AI forward.

Keywords: Convolution Neural Network (CNN); Ablation-CAM; Grad-CAM; Explainable AI

1. INTRODUCTION

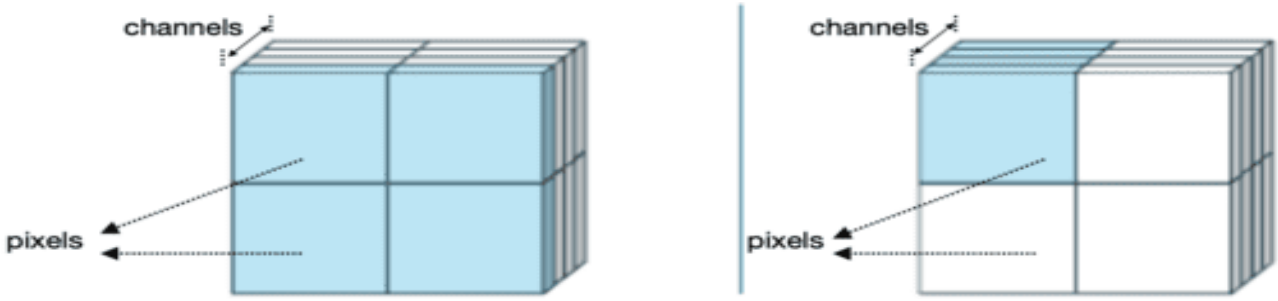
In essential industries, automated decision-making systems-especially those driven by deep learning are being utilised more and more. However, since blind trust is uncommon, users anticipate justifications for these choices. Numerous elements, such as trust, causation, transferability, accountability, and egalitarian decision-making, are what motivate this requirement for transparency[1]. Explainability is therefore essential to building confidence in these systems, particularly in high-stakes domains such as banking, healthcare, autonomous driving, and defence.

Convolutional Neural Nets (CNNs)[2] have significantly enhanced performance in several computer vision tasks, such as semantic segmentation[5], object detection[4], and image classification[3]. CNN designs have improved in accuracy and efficiency over time[6]. CNNs function as "black boxes," making it challenging for consumers to believe their predictions even in spite of their success[7]. Therefore, enhancing CNNs[8] interpretability is crucial for their incorporation into sectors where dependability and trust are unavoidable.

A number of approaches have been put up to solve CNN interpretability; gradient-based approaches, such as Grad-CAM[9] and Grad-CAM++[10], have drawn a lot of interest. By determining the areas of an image that have the greatest influence on the model's conclusion, these techniques produce visual explanations. Grad-CAM is not without its limits, though, especially when it comes to precisely localising numerous instances of an item or adequately capturing the area of an object, which can result in incomplete or misleading explanations[11]. Non-gradient-based techniques like Ablation-CAM[12] have been developed in response to these difficulties. In order for ablation-CAM to function, individual channels in the final convolutional layer are methodically removed, and the resulting changes in activation scores are monitored. This channel-wise ablation method is useful in many situations, but it is still too coarse and frequently falls short of offering a complete picture of the CNN's decision-making process.

We present a new method called Pixel Ablation-CAM, which extends ablation to individual pixels in the last convolutional layer to provide a more detailed interpretation. Rather of eliminating entire channels, our method systematically eliminates each pixel and evaluates the shifts in activation ratings to assign relevance to each one. This pixel-by-pixel ablation highlights the areas of a picture that have the biggest impact on predictions, enabling a more accurate understanding of the CNN's behavior. By concentrating on pixel vectors, Pixel Ablation-CAM not only overcomes the dimensionality constraints of conventional Ablation-CAM but also raises the overall resolution of visual explanations. This approach is more suitable for real-world applications where trust and transparency are essential since it provides more thorough localisation of significant features. For example, in autonomous driving, clearer explanations can enable safer navigation by detecting essential environmental cues, and in healthcare, more accurate interpretability can help with medical picture diagnosis by highlighting crucial regions in an X-ray or MRI scan. These uses highlight Pixel Ablation-CAM's wider

Received : 06 August 2024, Revised : 25 November 2024
Accepted : 06 January 2025, Online published : 24 March 2025



(a) Ablation-CAM approach (b) Pixel Ablation-CAM approach

Figure 1. Pixel Ablation-CAM overview: Given an image as input, the left image shows how the image is handled in Ablation-CAM, where each channel is evaluated by removing it to assert its importance. The right image illustrates the proposed method, where each pixel is treated as a 1D vector of channels and dropped out to determine its significance, offering a new perspective on CNN layer evaluation.

significance and its capacity to close the gap between model accuracy and interpretability in systems that are vital to the mission. Pixel Ablation-CAM can improve transparency in the face recognition space by pinpointing the precise facial features that influence a match or mismatch determination the most. By emphasizing the most influential regions, this not only increases the system's trustworthiness but also helps refine models, guaranteeing that the system makes judgments based on consistent and dependable qualities.

In this research, we present a comprehensive methodology for Pixel Ablation-CAM, empirically compare it with other approaches, and show that it can improve CNN interpretability on a range of tasks. An overview of our methodology is shown in Fig. 1, which compares our pixel-wise approach with conventional channel-based ablation. This contribution is expected to make a significant impact in the field of explainable AI by providing practitioners and researchers with an effective tool for understanding the inner workings of deep learning models.

2. LITERATURE REVIEW

The phrases "black box," "gray box," and "white box" relate to various degrees of closure of a component's internal essence in science, computing, and engineering. Specifically, a white box component presents all of its internal design, structure, and implementation to the user, while a black box component keeps all of this information hidden. Depending on the amount of information provided, there may be varying degrees of gray box components in between. Technological businesses have used the black box idea for commercial purposes; often, this has been done to preserve efficiency and protect intellectual property. The black box problem in AI refers to the inability of the system to adequately explain how it concluded. The three primary tenets of responsible AI are transparency, accountability, and responsibility.

Two visual aids for assisting in the interpretation of trained neural nets were presented by Yosinski[13], *et al*. One of the tools is designed to show the activations generated on every layer of a convolutional neural network (CNN) while it processes an image or video, such as a live webcam clip. They have made clear that observing real-time activations that alter in response to user input is a useful tool for developing insightful intuitions about convolutional neural networks. The second tool uses regularized optimization in image space to allow the visualization of features at each layer of a DNN. They included some novel regularisation techniques here because earlier iterations of this concept yielded less identifiable images. These techniques together result in visuals that are more interpretable and qualitatively clear.

The topic of comprehending classification judgments by pixel-wise breakdown of nonlinear classifiers was addressed extensively by Bach[14], *et al.* They presented a system that enables the visualization of individual pixel contributions to predictions for multi-layered neural networks and kernel-based classifiers over Bag of Words data. These pixel contributions are given to a human expert, who may intuitively confirm the accuracy of the classification conclusion and direct additional investigation toward areas of possible interest. These contributions can be visualized as heatmaps.

Zeiler and Fergus[15] have suggested massive convolutional neural network models that undergo conditioning for image classification. They started by demonstrating a fresh method for representing the activity inside the model. This demonstrates that the features are not at all random or illogical patterns. Instead, as the layers rise, they exhibit several naturally appealing characteristics like compositionality, growing invariance, and class discrimination. They noticed that there are ways to use this visualization to pinpoint model flaws and improve outcomes. SHAP (SHapley Additive exPlanations), a unified paradigm for interpreting predictions, was given by Lundberg and Lee[16]. For a given prediction, SHAP allocates an importance value to every feature. To provide visual explanations for the classification networks, Zhou[17], *et al.* proposed a new framework called Interpretable Basis Decomposition. This framework can quantify the contribution of each piece of evidence to the final prediction and disentangle the evidence encoded in the activation feature vector.

The survey papers[25-26,30] examine the evolution and importance of XAI research across a range of applications and disciplines. Throughout this research, a comprehensive collection of explainability summaries and classifications, as well as their applications and real-world use cases, has been created. The paper[27] offers a thorough examination of several XAI solutions, stressing both their advantages and disadvantages. While XAI has several benefits, such as enhanced stakeholder acceptance and trust in AI systems, the

capacity to recognise and reduce potential biases in AI models, and enhanced debugging and model-improvement capabilities, it also has drawbacks. Explainable artificial intelligence (XAI) in deep learning-based medical image processing was covered in the survey paper[28]. Images were categorized based on imaging technique and anatomical location inside an XAI framework. The study covered XAI evaluation techniques, recent criticisms of XAI, and potential applications of XAI in medical image processing.

The significance of the notion of XAI was examined by Reddy and Pavan[29]. Details are provided on some techniques, including saliency mapping, feature visualisation, and model interpretation. LIME and SHAP approaches were given particular consideration. The suggested study expands upon notable developments in saliency method evaluation[18], model interpretability[18], and CNN visualization techniques[6]. Here, we go through the major achievements made in these fields as well as their shortcomings, which highlight the need for better techniques.

2.1 Visualizing CNNs

Early efforts in CNN interpretation, such as those by Adebayo[19], *et al.*, utilised deconvolution approaches to trace neuron activations back to input pixels, providing insights into which image regions were responsible for specific neuron activations. Simonyan[20], *et al*. introduced class-specific saliency maps derived from the partial derivatives of class scores concerning input pixels, but these methods often lacked class discriminability, producing nearly identical visualizations across different classes. To increase the transparency and explainability of decisions made by a wide class of Convolutional Neural Network (CNN)-based models, Selvaraju[9], *et al*. suggested a method for creating "visual explanations." Their gradient-weighted Class Activation Mapping (Grad-CAM) creates a coarse localisation map that highlights the crucial areas in the image for idea prediction by using the gradients of each target concept flowing into the final convolutional layer. Grad-CAM's visualisation[9] provides fine-grained details of the projected class to illustrate the CNN-based model prediction, but it is not able to localize numerous instances of the same class. Additionally, the Grad-CAM heatmap's localisation is not particularly good when it comes to encompassing the class region in an image. To fix the limitations of the Grad-CAM process Grad-CAM++, a modified method, was presented by Chattopadhyay[10], *et al*. Their approach tackles the drawbacks of Grad-CAM, such as inadequate object localisations and many instances of the same class in one image. A time-efficient method called Ablation-CAM++ was presented by Salama[31], *et al*.; it can produce smooth explanations that are equivalent to Ablation-CAM in a lot less time. These techniques do, however, still have drawbacks, such as gradient saturation, which might mask the actual significance of particular features and lead to incomplete object renderings.

2.2 Non-Gradient Based Approaches

To address some of the limitations of gradient based methods, researchers have explored non-gradient based approaches like Ablation-CAM[21]. This method involves systematically removing channels in the final convolutional layer and observing the impact on the class activation score. While Ablation-CAM avoids issues related to gradient saturation, it often reduces the feature map's dimensionality, potentially losing fine-grained spatial information necessary for precise localisation.

2.3 Evaluating Trust in Models

Establishing confidence in CNN models is essential, especially for applications that demand great reliability. Approaches like LIME[7] and DeepLIFT[8] offered frameworks for importance attribution and local approximation, respectively. Although these approaches were less scalable for big, sophisticated CNNs and frequently required significant processing resources, they attempted to provide interpretable reasons for model predictions. Morcos[22], *et al*. and Zhou[12] *et al*. ablation investigations emphasized the significance of particular neurons and feature maps in defining network outputs. These findings show that the ablation of important neurons can have a substantial impact on network performance, highlighting the necessity for accurate interpretability techniques that can pinpoint and elucidate the significance of individual neurons.

2.4 Limitations of Existing Methods

Despite these advancements, current visualization techniques face several challenges. Gradient-based methods like Grad-CAM and Grad-CAM++ are prone to gradient saturation, which can result in suboptimal localisation and incomplete object highlighting. Non-gradient-based approaches such as Ablation-CAM, while avoiding gradient-related issues, may lack the necessary resolution and specificity due to the ablation of entire channels rather than more granular elements.

3. PROPOSED PIXEL-WISE ABLATION-CAM STRATEGY

Traditional convolutional neural network (CNN) architectures operate on the principle of learning 2D feature filters that capture local patterns in the input data. The activation map at the final convolutional layer is typically viewed as a collection of these 2D feature maps, with each map representing the activation of a particular learned filter. An alternative perspective is to consider the activation map as a collection of 1D feature vectors, where each spatial location i, j in the activation map is represented by a 1D vector of length equal to the number of feature channels c. This 1D vector encodes the activations of the c feature channels at that particular spatial location.

Inspired by this different perspective, we suggest an Ablation-CAM-inspired visualisation method that we apply to the 1D feature vectors in the activation map. Gradient information is used by conventional techniques like Grad-CAM and Grad-CAM++ to produce visual explanations. Nevertheless, they frequently experience problems such as noisy gradients and gradient saturation, which produce less accurate and understandable findings[23]. Some of these constraints have been addressed by non-gradient-based

techniques like Ablation-CAM, which systematically ablate channels in the final convolutional layer and measure the effect on the activation score. This approach eliminates problems with gradients but also lowers the dimensionality of the feature maps, which may result in the loss of fine-grained spatial information.

By executing pixel-wise ablation rather than channel-wise, the proposed Pixel Ablation-CAM expands on this idea and offers a more thorough and accurate analysis. By preserving spatial features down to the pixel level, this technique not only avoids the drawbacks of gradient-based techniques but also improves the resolution and accuracy of the visual explanations. This spatial ablation approach offers a complementary perspective to typical 2D filter-based activation map representations, primarily in its ability to identify the most significant spatial locations in the activation map. Compared to conventional methods like Grad-CAM or Ablation-CAM, this can provide more insights and a clearer grasp of the model's decision-making process.

Instead of treating the n channels as distinct filters, the suggested methodology treats each pixel in the final convolutional layer as a one-dimensional vector in n directions (where n is the number of channels). Improved interpretability results from this method's ability to evaluate pixel relevance at a finer and more accurate level.

3.1 Feature Map Interpretation

We interpret the feature map as an array of $h \times w$ one-dimensional vectors:

$$FeatureMap = \left\{ V_{i,j} \middle| V_{i,j} \in \mathbb{R}^n \right\} \tag{1}$$

Here, $V_{i,j}$ represents the vector at the spatial location (i,j) in the feature map, where each vector has n components corresponding to each channel. This is analogous to any regular one-dimensional vector where n components present the n directions

3.2 Ablation Process

For each pixel (i,j), we set its vector components to zero to ablate it.

$$V_{i,j} = 0 \tag{2}$$

We then forward propagate this modified feature map through the remaining layers of the network to observe the change in the final class activation score S_c:

$$\Delta S_c = S_c(Original) - S_c(Ablated) \tag{3}$$

This change in score provides a measure of the importance of the pixel (i,j) in the network's decision-making process.

3.3 Importance Assignment

The importance score w_{ij} for each pixel at position (i,j) is calculated as:

$$w_{ij} = \frac{Y^c - Y^c_{ij}}{Y^c} \tag{4}$$

Here, Y^c is the original class activation score for class c, and Y^c_{ij} is the class activation score after setting the activation at spatial location (i,j) to zero.

This score w^c_{ij} represents the fraction of drop in the activation score of class c when the activation at the spatial location (i,j) is removed.

3.4 Visualisation Process

The computed importance scores are then used to generate a heatmap that highlights the regions of the image most influential to the network's prediction. The heatmap $L_c^{PixelAblation-CAM}$ is given by:

$$L_c^{PixelAblation-CAM} = \mathrm{Re}\,LU\left(\sum_{i,j} w^c_{ij} \cdot A^k_{ij}\right) \tag{5}$$

This visualisation retains only the positive contributions, ensuring that the final map highlights areas with a significant positive influence on the class score. The visual impact consequences of different approaches are demonstrated in Section 4.

4. EXPERIMENTAL RESULTS AND DISCUSSION

We assess Pixel Ablation-CAM using experiments that are both quantitative and qualitative. The purpose of these tests is to evaluate Pixel Ablation-CAM's performance in relation to other methods, specifically Grad-CAM, Ablation-CAM, and Ablation-CAM++, and to determine how well it provides interpretable and accurate class-conditional localisation maps. For our trials, we employed pre-trained VGG-16 and Resnet-50 models, which were trained using ImageNet24.

4.1 Visualisation Impact

The visual impact of the proposed strategy is presented in Fig. 2 and compared with other methods. Here, the inference is performed using Resnet-50, a pre-trained model on the ImageNet dataset. Original Images of targets are shown in Fig. 2(a), Fig. 2(f), Fig. 2(k), Fig. 2(p), and Fig. 2(u), in that order. Figures 2(b), Fig. 2(g), Fig. 2(l), Fig. 2(q), and Fig. 2(v) display the target heatmaps generated by Grad-CAM, in that order. Similarly, the heatmaps of the targets via Ablation-CAM are displayed in Fig. 2(c), Fig. 2(h), Fig. 2(m), Fig. 2(r), and Fig. 2(w), respectively. The heatmaps generated by Ablation-CAM++[31], an optimized version of Ablation-CAM that uses clustering and tree pruning techniques to reduce the number of forward passes and improve efficiency, are displayed in Fig. 2(d), Fig. 2(i), Fig. 2(n), Fig. 2(s), and Fig. 2(x). Finally, the heatmaps of the targets by the proposed Pixel Ablation-CAM are displayed in the last column in Fig. 2(e), Fig. 2(j), Fig. 2(o), Fig. 2(t), and Fig. 2(y). The most significant discriminative zones are highlighted in red on this heatmap. The explanation map's core concept is to create a new image that only includes the original image's highlighted subregions through the use of visualization techniques. It has been observed that, compared to the other methods, the suggested method produces more precise and targeted maps, offering superior localisation and visualization of the discriminative regions.

We tested Pixel Ablation-CAM's performance using the ResNet-50 model, which was previously trained on the ImageNet dataset. Among the crucial actions in our testing procedure are as follows.

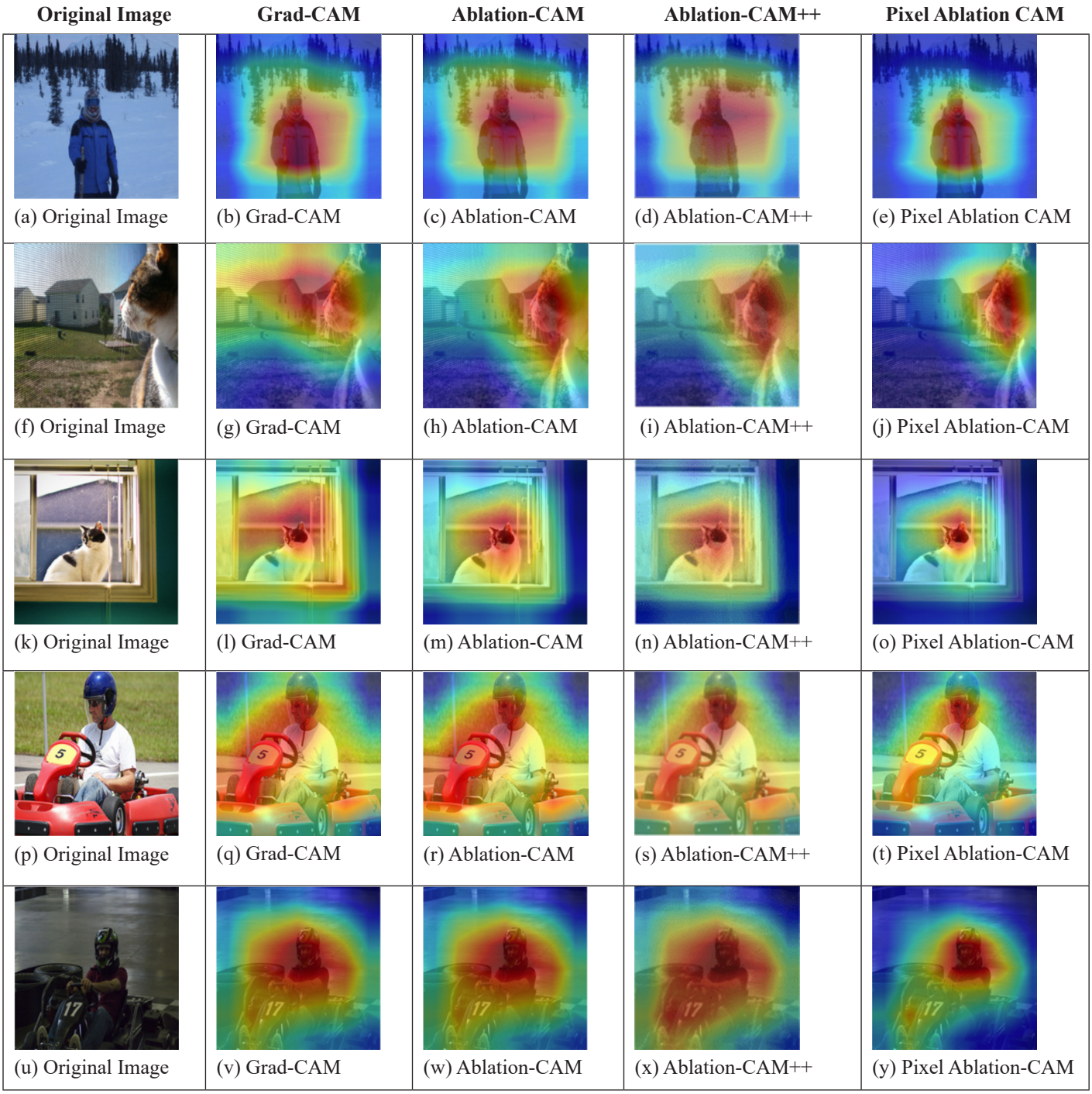


Figure 2. Visual impact: Fig. 2(a, f, k, p, u) Original image; Fig. 2(b, g, l, q, v) Grad-CAM visualization; Fig. 2(c, h, m, r, w) Ablation-CAM visualization; Fig. 2(d, i, n, s, x) Ablation-CAM++ visualization; and Fig. 2(e, j, o, t, y) Proposed Pixel Ablation-CAM visualisation.

4.1.1 Selection of Thresholds for Pixel Ablation

- The threshold for pixel ablation was selected through empirical testing, with a focus on retaining the most relevant features while minimizing noise.
- Based on their activation levels, we concluded that the top 20% of pixels from the localisation map should be kept. This decision guarantees that the most discriminative traits necessary for object recognition are highlighted in the explanation maps.

4.1.2 Evaluation Metrics

- Using a number of criteria, such as the average percentage rise in confidence and activation ratings as well as the average percentage decrease in confidence, we evaluated the visual impact of the heatmaps that were generated.
- Better localisation is shown by a lower confidence decrease, and improved explanatory power of the visualisation approach is indicated by a bigger percentage rise in scores.

4.1.3 Key Experimental Observations

- Pixel Ablation-CAM consistently produces more precise and targeted maps compared to Grad-CAM, Ablation-CAM and Ablation-CAM++.
- The explanation maps facilitate better understanding of the model's decision-making process by clearly highlighting the critical regions within the images.

4.2 Comprehensive Assessment of Pixel Ablation-CAM

We take inspiration from the Grad-CAM assessment approaches and adapt and expand the methodology from the Ablation-CAM publication to empirically assess the performance of Pixel Ablation-CAM. Our goal is to produce class-conditional localisation maps and evaluate them with several quantitative criteria. We use Pixel Ablation-CAM to create a class-conditional localisation map (heatmap) for each image I, which we then compare to maps produced by Grad-CAM and Ablation-CAM. The most significant discriminative areas of the image are highlighted in these heatmaps and are usually indicated in red. Tables 1 and 2 demonstrate that, on this criteria, Pixel Ablation-CAM outperforms Grad-CAM and Ablation-CAM due to a smaller output score loss by the pre-trained VGG-16 and Resnet-50 models on ImageNet[24].

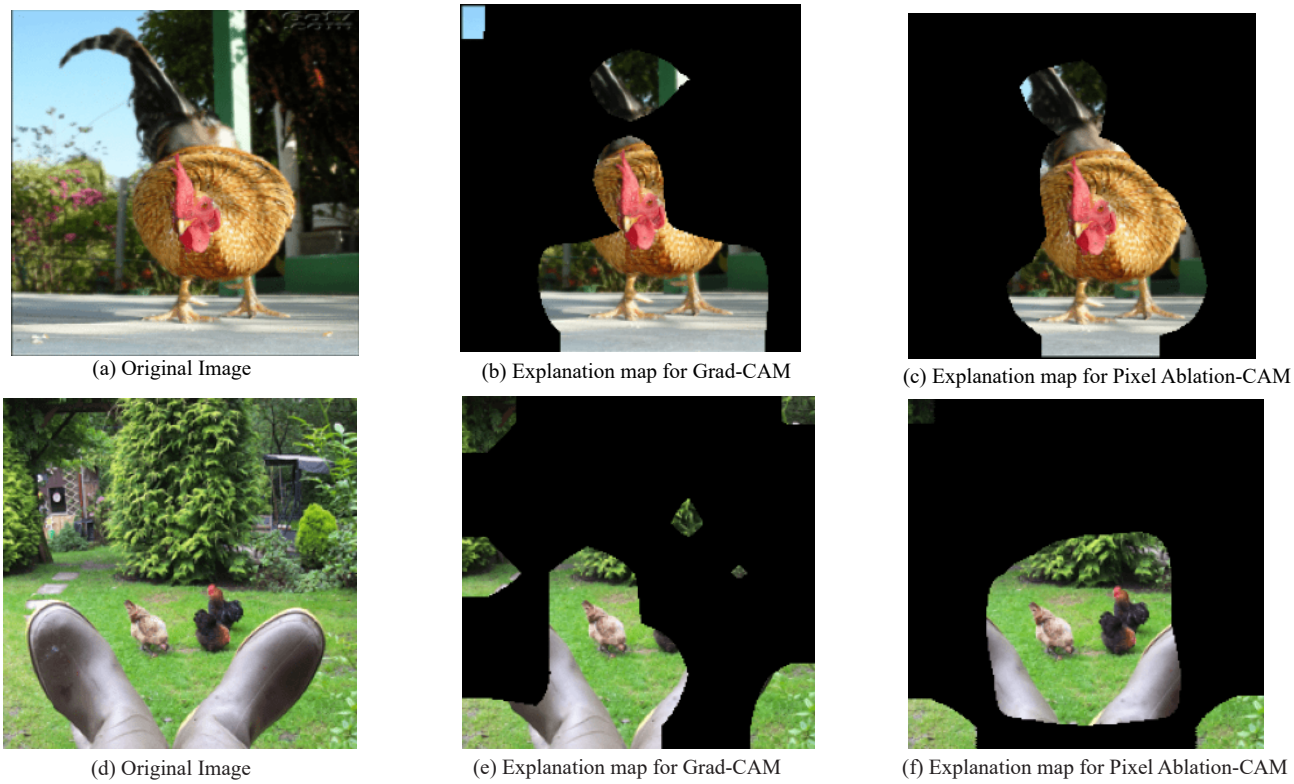
Next, we create explanation maps by retaining the top 20 % of pixels from the localisation map, setting these pixels to 1 and all other pixels to 0. This thresholding ensures a fair comparison between methods based on the number of highlighted pixels. The explanation map is generated by performing a point-wise multiplication of the original image *I* with the thresholded localisation map, retaining only the most relevant regions emphasized by the visualisation technique. We utilize the following metrics to assess the effectiveness of the explanation maps generated by Pixel Ablation-CAM: average

Table 1. Results for Resnet-50 on ImageNet 2012 validation data[24]

Metric	Grad-CAM	Ablation-CAM	Pixel Ablation-CAM
Average % drop in confidence (lower is better)	47.03	46.85	32.64
Average % in activation (lower is better)	33.87	33.58	24.82
Percent increase in confidence (higher is better)	12.70	14.26	23.36
Percent increase in activation (higher is better)	8.86	8.96	17.98

Table 2. Results for VGG-16 on ImageNet 2012 validation data[24]

Metric	Grad-CAM	Ablation-CAM	Pixel Ablation-CAM
Average % drop in confidence (lower is better)	49.09	46.99	33.43
Average % in activation (lower is better)	35.77	33.78	25.76
Percent increase in confidence (higher is better)	15.92	17.10	26.43
Percent increase in activation (higher is better)	9.69	11.71	18.45



(a) Original Image (b) Explanation map for Grad-CAM (c) Explanation map for Pixel Ablation-CAM

(d) Original Image (e) Explanation map for Grad-CAM (f) Explanation map for Pixel Ablation-CAM

Figure 3. Success of Pixel Ablation-CAM: (a) Original image. (b) Explanation map for Grad-CAM, (c) Explanation map of proposed Pixel Ablation-CAM, (d) Original image with multiple instances of subject of interest, (e) Explanation map for Grad-CAM and (f) Explanation map of proposed Pixel Ablation-CAM.

reduction in confidence and activation score, a percentage increase in confidence and activation score, and a percentage of wins in confidence and activation score.

A high-quality explanation map should cover the most relevant parts of the object, resulting in a lower drop in the model's output scores when the explanation map is used as input instead of the full image. This metric is calculated as:

$$\text{Average drop \%}=\frac{1}{N}\sum_{i=1}^{N}\max\left(0,\frac{Y_i^c-O_i^c}{Y_i^c}\right)\times 100 \quad (6)$$

where, Y_i^c is the output score (confidence) for the original image and O_i^c is the score for the explanation map. N is the total number of images. The max function ensures that negative values (where $O_i^c > Y_i^c$) are set to zero.

Another important metric is the percent increase in confidence and activation score, which assesses how often the explanation map increases the model's output scores, especially when the context in the full image acts as noise. This is defined as:

$$\text{Rate of increase in scores}=\left(\frac{1}{N}\sum_{i=1}^{N}1_{Y_i^c<O_i^c}\right)\times 100 \quad (7)$$

where, the function is an indicator that returns 1 in the case that the argument is true and 0 otherwise.

Figure 3 shows the effectiveness of Pixel Ablation-CAM for accurate target identification. The original image with the subject is displayed in Fig. 3(a). Figs. 3(b) and 3(c) depict the explanatory map for Grad-CAM and the suggested technique Pixel Ablation-CAM, respectively. Pixel Ablation-CAM has been found to trace the subject of interest more accurately than Grad-CAM. Comparably, Fig. 3(d) is another set image that features multiple instances of the subject of interest. Figure 3(e) and Fig. 3(f) depict the explanatory map for Grad-CAM and the suggested technique Pixel Ablation-CAM, respectively. It is evident that whereas the suggested Pixel Ablation-CAM successfully localizes all instances of the subject of interest, Grad-CAM only manages to localize one. It should be mentioned that the VGG-16 pre-trained model on the ImageNet dataset was utilized for inference. The experiments conclude that the suggested Pixel Ablation-CAM outperforms the Grad-CAM in terms of highlighting the object of interest.

In order to assess a channel's significance, Ablation-CAM routinely eliminates (ablates) the entire channel from the final convolutional layer. Although this reveals broad contributions at the channel level, it oversimplifies the activity of the network and eliminates fine-grained pixel-level information that are essential for comprehending intricate visuals. In reality, this can result in less precise renderings, particularly when examining thick or detailed objects.

On the other hand, Pixel Ablation-CAM works at the level of individual pixels by methodically ablating every pixel in the final convolutional layer and tracking the resulting change in activation. This approach offers the following significant benefits:

- Granularity: It offers interpretability down to the pixel level, allowing for more granular and localized insights on model choices.
- Comprehensiveness: It recognizes many occurrences of an object or finer structures that channel-based approaches miss by capturing small changes within images.
- Precision: Pixel Ablation-CAM produces heatmaps that are crisper and more accurate, which is important for jobs like autonomous driving or medical imaging where precise item recognition is necessary.

By overcoming the dimensionality reduction inherent in channel-wise ablation, Pixel Ablation-CAM produces more comprehensive visual explanations, enhancing transparency and trust in CNN models.

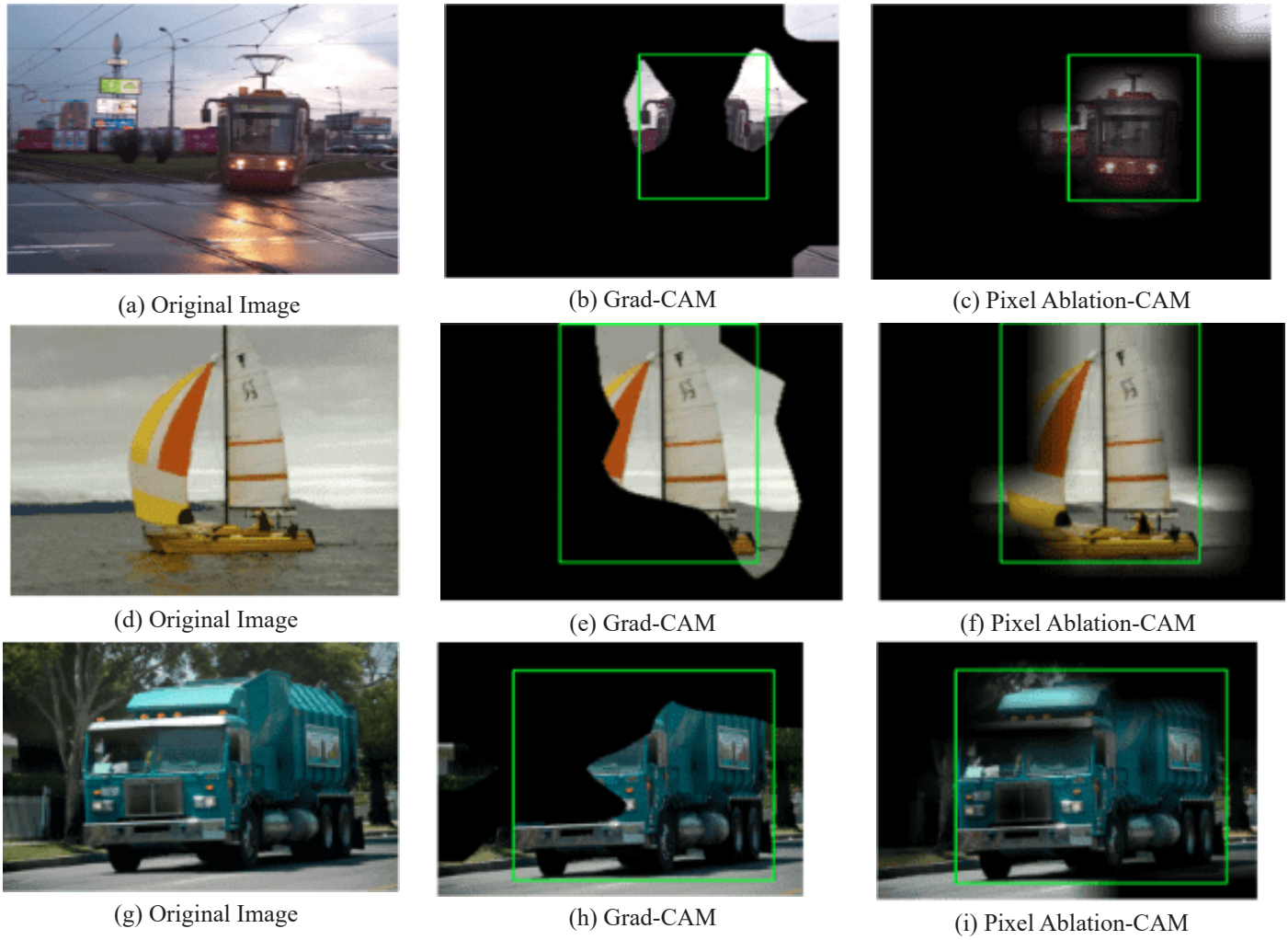
4.3 Subjective Assessment of Pixel Ablation-CAM

In this section, we assess how trustworthy or human-interpretable the suggested Pixel Ablation-CAM explanations are. Using ResNet-50 as the underlying model, we empirically investigated the reliability of the explanations produced by Pixel Ablation-CAM, Ablation-CAM, and Grad-CAM. Five classes with the greatest F1-scores (above 0.94), each from the ImageNet validation set[24], were chosen for this assessment. This selection produced a total of 250 images (50 images per class), ensuring that the underlying model worked well on these categories.

For every image, explanation maps were produced using Grad-CAM, Pixel Ablation-CAM, and Ablation-CAM. Fifteen human volunteers were shown these maps and the associated original images; none of them had any prior information about deep learning or the topic matter. Each image's class title was given to the subjects, who were then asked to choose the explanation map they thought best matched the object in that image. Additionally, if they thought the explanation maps that were generated were similar, they might have chosen "same." The replies were standardized for each image so that the maximum possible score for each image was 1.00. After summing these normalized ratings for each image, a maximum score of 250 might be attained. Based on the results, Pixel Ablation-CAM scored 115.75, whereas Ablation-CAM scored 78.50 and Grad-CAM scored 40.25. The remaining 15.50 were classified by the subjects as "same."

This empirical investigation offers compelling evidence in favour of our hypothesis, which states that the suggested enhancement to Pixel Ablation-CAM facilitates human-interpretable image localisation and hence increases confidence in the model's judgment. Because Pixel Ablation-CAM is based on the principles of both Ablation-CAM and Grad-CAM, it outperformed these techniques in about 49.35% of the cases. Nonetheless, Pixel Ablation-CAM was chosen in most cases, demonstrating its ability to produce more reliable explanations.

Figure 4 shows a comparison between Grad-CAM and the proposed Pixel Ablation-CAM method for object localisation across three different images using explanation map thresholds of 10 %, 30 %, and 50 %, respectively. The original image, Grad-CAM explanation maps, and the Pixel Ablation-CAM explanation maps are displayed in the first, second, and third columns, respectively, and the green boxes represent ground truth annotations that show the actual region of importance in each image. It has been noted that, in comparison to Grad-CAM,



(a) Original Image (b) Grad-CAM (c) Pixel Ablation-CAM

(d) Original Image (e) Grad-CAM (f) Pixel Ablation-CAM

(g) Original Image (h) Grad-CAM (i) Pixel Ablation-CAM

Figure 4. Object localisation capabilities of Grad-CAM and pixel ablation-CAM, with explanation map threshold of 10 %, 30 %, and 50 % successively. The green boxes represent ground truth annotations for the images.

Pixel Ablation-CAM offers more precise and comprehensible localisation of the regions of interest in the images. The Pixel Ablation-CAM maps show a better capacity to identify the true regions of relevance in distinct objects at varying thresholds, as they line more closely with the ground truth bounding boxes.

4.4 Pixel-wise Ablation-CAM for Model Selection

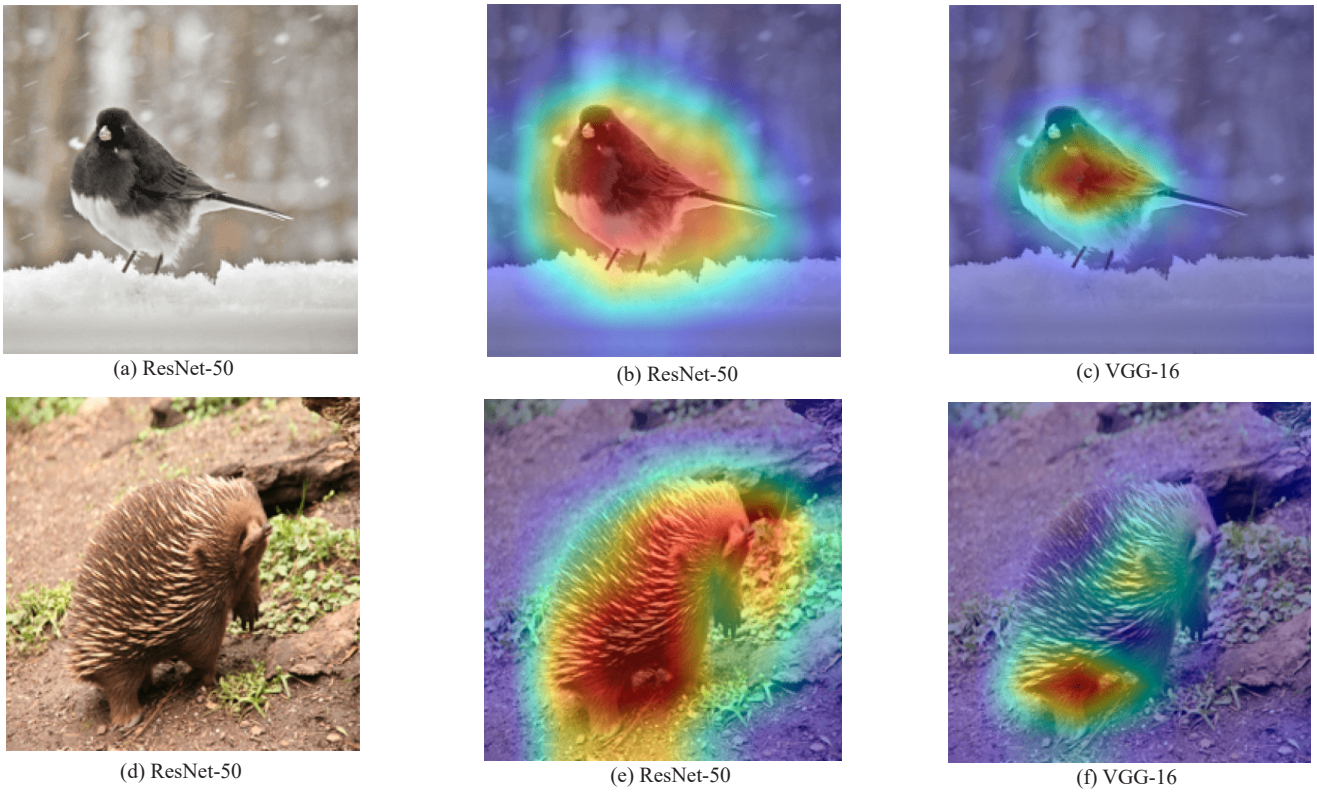
We start by looking at the Pixel Ablation-CAM maps that were created for two different models, and then we compare the maps that were created for different classes using the same model. The purpose of this comparison is to investigate the theory put forward in the paper[8] that better feature maps will be produced by a model with greater generalization capability. In particular, we contrasted the Pixel Ablation-CAM representations of VGG-16 with those of ResNet-50, a model that is recognised for its superior performance, with a top-1% test error of 1.69 as opposed to 2.68 for VGG-16.

The comparison of the suggested Pixel Ablation-CAM visualizations' effects using the Resnet-50 and VGG-16 models is displayed in Fig. 5(a) depicts the original image, and Fig. 5(b) and Fig. 5(c) show the matching heatmap created with Resnet-50 and VGG-16, respectively. In a similar vein, Fig. 5(d) displays the original image, while Fig. 5(e) and Fig. 5(f) display the equivalent heatmap created with VGG-16. For every sample, the subjects were asked to select which of the two explanation maps they believed to be most reliable. ResNet-50 generates a more trustworthy explanation than VGG-16 for the examples provided.

From the ImageNet dataset, we chose the top 1 % of classes with the highest F1 scores for our experiment. We were able to confirm our findings by making direct comparisons on the same collection of images after identifying five common classes among them. We only took into consideration images for which both models produced predictions that agreed with the ground truth to separate the effectiveness of the visualisations from the precision of the models.

Fifteen human participants were requested to score the dependability of the Pixel Ablation-CAM representations for ResNet-50 and VGG-16, without any prior understanding of deep learning or the subject matter. On a scale of +2/-2 one visualization was more/less dependable than the other, +1/-1 one visualization was marginally more/less reliable, and 0 indicated equally reliable, they judged which visualization best described the object in the image. To avoid bias, the visuals were shown at random, with no indication of which model generated which visualization.

According to the data, respondents gave ResNet-50 a score of 1.48, demonstrating its unquestionably higher level of



(a) ResNet-50 (b) ResNet-50 (c) VGG-16

(d) ResNet-50 (e) ResNet-50 (f) VGG-16

Figure 5. Pixel Ablation-CAM visualisations for models ResNet-50 and VGG-16. (a, d) original image, (b, e) heatmap by using Resnet-50, (c, f) heatmap by using VGG-16.

reliability than VGG-16. This result validates our hypothesis by showing that Pixel Ablation-CAM can distinguish the more accurate model based on how well its visuals are rendered. ResNet-50 often outperformed VGG-16 in highlighting pertinent areas of the object in the images. As a result, Pixel Ablation-CAM can support model selection and help users establish confidence in model results.

4.5 Computational Complexity

We mainly use the approach in the last convolutional layer, where the feature map dimensions are down sampled considerably, to assess the effect of pixel-wise ablation on computational resources. With n and m standing for height and width, respectively, a feature map of size $n \times m$ calls for $n \times m$ forward passes per picture. At this point, the computational effort scales linearly yet maintains efficiency due to the smaller size. The decision to concentrate on the final layer for an ideal balance is reinforced by testing pixel-wise ablation at earlier layers, which revealed that larger feature map sizes result in higher computational demands with little additional interpretative benefits.

5. CONCLUSION

To improve the interpretability of CNN-based models, we provide Pixel Ablation-CAM, a newly created pixel-wise ablation technique. In contrast to conventional channel-wise ablation techniques, the suggested method offers a more precise and detailed evaluation of pixel relevance by treating every pixel in the final convolutional layer as a one-dimensional vector over several channels. This change makes it possible to visualize data more finely and precisely, which successfully addresses the drawbacks of earlier techniques like Grad-CAM, Grad-CAM++, and Ablation-CAM.

Our comprehensive tests show that Pixel Ablation-CAM outperforms these techniques in terms of interpretability and localisation, providing more precise and comprehensive visual explanations. Through the resolution of problems with gradient saturation and spatial information loss, Pixel Ablation-CAM offers a strong and dependable instrument for comprehending and relying upon CNN judgments. Subsequent research endeavours will centre around refining the suggested methodology's computational efficacy and investigating its possible integration in increasingly intricate and varied CNN structures and assignments.

REFERENCES

1. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018, **16**(3), 31-57. doi:10.1145/3236386.3241340.
2. LeCun, Y.; Bengio, Y. & Hinton, G. Deep learning. *Nature*, 2015, **521**, 436-444. doi:10.1038/nature14539.
3. He, K.; Zhang, X.; Ren, S. & Sun, J. Deep residual learning for image recognition. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. doi:10.1109/CVPR.2016.90.
4. Ren, S.; He, K.; Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transact. Pattern Analysis and Machine Intell.*, 2017, **39**(6), 1137-1149.

doi:10.1109/TPAMI.2016.2577031.

5. Long, J.; Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3431-3440.
doi:10.1109/CVPR.2015.7298965.
6. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *In* 3rd International Conference on Learning Representations (ICLR), 2015, San Diego, CA, USA, 2015, pp. 1-14.
7. Ribeiro, M.T.; Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the predictions of any classifier. *In* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135-1144.
doi:10.1145/2939672.2939778.
8. Shrikumar, A.; Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *In* Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, **70**, pp. 3145-3153.
9. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D. & Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localisation. *Int. J. Comput. Vision*, 2020, **128**, 336-359.
doi:10.1007/s11263-019-01228-7.
10. Chattopadhyay, A.; Sarkar, A.; Howlader, P. & Balasubramanian, V.N. Grad-CAM++: Generalised gradient-based visual explanations for deep convolutional networks. *In* Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839-847.
doi:10.1109/WACV.2018.00097.
11. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. *CoRR*, 2017. abs/1706.03825.
12. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A. & Torralba, A. Learning deep features for discriminative localisation. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921-2929.
doi:10.1109/CVPR.2016.319.
13. Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. The deep learning workshop: 31st International Conference on Machine Learning, Lille, France, 2015. arXiv:1506.06579.
14. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K-R. & Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 2015, **10**(7), 1-46.
doi:10.1371/journal.pone.0130140.
15. Zeiler, M.D. & Fergus, R. Visualising and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Springer, Cham, 2014, **8689**.
doi.:10.1007/978-3-319-10590-1_53.
16. Lundberg, S.M. & Lee, S.I. A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, **30**, pp. 4768-4777.
17. Zhou, B.; Sun, Y.; Bau, D. & Torralba, A. Interpretable basis decomposition for visual explanation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018, Lecture Notes in Computer Science, Springer, Cham, 2018, **11212**.
doi:10.1007/978-3-030-01237-3_8.
18. Zhang, Q. & Zhu, S. Visual interpretability for deep learning: A survey. *Front. Inf. Technol. Electron. Engin.*, 2018, **19**, 27-39.
doi:10.1631/FITEE.1700808
19. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M. & Kim, B. Sanity checks for saliency maps. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 9525 - 9536.
doi:10.48550/arXiv.1810.03292
20. Simonyan, K.; Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, 2013. abs/1312.6034.
21. Desai, S. & Ramaswamy, H.G. Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localisation. *In* IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 2020, pp. 972-980.
doi: 10.1109/WACV45572.2020.9093360.
22. Morcos, A.S.; Barrett, D.G.T.; Rabinowitz, N.C. & Botvinick, M.M. On the importance of single directions for generalization. *In* 6th International Conference on Learning Representations (ICLR), Vancouver, Canada, 2018.
23. Sundararajan, M.; Taly, A. & Yan, Q. Axiomatic attribution for deep networks. ICML'17: Proceedings of the 34th International Conference on Machine Learning, 2017, **70**, pp. 3319-3328.
24. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C. & Li, F.F. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision*, 2015, **115**, 211-252.
doi: 10.1007/s11263-015-0816-y.
25. Nagahisarchoghaei, M.; Nur, N.; Cummins, L.; Nur, N.; Karimi, M.M.; Nandanwar, S.; Bhattacharyya, S. & Rahimi S. An empirical survey on explainable AI technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics*, 2023, **12**, 1092.
doi:10.3390/electronics12051092
26. Saranya, A. & Subhashini, R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analyt. J.*, 2023, **7**, 1-14.
doi:10.1016/j.dajour.2023.100230
27. Sharma, N.A.; Chand, R.R.; Buksh, Z.; Shawkat Ali,

A.B.M.; Hanif, A. & Beheshti, A. Explainable AI frameworks: Navigating the present challenges and unveiling innovative applications. *Algorithms*, 2024, **17**, 227.
doi:10.3390/a17060227

28. Van Der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.A. & Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Analysis*, 2022, **79**.
doi:10.1016/j.media.2022.102470
29. Reddy, G.P. & Pavan Kumar, Y.V. Explainable AI (XAI): Explained. *In* IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 2023, pp. 1-6.
doi: 10.1109/ESTREAM59056.2023.10134984
30. Pascal, H.; Michael K.; Patricia C. & H. Felix, W. Explanation matters: An experimental study on explainable AI. *Electron. Markets*, 2023, **33**, 17.
doi:/10.1007/s12525-023-00640-9
31. Salama, A; Adly, N. & Torki, M. Ablation-CAM++: Grouped recursive visual explanations for deep convolutional networks. *In* IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 2011-2015.
doi:10.1109/ICIP46576.2022.9897617

CONTRIBUTORS

Mr Akash Samanta is a BTech Computer Science final year student at Techno India University, West Bengal. His research interests include: Machine learning, computer vision, artificial intelligence, explainable AI, and pattern recognition.
In the current study he implement the project, and test the suggested procedure.

Dr Debasis Chaudhuri obtained PhD in image processing and pattern recognition from the Indian Statistical Institute, Kolkata. He is working as a Professor of computer science and engineering at Techno India University in Kolkata. His research interests include: Image processing, pattern recognition, computer vision, remote sensing, machine learning, deep learning, and target detection from satellite image, SAR, thermal, and MMW imageries.
In the current sutdy he assisted the project scientists and developed the algorithm, which led to the current research.

Mr Aniket Kumar Singh obtained BTech in electronics and communication engineering, from the Heritage Institute of Technology, Kolkata, West Bengal, India. He is currently working at Techno India University, Kolkata, as a Project Scientist. His current research interests include: Computer vision, image processing, machine learning, and remote-sensing.
In the current sutdy he was involved in the project's implementation as well as the testing of the recommended process.

Mr Manish Pratap Singh is presently working as Director of DRDO Young Scientist's Laboratory-Cognitive Technologies at Chennai. Manish Pratap Singh did his MTech from IIT Kharagpur in Computer Science and Engineering. His areas of interest include: Pattern recognition, satellite image processing, and cognitive surveillance.
In the current sutdy he worked on the algorithm's overall architecture has benefited the ongoing study.