2024 World Journal of Innovation and Modern Technology, Vol.7, Issue 2 (Apr)
ISSN 2682-5910
DOI: 10.53469/wjimt.2024.07(02).08

Gradient - Based Convolutional Neural Network Feature Visualization

Liangyu He

School of Computer engineering, Weifang University, Weifang, Shandong, China

Abstract: ***This paper proposes a novel method for evaluating the robustness of visualization. In terms of effectiveness, the paper assesses visual coherence, visual resolution, and multi-target visualization. Regarding the robustness of visualization methods, a new evaluation approach is introduced by applying the model's adversarial resistance assessment from image classification to visualization methods. The paper incorporates deconvolution techniques to showcase image features extracted by different convolutional layers of Convolutional Neural Networks (CNN) and employs class activation techniques to visualize critical regions influencing CNN decisions. Additionally, for multi-target input images, a gradient-class activation-based visualization method is proposed and compared across different models, demonstrating its applicability to any model and its superior ability to retain target information compared to other visualization methods. Finally, the paper introduces an evaluation method for visualization effects, focusing on both effectiveness and robustness. In practical applications, this work provides valuable insights for the interpretability of deep learning models and handling complex scenarios.***

Keywords: Convolutional neural network; Robustness; Feature visualization.

1. INTRODUCTION

Visualization of Convolutional Neural Network (CNN) features involves various methods. This paper conducts a visual exploration of the focal points in the image regions that different CNN models emphasize. The study primarily selects methods based on gradient analysis, including backpropagation and gradient-class activation, for feature visualization.

The paper presents the utilization of deconvolution techniques for showcasing image features extracted from various CNN convolutional layers. It also employs class activation techniques to visualize crucial regions influencing CNN decision-making. Additionally, a gradient-class activation visualization method is introduced for multi-target input images, and its results are compared across different models. The paper introduces an evaluation method for visual effects, assessing effectiveness in terms of visual coherence, resolution, and multi-target visualization. Regarding robustness, the paper incorporates a model deception resistance evaluation method from image classification, proposing a novel approach for assessing the robustness of visualization methods.

2. RELATED WORKS

In the realm of convolutional neural networks (CNN), a series of studies have advanced our understanding of robustness and feature visualization. Cao et al. (2020) analyze the noise robustness of deep neural networks [1], while Becker et al. (2020) propose robust dimensionality reduction for data visualization using deep neural networks [2]. Lou et al. (2021) introduce a CNN-based approach to predict network connectedness robustness [3]. Shao et al. (2021) investigate the adversarial robustness of vision transformers [4]. Wang et al. (2020) contribute CNN explainer for learning CNNs with interactive visualization [5]. Tang et al. (2020) present a novel CNN for low-speed structural fault diagnosis [6], and Chen et al. (2021) propose a multiscale CNN for effective bearing fault diagnosis [7]. Wang et al. (2022) introduce an attention-guided joint learning CNN for bearing fault diagnosis and vibration signal denoising [8]. Zhao et al. (2022) present a multiscale inverted residual CNN for intelligent bearing diagnosis [9], and Fang et al. (2021) offer LEFE-Net, a lightweight feature extraction network for bearing fault diagnosis [10]. Dyrba et al. (2021) focus on improving the comprehensibility of 3D CNNs in Alzheimer's disease research [11]. Pandey and Jain (2022) contribute a robust CNN for plant leaf disease identification from smartphone images [12]. Zhang et al. (2021) propose integrated neural networks for underwater target recognition [13]. Ding et al. (2020) explore indexing electron back-scatter diffraction patterns using a CNN [14]. Jang and Tong (2024) enhance human vision modeling in CNNs by incorporating robustness to blur [15].

3. METHODS

3.1 Feature Visualization Based on Deconvolution Technique

Deconvolution is a computation process based on algorithms in mathematics, used to represent the relationship between recorded data and input data, as well as to understand the computational process of algorithms. In the field of image processing, to interpret the features learned by each layer of a convolutional neural network (CNN), it is necessary to understand the intermediate features of the CNN. In response to this, the deconvolution neural network is proposed.

Deconvolution technique can be better understood through mathematical formulas, as shown in the following equation:

$$h = f \times g \tag{1}$$

Here, f represents the original input, g represents a series of convolution operations performed by the convolutional network on this input image, and *h* represents the output feature values.

Deconvolution, the inverse of convolution, reverses the mapping of output feature values to the original image space. This aids in understanding learned features and activated regions in the original image by each unit in the convolutional network. In practical experiments, deconvolution is achieved by transposing the filters in the convolutional neural network. Applying these transposed filters to the output features of each layer yields feature visualization results. The convolution and deconvolution processes are depicted in Figure 1.

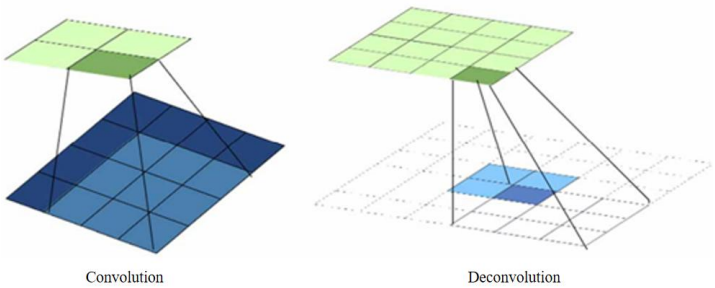

Figure 1: Operation process of convolution and deconvolution

The deconvolution process, depicted in the schematic diagram, is irreversible due to the max-pooling method used in network operations. This method preserves information about maximum values in the feature map, resulting in the loss of data for non-maximum pixel values. During forward propagation, the deconvolution technique records the positions of maximum values in the feature map, aiding in identifying original pixel positions during deconvolution and filling in lost pixel values with zeros.

3.2 Gradient-Activated Feature Visualization

3.2.1 Gradient-Activated Visualization Principle

For a given image with a feature map A^k, the calculation formula for the weights of the image belonging to class C is as follows:

$$W_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial Y^c}{\partial A_{i,j}^k} \tag{2}$$

In this context, Z is a constant representing the number of pixels in the activation map, and Y^c is a differentiable function of feature A^k. In comparison to the previous Class Activation Mapping (CAM) that used global average pooling instead of fully connected layers, this algorithm modifies the weight calculation approach. By utilizing the formula mentioned above to obtain the weights for all feature maps, the final class activation map is obtained by multiplying the weights with the feature maps. The calculation formula is as follows:

$$L^{c}_{Grad-CAM} = \mathrm{Re}\,lu\left(\sum_{k} W^{c}_{k} A^{k}\right) \tag{3}$$

3.3 Feature Visualization Based on Target Selection Gradient

3.3.1 Principles of Target Selection Gradient Visualization

The ultimate goal of visualizing features for interpretability is to highlight crucial characteristics contributing to the network's predictions. Previous class activation maps only focused on the target selection region, neglecting fine-grained propagation. In the Target Selection Gradient (TSGB) algorithm, class activation maps simultaneously meet both the conditions of target selection region and fine-grained propagation, as illustrated in Figure 2, representing the algorithm framework.

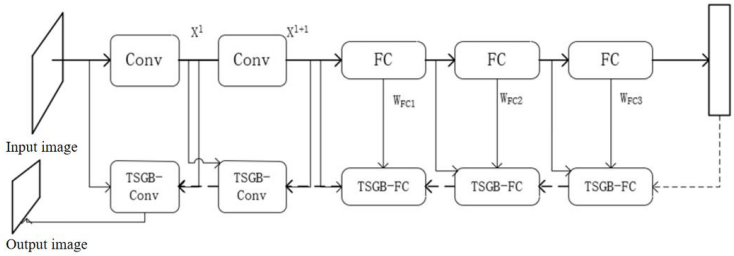

Figure 2: Framework diagram of TSGB algorithm

The CNN visualization comprises two crucial modules: the Target Selection Module and the Fine-grained Propagation Module. The Target Selection Module prioritizes fully connected layers by boosting negative connections based on the positive-to-negative contribution ratio, thereby enhancing focus on the target region. Simultaneously, the Fine-grained Propagation Module incorporates gradient weights and feature maps to generate high-resolution class activation maps.

Target Selection Module: Exclusively applied to fully connected layers within the convolutional neural network, this module modifies the gradients of the last layer of fully connected layers. The standard gradient propagation calculation formula is as follows:

$$g^{l}_{i} = \sum_{j} w_{ij} g^{l+1}_{j} \tag{4}$$

In this context, g^{l}_{i} represents the i-th node in the l-th layer, and g^{l+1}_{i} represents the backpropagation gradient of the j-th node in the (l+1)-th layer. In the corrected gradient computation method, the initial gradient of the target node 'c' in the output layer is set to 1, i.e., $g^{l+1}_{i=c} = 1$. The initial gradients of all other nodes are set to 0. Subsequently, the gradients are computed layer by layer in a top-down manner, following the specific calculation formula outlined below.

$$g^{l}_{i} = \sum_{j} \left(w^{+}_{ij} + E_{j}\left(x^{l}, w\right) w^{-}_{ij}\right) g^{l+1}_{j} \tag{5}$$

In this context, w_{ij} represents the connection weight, $w^{-}_{ij} = w_{ij} - w^{+}_{ij}$ and x^{l}_{i} denote the features of the i-th node in the l-th layer, and "d" represents the enhancement factor. The value of $E_{j}\left(x^{l}, w\right)$ is obtained by the ratio of positive contribution to negative contribution, and its calculation formula is as follows:

$$E_{j}\left(x^{l}, w\right) = \alpha \frac{\sum_{i} x^{l}_{i} w^{+}_{ij}}{\sum_{i} \left| x^{l}_{i} w^{-}_{ij} \right|} \tag{6}$$

Here, α is a constant value used to adjust the enhancement ratio. When the ratio of the enhancement factor is greater than 1, to avoid excessive suppression of the foreground image, this constant is employed to adjust the magnitude of the enhancement factor.

Fine-Grained Propagation Module: Tailored for the convolutional layers within the Convolutional Neural Network (CNN) model, this module computes the gradient of the lower layer using the following formula:

$$g_i^l = sign\left(x_i^l\right)\sum_j\left[x_i^{l+1}u_{ij} \Big/ \sum_i\left|x_i^l u_{ij}\right|\right]g_i^{l+1} \tag{7}$$

If x_i^l is within the receptive field of x_i^{l+1}, then $u_{ij}=1$ is considered; otherwise, it is zero. The feature map of x_j^{l+1} is determined by model parameters. Transforming the above formula into matrix operations to obtain gradients in the lower layers of a convolutional neural network, let $X^1 \in R^{M\times H_l\times W_l}$ represent the feature map of layer "l," $X^{1+1} \in R^{M\times H_l\times W_l}$ represent the feature map of layer "l+1," and $U^{1+1} \in R^{M\times N\times K_n\times K_w}$ represent a set of convolutional kernels of layer "l" with size $K_n \times K_w$. After transformation into matrix form, the gradient calculation formula is as follows:

$$G_m^l = \left[X^{l+1}\bullet G^{l+1} \Big/ \left|X^l\right| * U^l\right] * \left(U_m^l\right)^T \bullet sign\left(X_m^l\right) \tag{8}$$

3.4 Multi-Target Feature Visualization Improved with Gradient-Class Activation

3.4.1 Principles of Multi-Target Visualization

Most visualization methods focus on displaying specific target regions or fine-grained detection in single-target input images. However, there is limited research on visualization methods for input images containing multiple objects of the same class. It is challenging to display as many targets as possible in the heatmap. To address this issue, a novel gradient-weight calculation method for multi-target images is designed. The weight calculation formula is as follows:

$$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \bullet relu\left(\frac{\partial Y^c}{\partial A_{i,j}^k}\right) \tag{9}$$

In this context, relu refers to the rectified linear unit activation function, and α_{ij}^{kc} is the weighted average coefficient of the pixel-level gradients of class "c" and convolutional layer feature map A^k. The calculation formula is as follows:

$$\alpha_{ij}^{kc} = 1 \Big/ \sum_{l,m}\frac{\partial Y^c}{\partial A_{l,m}^k} \tag{10}$$

This method derives a novel computation approach for the gradient weights concerning a specific class "c" and activation feature map "k." The calculation formula for the classification score of the specific class "c" is given by:

$$Y^c = \sum_k w_k^c \bullet \sum_i\sum_j A_{ij}^k \tag{11}$$

After a series of substitutions, a new weight computation method is obtained, and its equation expression is as follows:

$$w_k^c = \sum_i \sum_j \left[\frac{\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3}} \right] \cdot relu\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{12}$$

The method of obtaining heat maps with Grad-CAM is similar, multiplying weights with feature maps to obtain the final visualized heat map. The calculation formula is as follows (Equation 13).

$$L_{Grad-CAM++}^c = relu\left(w_k^c A_{ij}^k\right) \tag{13}$$

The formula computes high-order derivatives of gradient weights, enabling the visualization method to be applied to input images with multiple similar targets. This allows the heat map to display as many similar targets as possible.

3.5 Feature Visualization Effect Evaluation

3.5.1 Effectiveness

Visual Coherence refers to the saliency map or heat map formed should only focus on the region of interest, suppressing other regions in the input image that are unrelated to the target area. Under this metric, the more information about the target area retained in the heat map or saliency map and the less irrelevant information from other areas, the finer the heat map of the target area, indicating a higher visualization effect.

Visual Discriminability refers to the ability of the visualization method to accurately generate saliency maps or heat maps for specific targets based on different categories. This evaluation criterion corresponds to the characteristics of image classification, measuring whether the visualization method can visualize specific targets and generate information related only to the features of the target area, while suppressing information irrelevant to the target area.

3.5.2 Robustness

The study evaluates the visualization method's robustness by attacking it directly, adding imperceptible perturbations to the input and assessing whether the interpretation results can withstand these attacks to provide accurate explanations. Assume the input image is x, add a perturbation δ to obtain an adversarial image x_{adv}. This adversarial image has almost no visual difference from the original input image, i.e., satisfying $\|\delta\| = \|x_{adv} - x\| \langle\langle \varepsilon$ (ε representing a very small constant), ensuring the visual invariance of the adversarial image and the original input image. Assume the visualization result of the adversarial image is $g\left(x_{adv}\right)$, and the visualization result of the input image is $g\left(x\right)$. The method evaluates visualization effectiveness using mean squared error as the loss function. This ensures alignment between the adversarial sample's visualization and that of the original input image after perturbations, as defined by the following formula:

$$Loss = \lambda_1 \left\| g\left(x_{adv}\right) - g\left(x\right) \right\|^2 + \lambda_2 \left\| f\left(x_{adv}\right) - f\left(x\right) \right\|^2 \tag{14}$$

In this context, λ_1 and λ_2 represent the weight parameters for the visualization interpretation and the original input image, respectively. In this experiment, perturbed images were generated by adding random noise and Gaussian noise to the original input images. This was done to observe whether there are changes in the visualization interpretation of the visualization methods.

4. RESULTS

4.1 Feature Visualization Based on Deconvolution Technique

First, the CNN model is trained. Then, feature visualization analysis is conducted on input images to understand which image features each layer of the CNN extracts. In this experiment, visualizations were performed on both the first and fourth layers. Since the first layer captures low-level features and the fourth layer represents high-level features, selecting these two layers is more representative. The results indicate that low-level layers extract simple features like color and texture, while high-level layers capture more complex features:

Visualization results of the first layer: In this experiment, a convolutional neural network model was trained on the PASCAL VOC2012 dataset using the TensorFlow framework, and the results of the visualization of the first layer features are shown in Figure 3.

Figure 3: Visualization results of the first layer features

From the graph, it can be observed that each convolutional kernel in the convolutional neural network extracts different features. Visualizing the features from the first layer in Figure 3, it is evident that the network captures relatively simple image features. These features predominantly outline the contours of airplanes, as seen in the visualization of the wing contours of an airplane in the fifth row and third column.

Figure 4 shows the fifth activation maps for ten images per class after passing through the first convolutional layer of a neural network. These maps highlight contour and shape features, retaining much of the image's original information, indicating strong capability in detecting basic object attributes.

Figure 4: Visualization results of different images in the fifth characteristic map of the first floor

The fifth activation value of the first layer in the convolutional neural network, in addition to activating color features, roughly outlines and shapes the main objects for certain images. For example, in the tenth row and tenth

column, the contours of the characters and the dog are clearly visible. In the visualization result for airplanes in the second row, most of the external contour features of the airplanes are evident, especially the wing contours. In the image at the eighth column of the second row in Figure 4, the overall outline of the airplane is distinctly visible. This indicates that the first layer of the convolutional neural network extracts relatively simple image features.

The visualization result for the fourth layer is shown in Figure 5. It is evident from the image that the features extracted at this layer are markedly different from those in Figure 3. In the visualization result for this layer, the contour features of the target objects are almost invisible.

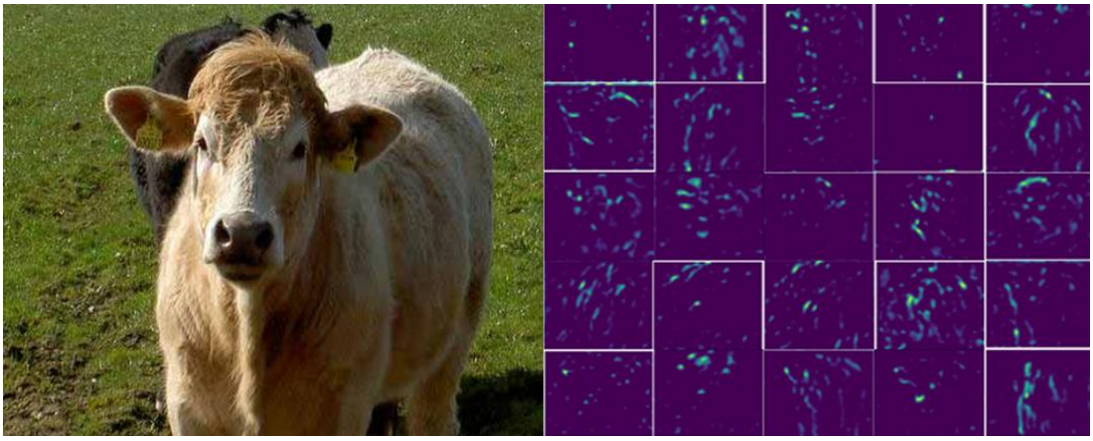

Figure 5: Visualization results of the fourth layer feature map

Experimental results show that as convolutional neural networks (CNNs) deepen, extracted image features become progressively more complex, resembling human brain functionality. Like human cognition, these networks initially focus on simple information, such as object contours, colors, and textures, before delving into more detailed features. Shallow convolutional layers primarily extract basic image features, while deeper layers can capture intricate details, including nuanced aspects of animals. Opting for a deeper CNN model structure is thus more suitable for tackling complex computer vision tasks.

4.2 Gradient-Activated Feature Visualization

This algorithm, an enhancement of CAM, effectively emphasizes features highlighted by the convolutional neural network for each image class through gradient-class activation. By utilizing the ReLU function in the final weighted sum, the model focuses solely on class-relevant regions, eliminating irrelevant background areas in the activation map and improving visualization accuracy. The interpretative results are demonstrated in Figure 6:

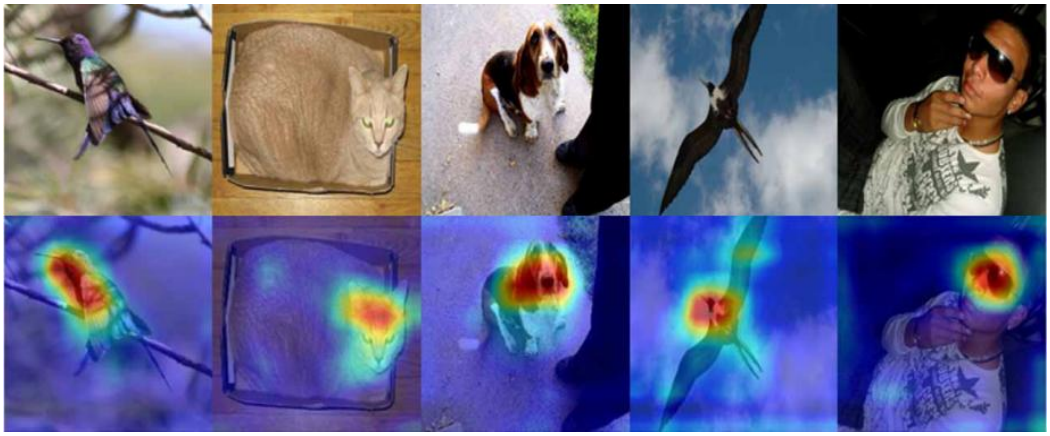

Figure 6: Visualization results based on gradient-class activation

According to Figure 6, the Grad-CAM visualization algorithm effectively highlights the convolutional neural network's focus on the head features of animals in the images. This indicates that head features are crucial information for the network to distinguish between animal categories. Experimental results demonstrate that in the

visualizations of images from different categories, the highlighted areas primarily concentrate on the head features of animals, such as the beak and wings for birds, the head and ears for cats, the mouth for dogs, and the glasses and nose for humans. However, it is worth noting that the algorithm sometimes encompasses broader regions, potentially including irrelevant background information.

4.3 Feature Visualization Based on Target Selection Gradient

The algorithm comprises two modules: target selection and fine-grained propagation. Through a single backward propagation process, it simultaneously generates saliency maps for target selectivity and fine-grained details. By addressing negative contributions, like background image weights, it refines fully connected layer gradients for precise target region selection. The fine-grained propagation module enhances visual interpretation by leveraging feature response ratios between consecutive layers. Visualization results using TSGB are illustrated in Figure 7:

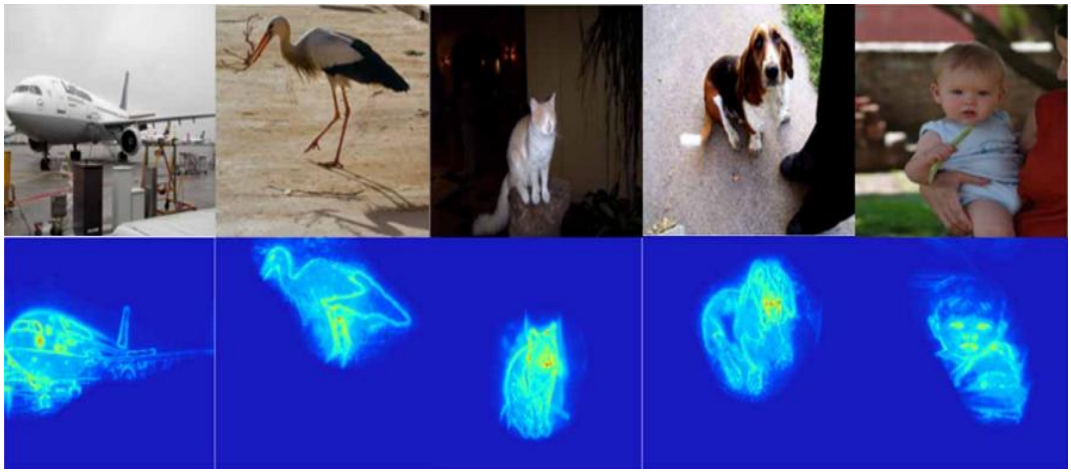

Figure 7: Visualization Results Based on TSGB

Figure 7 shows succinct TSGB visualization results, accurately emphasizing the target region while minimizing irrelevant background details. Notably, it highlights key features like the baby's nose and eyes in the second row and fifth column. In contrast to Grad-CAM, TSGB focuses solely on the baby, offering more detailed and relevant information about the target region.

The algorithm has limitations, particularly in scenarios involving occluded or cluttered background images, resulting in potential information loss in the target region. Furthermore, its visualization efficacy is constrained when dealing with images featuring multiple targets of the same class, as it may not encompass all similar targets.

4.4 Multi - Target Feature Visualization Improved with Gradient-Class Activation

This method improves gradient-class activation visualization by incorporating second-order partial derivatives of pixel gradient weights, introducing new weighted coefficients with third-order derivatives. This enhancement is particularly effective for images with multiple similar objects, showcasing superior performance across various models. It retains more target information, suppresses irrelevant background details, and highlights intricate features within the target region. Experimental results suggest the suitability of all visualization methods for VGG models, whereas heatmaps from other models may contain additional irrelevant information.

4.5 Feature Visualization Effect Evaluation

4.5.1 Effectiveness

The evaluation of the performance of various visualization algorithms under the metric of visual coherence is illustrated in Figure 8.

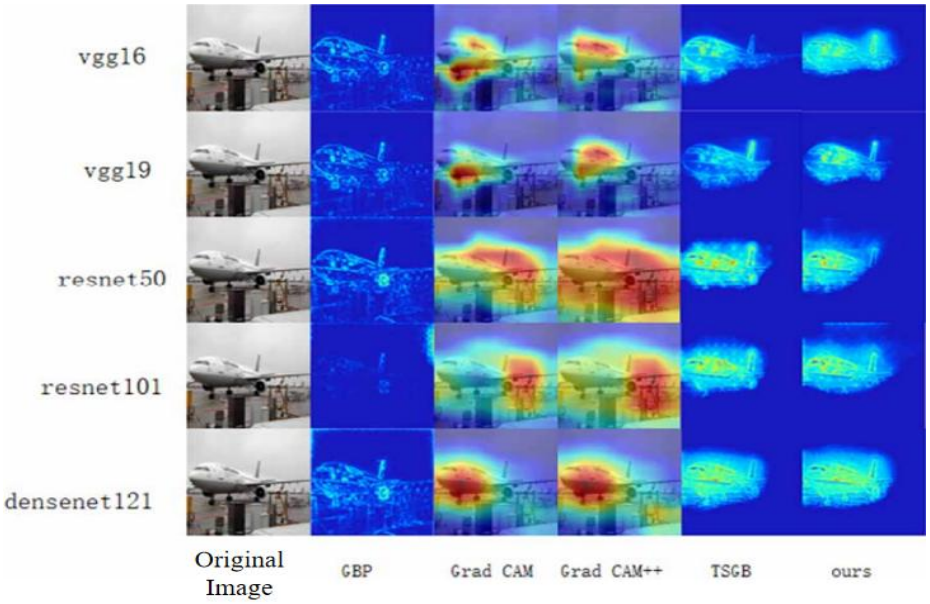

Figure 8: Visualization Effects of Different Visualization Methods on Different Models

Figure 8 illustrates that all visualization techniques predominantly emphasize key features situated on the aircraft's head. Notably, except for Grad-CAM and Grad-CAM++, which operate at the region level, other methods prioritize fine-grained representation of image features. These techniques exhibit minimal activation in irrelevant image regions, showcasing strong visual coherence and minimal inclusion of background information. Additionally, visualization algorithms generally surpass convolutional neural network models, particularly VGG, compared to ResNet and DenseNet.

For this evaluation, five randomly chosen images from different classes in the PASCAL VOC dataset were used with the VGG16 model. Various visualization methods were applied, and Figure 9 showcases their effects, evaluated based on visual discriminability:

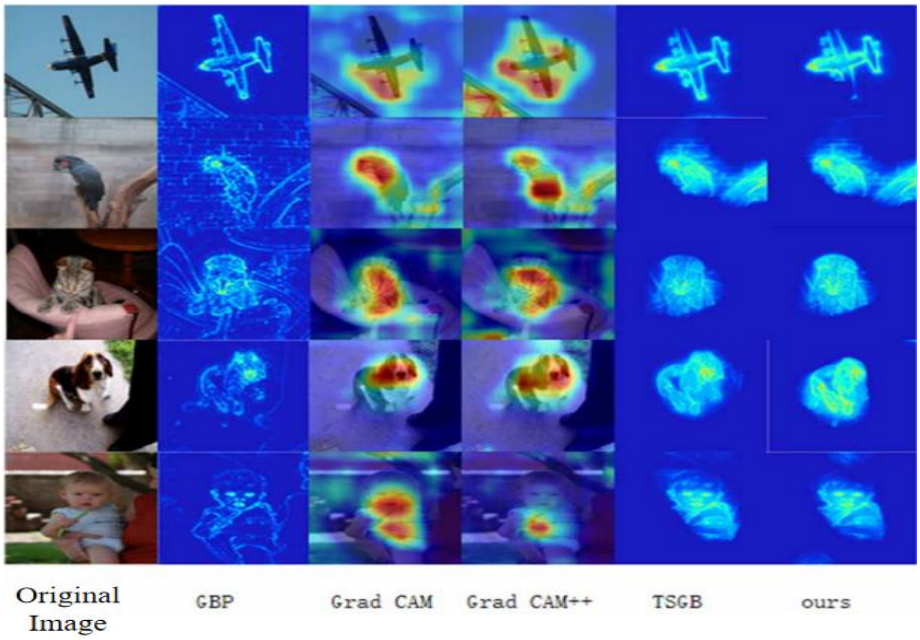

Figure 9: Visual renderings of different classes

Figure 9 demonstrates varied outcomes with different visualization methods across categories. Gradient-class activation effectively highlights targets in birds, cats, and dogs but shows less emphasis on humans and airplanes,

potentially including more irrelevant background details. GBP visualization tends to incorporate extraneous information across all categories, lacking precise focus on targets. Other gradient analysis-based methods effectively concentrate on targets but may partially include background features, like branches in the second row, possibly due to the network treating branches as a feature for bird classification. In interpreting CNN models through visualization, branches are presented as part of the heatmap.

The multi-target visualization experiment evaluates the effectiveness of various visualization methods on images featuring multiple instances of the same category. Using the PASCAL VOC dataset and the VGG16 model, the goal is to optimize localization accuracy and minimize omission.

4.5.2 Anti - Deception

Regarding robustness to adversarial attacks, TSGB and Grad-CAM++ exhibit strong resistance. These methods generate different visual explanations for original input images and adversarial images while maintaining consistent model prediction results. In contrast, Grad-CAM provides similar visual explanations for both original input images and adversarial images, even highlighting the target areas in a nearly identical manner. This indicates that Grad-CAM has weaker robustness, i.e., resistance to adversarial attacks, compared to other visualization methods. However, the experimental results also show consistent visual effects for images of cats and dogs, indicating relatively weak robustness of visualization methods when dealing with complex animal targets.

5. SUMMARY

This paper introduces a method that utilizes deconvolution and class activation techniques for visualizing the image features and critical decision regions extracted by CNNs across different convolutional layers. Additionally, for multi-target input images, a gradient-class activation-based visualization approach is proposed and compared across various models, demonstrating superior effectiveness. The article also provides a comprehensive evaluation method for visual effects, assessing results from the perspectives of both effectiveness and robustness, including visual coherence, resolution, and multi-target visualization. Finally, by incorporating a model's adversarial robustness evaluation method into the visualization assessment, a novel robustness evaluation approach is proposed. These innovative methods contribute to the advancement of research in the interpretability of deep learning models.

REFERENCES

[1] Cao, Kelei, et al. "Analyzing the noise robustness of deep neural networks." IEEE Transactions on Visualization and Computer Graphics 27.7 (2020): 3289-3304.
[2] Becker, Martin, et al. "Robust dimensionality reduction for data visualization with deep neural networks." Graphical Models 108 (2020): 101060.
[3] Lou, Yang, et al. "A convolutional neural network approach to predicting network connectedness robustness." IEEE Transactions on Network Science and Engineering 8.4 (2021): 3209-3219.
[4] Shao, Rulin, et al. "On the adversarial robustness of vision transformers." ar**v preprint ar**v:2103.15670 (2021).
[5] Wang, Zijie J., et al. "CNN explainer: learning convolutional neural networks with interactive visualization." IEEE Transactions on Visualization and Computer Graphics 27.2 (2020): 1396-1406.
[6] Tang, Haihong, et al. "A novel convolutional neural network for low-speed structural fault diagnosis under different operating condition and its understanding via visualization." IEEE Transactions on Instrumentation and Measurement 70 (2020): 1-11.
[7] Chen, Junbin, et al. "Multiscale convolutional neural network with feature alignment for bearing fault diagnosis." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-10.
[8] Wang, Huan, et al. "Attention-guided joint learning CNN with noise robustness for bearing fault diagnosis and vibration signal denoising." ISA transactions 128 (2022): 470-484.
[9] Zhao, Wenlei, et al. "Multiscale inverted residual convolutional neural network for intelligent diagnosis of bearings under variable load condition." Measurement 188 (2022): 110511.
[10] Fang, Hairui, et al. "LEFE-Net: A lightweight efficient feature extraction network with strong robustness for bearing fault diagnosis." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-11.
[11] Dyrba, Martin, et al. "Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease." Alzheimer's research & therapy 13 (2021): 1-18.

[12] Pandey, Akshay, and Kamal Jain. "A robust deep attention dense convolutional neural network for plant leaf disease identification and classification from smart phone captured real world images." Ecological Informatics 70 (2022): 101725.
[13] Zhang, Qi, et al. "Integrated neural networks based on feature fusion for underwater target recognition." Applied Acoustics 182 (2021): 108261.
[14] Ding, Z., E. Pascal, and M. De Graef. "Indexing of electron back-scatter diffraction patterns using a convolutional neural network." Acta Materialia 199 (2020): 370-382.
[15] Jang, Ho**, and Frank Tong. "Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks." Nature Communications 15.1 (2024): 1989.