Dean&Francis
ISSN 2959-6157

A Comparative Visualization Analysis of Neural Network Models Using Grad-CAM

Yuhao Shen[1,*], **Xiyan Huang**[2]

[1]Department of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
[2]School of Computer Science and Technology, Tongji University,Shanghai,China

*Corresponding author: ssyee@hdu.edu.cn

Abstract:

As the area of deep learning is advancing at the speed of light, the issue of model interpretability is now a priority for a better understanding and enhancement of the decision-making processes implemented by sophisticated neural networks. So, as deeper learning models create, it becomes vital to guarantee greater transparency and interpretability, in particular, in applications as medical image analytics, Auto-Driving, and Security Systems. The process of visualization of these decisions is assisted by Grad-CAM which is a powerful visualization tool. The rationale behind this work stems from the growing concern on the interpretability of deep learning models with the hope of systematically evaluating how different models attend to certain regions of images during classification. In this study, the three deep neural network models were Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Swin Transformer, which were used to classify the images with the help of Grad-CAM in visualizing the heatmaps of the important regions of the input images important for the models' decision-making. The conclusively featured experimental outcomes prove that Grad-CAM can help to improve the interpretability of these deep networks irrespective of the used architecture or type. This work also extends Grad-CAM, demonstrating its capability to offer insights of model processes toward achieving better and more interpretable AI models.

Keywords: Grad-CAM; ViTs; model decision; visualization of models; deep learning

1. Introduction

State-of-the-art models such as Deep Neural Networks (DNNs) as of now have greatly revolutionized computer vision by solving challenges such as image classification, object detection & segmentation. However, within the framework of these models, it is possible to achieve a result that is beyond human capabilities; at the same time, their work is often simply incomprehensible. It also becomes the "black box" issue, and this is a problem in such cases as medical diagnosis or an autonomous vehicle, where it is important to know how the model comes to the given conclusion [1]. Convolutional Neural Networks (CNNs) have been seen as the most popular

ISSN 2959-6157

models in these advancements; the VGG as well as the ResNet models. VGG incorporated deeper layers [2] so as to enhance the image classification while ResNet provided with an option to add more layers and get rid of the vanishing gradient problem by engaging shortcut connections. To this end, Vision Transformers (ViTs) are of a relatively new era that employ self-attention mechanisms for capturing local and global dependencies in images [3]. Nonetheless, it was seen that ViTs outperform CNNs in efficiency and accuracy; however, ViTs are much more demanding in terms of data and computations than CNNs and add to the problem of model interpretability. As models get more complicated, and as we build more complicated models, the need for interpretability and accountability only becomes much more paramount, leading to the call for tools that can explain how such models come up with their decisions [4].

However, interpretability challenges have manifested and Grad-CAM (Gradient-weighted Class Activation Mapping) has been developed to respond to those challenges. Grad-CAM produces heatmaps with the areas of the image which are the most significant to the model's prediction; this visualizes the gradients happening in the last convolutional layer to help the analyst understand how the decision was made [5]. However, as is shown below, Grad-CAM has some limitations and is best used in tasks such as classification, detection, and medical imaging. Moreover, it can have different performance results on different architectures, which indicates not only the need for deeper exploration of the ideas. Our study concerns Grad-CAM utilization comparison across different architectures of deep learning, including CNNs, Vision Transformers, and Swin Transformers. Therefore, by comparing Grad-CAM's interaction with these models systematically, we wish to learn more about its advantages and disadvantages with regard to giving visual explanations [6].

In this work, we explore the visualization outputs of Grad-CAM for different models VGG, ResNet, and ViTs. The objectives of the experiments are to investigate how Grad-CAM highlights relevant image parts and to investigate the manner in which these models perceive images. CNNs for instance apply feature extraction in apace hierarchy where the initial layers are responsible for detecting edges and textures while the deeper ones identify the objects. As for Vision Transformers, the images are instead divided into patches, and self-attention cues are utilized; thus, Grad-CAM may recognize dissimilar important areas. In this way, we evaluate Grad-CAM's effectiveness in improving model explainability across diverse architecture types. This paper shows how Grad-CAM works with models of different architectures and the research can identify possible areas for improvement in Grad-CAM visualization [7,8,9]. This work is useful to shed more light on the actual methods of applying Grad-CAM to tasks with more composite input data, including multi-modal or temporal data. Subsequently, the present research contributes to enhancing understanding on the applicability of Grad-CAM on or across different architectures of models.

2. Method

2.1 Network Models

In our study, we focused on three distinct deep learning architectures: Most popular architectures are CNN, Vits, and Swin Transformers (Fig.1). These models were selected to understand and compare the effectiveness of the Grad-CAM technique in different types of the neural networks, while each of them has different strategies for processing the images. In this work, we compare these models to offer a more complete picture of Grad-CAM's versatility and efficacy across different architectural styles.

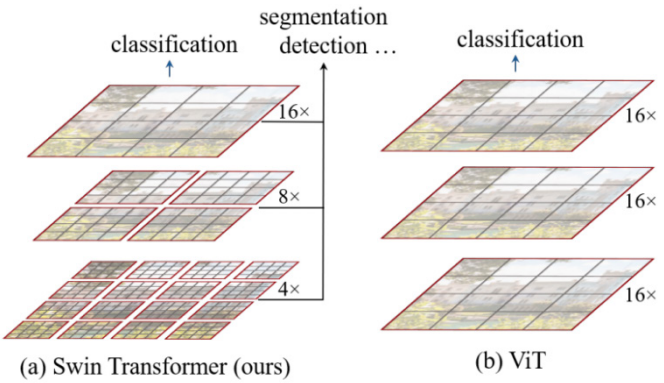

Fig. 1. Swin Transformer architecture [10].

CNNs are regarded as a staple in today's computer vision. They work through a set of filters to input images with a series of convolutions and extract these features at different layers in such a hierarchy. The initial layers detect

simple features such as edges and texture while the later layers detect high level features like shape and objects. CNNs, especially VGG, are basic and efficient; thus, they are suitable in comparing Grad-CAM on visual interpretation [8]. ViTs, thus, are a marked departure from the traditional convolutional way of approaching things. The ViTs employ the transformer architecture that has been used for natural language processing and then apply it to the image input. This architecture splits an image into patches and each patch is transformed into a token and the architecture models the interactions between token using self attention mechanisms. In our work, ViTs were chosen to understand Grad-CAM in architectures that use attention and are not based on convolution operations [4]. Swin Transformers evolutionized the ViT approach to image information processing, thus providing a more sophisticated means of operation. Swin Transformers organise filters hierarchically as well as apply sliding windows that are shifted with the scale which helps the model to gain both the local and global information at scale. This way Swin Transformers can reach very high performance on different vision tasks, even surpassing the CNNs or the basic ViT. We included Swin Transformers into our consideration in order to investigate whether Grad-CAM could produce semantically meaningful visualizations in the setting which encompasses the features of both CNNs and ViTs [9].

2.2 Grad-CAM Theory and Principles

Grad-CAM is an effective method for producing visual explanations of deep neural network predictions in the cases of CNNs. The concept of Grad-CAM is based on the notion of gradient of the target class with respect to the features in the last convolutional layer to determine the most important regions in an input image that has contributed to the final prediction made by the model. The process involves passing an input image through the model to get a forward pass while generating class scores for a particular task, like image classification or captioning. Grad-CAM, for a specific class of interest, sets the gradient of the remaining classes to zero while the gradient for the target class is set to one. These gradients are then passed back through the model, specifically to the convolutional layers, and quantifies the influence of each feature map with regard to the target class.

From mathematical perspective, Grad-CAM calculates a linear combination of the feature maps having weights normalized out of gradients passed back to the feature map (Fig.2). This makes sure that only influential features dominate the space. The obtained class-wise heatmap, that focuses only to the parts of the image most important for the class of interest, represents areas that have significant participation in the model decision-making process, represented by high intensity values. The heatmap is then overlaid on the input image and thus allowed the model to provide a much more easily interpretable output. From the diagram it is clear that Grad-CAM is general and can be used for multiple tasks such as image classification, image captioning, and visual question answering (VQA). It can also thus be generalized to architectures other than CNNs as shown in Its case in models with convolutional and task-specific layers where there are several interplays like with the recurrent neural networks (RNNs). In addition, the Guided Grad-CAM variant employs guided backpropagation for the refinement of the visualization. When Grad-CAM is combined with guided backpropagation the much clearer, more detailed coarse localization map is obtained, which reveal both spatial and concept-specific features improving interpretability of the internal representations of the model.

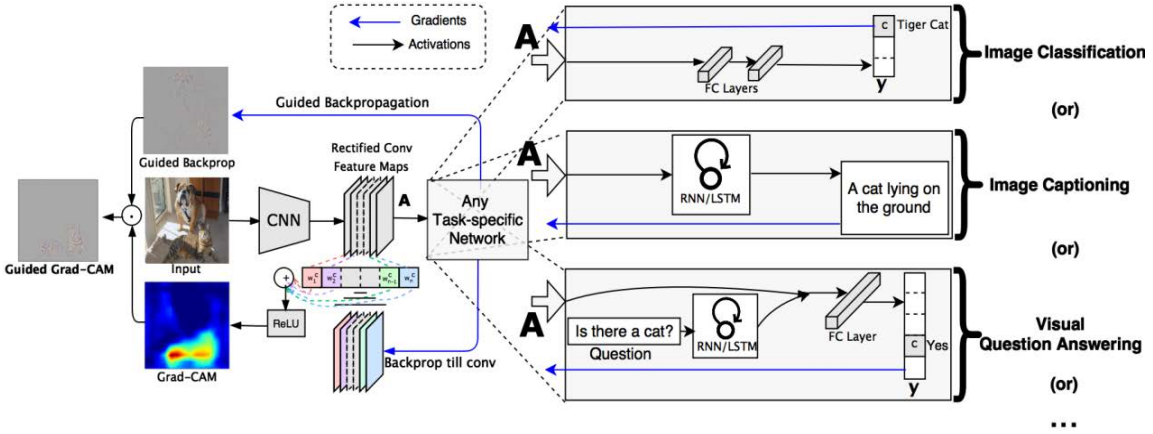

Fig. 2. Grad-CAM overview [11].

ISSN 2959-6157

2.3 Target Layer Selection

This is the reason why Grad-CAM visualizations largely depend on choosing the target layer inside network models. This layer is crucial for defining the resolution and usefulness of heatmaps produced, as their determinants affect the further comprehensibility of model's decision making. Due to the differences in structures between the CNN model and the two transformer models namely ViT and Swin Transformer, we decided to chose certain layers to increase the quality of the Grad-CAM images.

For CNNs, it was decided that the last convolutional layer is the most appropriate for investigation. In Neurons such as in the VGG network that contains multiple of convolutional layers connected with fully connected layers, the final convolutional layer is ideal for Grad-CAM. This layer extracts more general-level characteristics while benefiting from the spatial details, which are necessary to determine which areas of an entry image are involved in a model's decision-making process. Through concentration upon this layer, the heatmaps generated are specific and retain their well-understood semantic value and are thus so beneficial in providing an understanding of the model's interpretative function [12]. In case of ViTs, target layer was set as the output of Transformer Block. ViTs on the other hand work with the images in the form of sequences of patches where every patch is handled like an individual token in NLP. The final appreciation of the image before classification is formed here in the Transformer Block which combines global context and category details from multi head self attention and feed forward layers. This makes it an ideal candidate for Grad-CAM since it gives the interpreter an insight of the model's workings at a time when important decisions are being made [4]. In the Swin Transformer, the final stage product was chosen as the target layer. Swin Transformers use hierarchical structure with shifted windows to extract information from imagery in multiple resolutions. The last stage integrates both the local and the global data, so it is the best stage to select if you need to make Grad-CAM heatmaps. In this way, we focused on this stage so that the heatmaps capture the most important aspects at different scales, providing a detailed picture of the model's decisions [13].

3. Experiments and Results

3.1 Network Settings

MobileNetV3, VGG16, ResNet34, RegNet_Y_800MF, and EfficientNet_B0--which were the CNN models used and were all pretrained on the large-scale ImageNet dataset Grad-CAM was originally developed as a method for visualizing convolutional neural network features. We use its activation maps to form an approximate indication of where an image is being looked at by a trained network Invisalign cards serve another means of bringing MathML formulas together with the compiled preferably human-readable form It's possible to extract the important visualizations from CNN models because their final feature map layer is purpose-designed to operate on photo features. For example, resnet34 has a layer called "layer4" that produces this information and EfficientNet_B0 has a feature set labeled "features". This information is critical for creating meaningful class-discriminative localization maps that highlight those parts of the image most influential to what the model has output. By using such methods of storage standard preprocessing techniques, including image resizing and normalization, can ensure consistency with the models' predicted inputs, leading to reliable visualization outputs.

The ViT differs from CNNs in that it considers an image itself one sequence of patches, pictorially processing each as a token through self-attention machinery to capture long-range dependencies innately. By using the ViT-Base model and a patch size of 16, we could achieve an input resolution in Hue 224x224.Normalization is one of several crucial steps in attention-based feature extraction. With Grad-CAM integration, our focus returns to the last block and layer of `norm1', which has thus been targeted. This ensures better localization through CNN. Since Grad-CAM's undefined jigsaw does not quite fit into the ViT's mental images, we introduced a custom transformation called `ReshapeTransform' to fashion perfectly regular CNN feature maps. This allowed us to generate maps for class-discriminative localization in just a few short lines of code.Finally, we also combined Grad-CAM with the Swin Transformer, which integrates a hierarchical structure and shifted window attention. We employed the Swin-Base architecture and a patch size of 4 with window size of 7. For Grad-CAM we selected the norm layer's size, insuring we would have an accurate feature extraction stage before classification.A similar approach was used to reshape Swin's outputs as for the ViT, enabling us to look at any one point in the model's decision-making process. This combination of CNNs, ViT, and Swin Transformers with Grad-CAM allows us to visualize and interpret the internal workings of different model architectures, demonstrating how each network's attention mechanism affects classification outcomes. The targeted layers for each model play key roles in feature extraction, and the integration with Grad-CAM provides valuable insights into the regions of the input images that most strongly influence the models' predictions.

3.2 Datasets and Evaluation Metrics

To conduct this experiment, we employed the ImageNet-1000 dataset. It is a widely used benchmark in image classification, containing over 1.2 million training images and 50,000 validation images across 1,000 categories. With its comprehensive variety, Grad-cam can offer a model attention view covering multiple classes and fine-grained analysis of such views. All images were preprocessed to ensure consistency, including normalization and size scaling which correspond to protocols used in training of ImageNet-trained models. This guarantees that there is equal treatment of input data between different architectures regarding standard deviation or mean of intensities which could affect our comparisons adversely. We evaluate our Comprehensibility Metric primarily through qualitative means, stressing the lucidity Derived from Grad-CAM Pientudes. By covering the original picture with these heatmaps, we can assess visually whether highlighted areas correspond to a class of interest. Different models each display distinct reasoning processes and this approach in computing Grad-CAM bacterial landmines provides new insights for understanding the thinking behind them.

3.3 Training and Results

3.3.1 Iterative results analysis

The following is a method of processing images using a masked autoencoder (MAE). A Whole image is separated into patches first, a part of which patched with mask. Then in patched out patches are sent through an encoder, and output by the encoder is added to those patched with mask before this combines with other information from the decoder to recreate the original picture. In the course of our pretraining experiment, we achieve a loss of 0.03 after 400 iterations. Even with fewer training iterations than those reported in previous research (the original number was 2000), we still have good results and this indicates to us that the approach works effectively with a smaller number. Fig. 3 and Fig.4 present images of the results of our MAE approach as it processes images through a pretraining experiment. Images from different categories in the ImageNet dataset are shown in each row; this demonstrates MAE ‘s capacity to deal with the task of pretraining images crossing many kinds and types that are probably quite dissimilar to one another. Columns represent the changes in image reconstruction as a function of different numbers of sentences. They show how the model learns to restore the portions of input images blanked out by mask at various stages during training. The images in these rows show how the model learns to reconstruct the parts of the input images that are masked out as training continues but at different times.

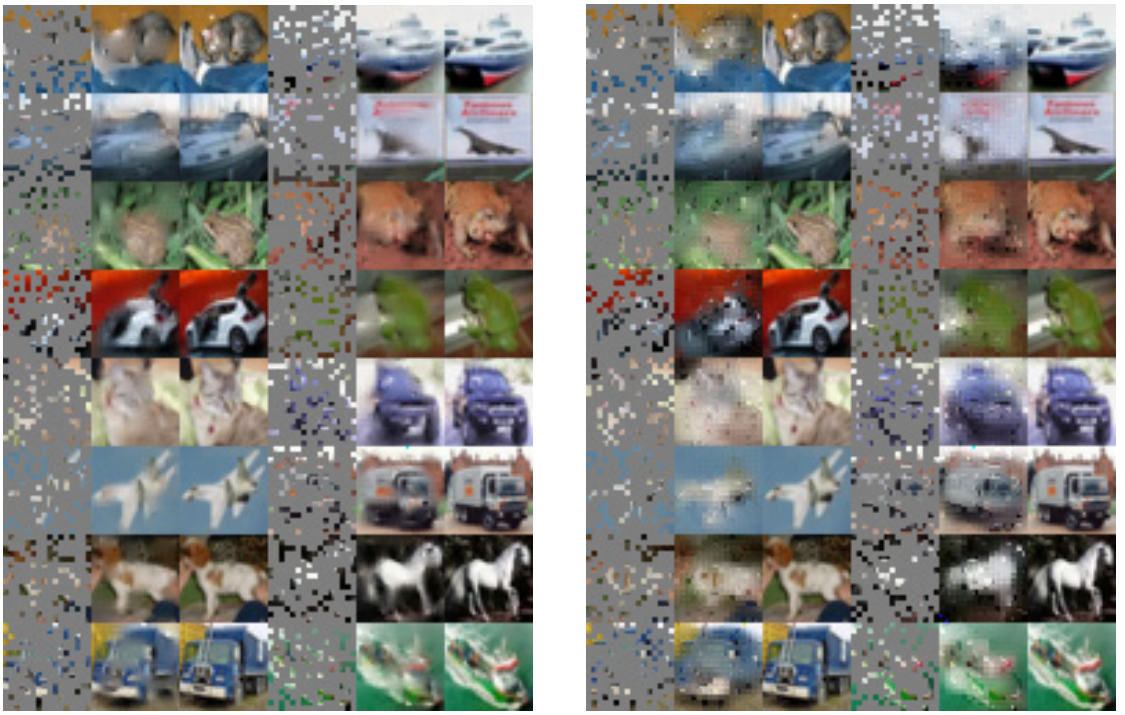

Fig. 3. Effect after 67 and 206 iterations.(Photo/Picture credit : Original)

ISSN 2959-6157

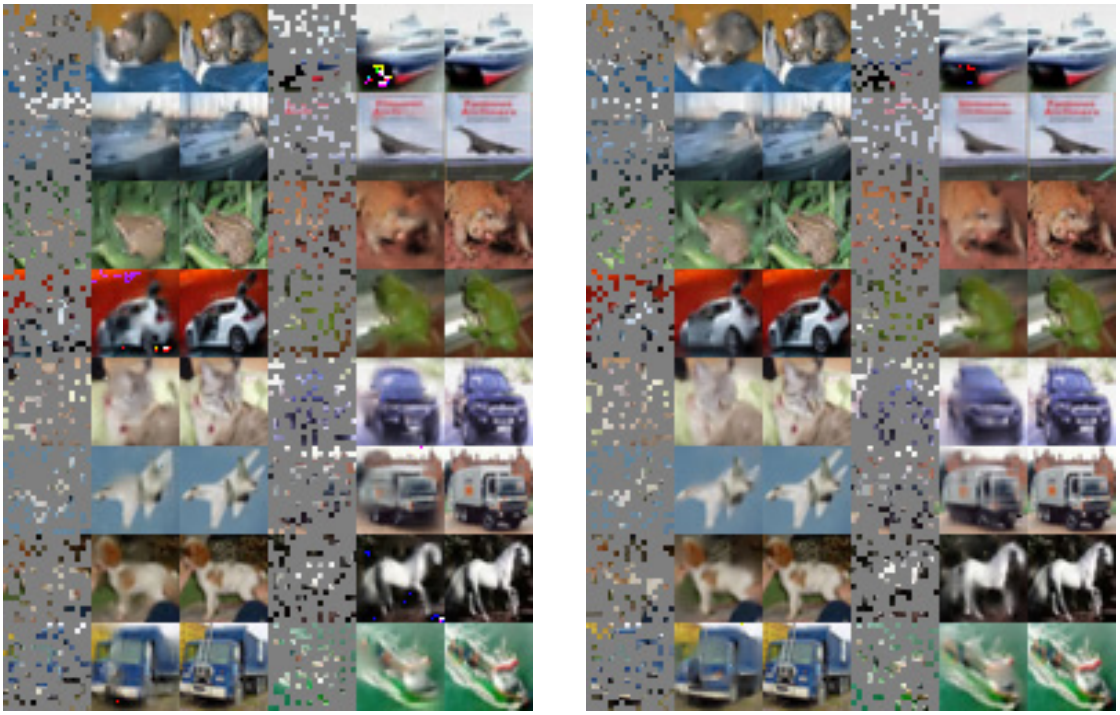

Fig. 4. Effect after 302 and 399 iterations.(Photo/Picture credit : Original)

3.3.2 The impact of masking ratio

We standardized data transformations to guarantee consistency and comparability of results across models. Meanwhile, adjustments were made according to the specifics of each model's parameters. Learning rates, batch sizes and dropout rates are important for optimizing performance while minimizing overfitting. In this regard, hyperparameter search has been performed to find best settings not only once but hundreds of times (Fig.5). For models like ViTs and Swin Transformers, adjustments like LayerNorm fine-tuning or adding dropout were vital to stabilizing training. For instance the former improved the flow of gradients and at the same time reduced reliance on any particular features.

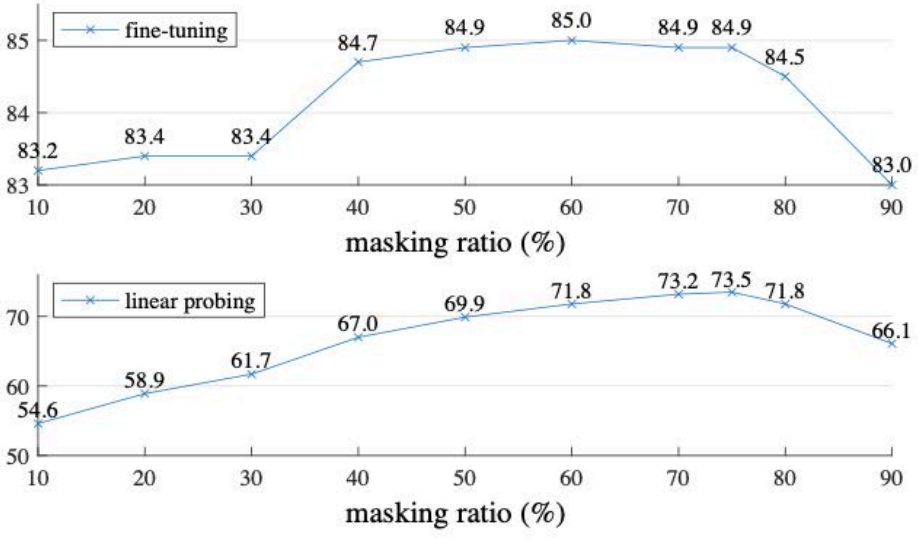

Fig. 5. The impact of masking ratio (Photo/Picture credit : Original).

Thanks to its mask ratio of 0.75, the method was highly successful. On the one hand, because this figure strikes a balance between forcing the model to learn meaningful representations from unmasked patches and leaving

enough information for a successful reconstruction. The higher mask ratio likely encourages the model to focus more on global and structural features rather than localized or trivial details. This abstraction under duress, results in an improvement that helps the model to generalise. It works well in both pretraining stages and in subsequent tasks. During the second round of fine-tuning, adjustments to hyperparameters such as reducing the LayerNorm learning rate from 0.05 to 0.001 and adding a dropout rate of 0.2 further improved performance (Fig.6). The decrease in the learning rate for LayerNorm helped to prevent instability in the normalization process, but particularly in the more depth layers where small changes of normalization parameters can produce large fluctuations on gradients. This sort of stability is particularly important for models like the ViT and Swin, which have attention mechanisms that are highly sensitive to minor variations in input distributions. The introduction of dropout was another regularisation touch. Added at 0.2, it takes the risk out of overfitting by making it more likely that any given neuron or pattern isn't too much relied upon during training.

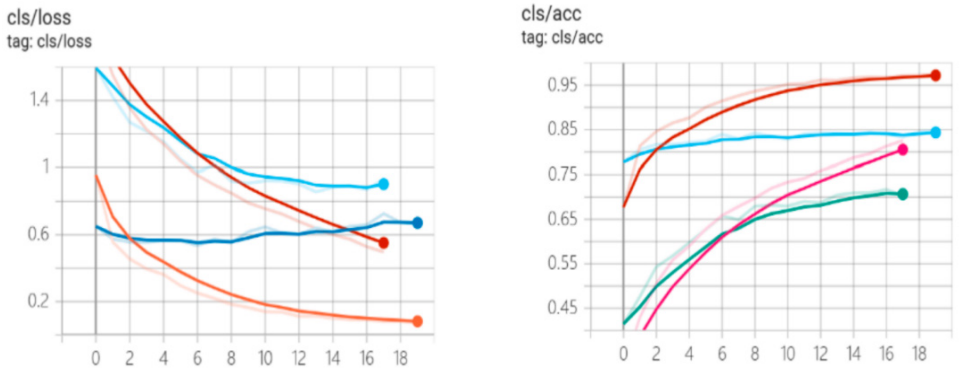

Fig. 6. The impact of Fine Tuning (Photo/Picture credit : Original).

3.3.3 The visualization of models

Three ablation experiments were designed to compare the performance of CNN, ViT, and Swin models in visualization tasks. The first experiment was completed by different models and shows Grad-CAM heatmaps for the class "cat. The second experiment compares heatmaps of both the class "cat" and the class "dog," ViT and ResNet. The third experiment set up this example that we discuss in the next section by using heatmaps based on ResNet have been arranged for many classes including "cat." By analyzing the heatmaps, we found that different models focus on distinct parts of an image depending in this example which class is being depicted at the time. This strong evidence for the interpretability of deep neural networks can be seen from Fig.7, Fig.8, Fig.9 and Fig.10.

Fig. 7. CNN-RESNET34 (Photo/Picture credit : Original).

ISSN 2959-6157

Fig. 8. VGG-16.(Photo/Picture credit : Original)

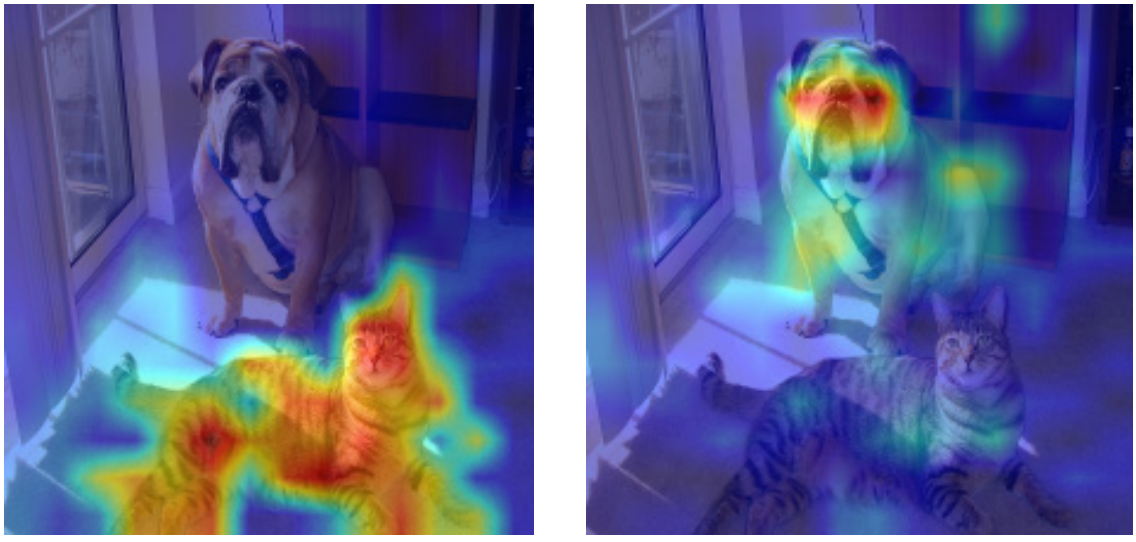

Fig. 9. VGG (Photo/Picture credit : Original)

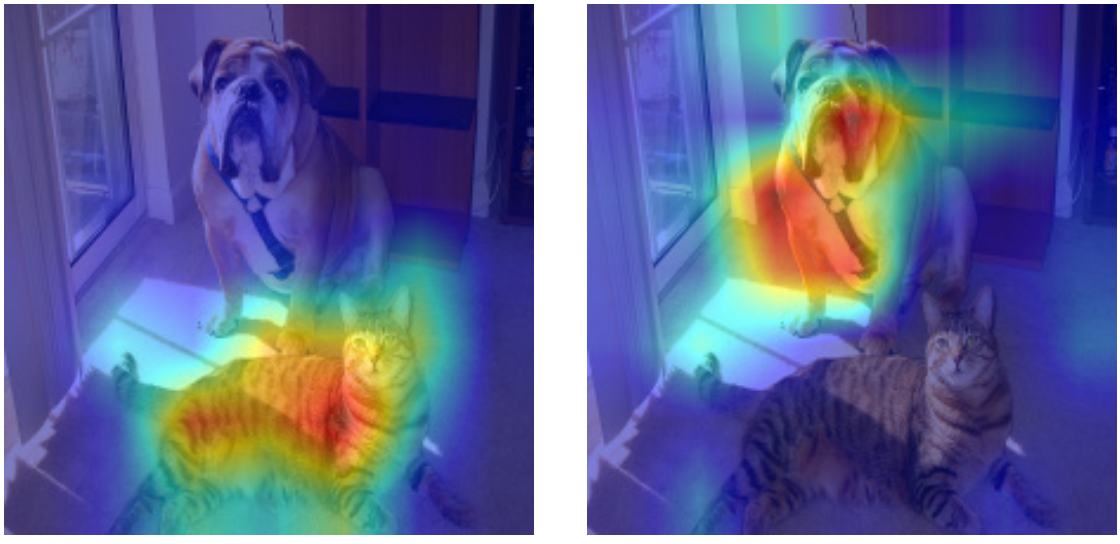

Fig. 10. Swin (Photo/Picture credit : Original)

4. Conclusion

We evaluated three popular deep learning architectures – CNNs, Vits, and Swin Transformers – in detail by employing Grad-CAM to visualize their decision-making processes on image classification tasks. Our experiments showed that using class-discriminative heatmaps to obtain where each model paid attention in an input image, we successfully compared their attention models. We demonstrate through an experimental analysis that Grad-CAM works effectively to make these models easy to interpret. Different attention modes among architectures point up their deep strengths and shallow weaknesses in specific. These findings underscore that Grad-CAM is a useful mechanism for increasing the clarity of deep learning, which is especially necessary in cases if explication brings indispensable benefits. Future work might extend successful demonstration of Grad-CAM onto emerging architectures or fast-expanding research areas such as biomedicine, and highly regulated system classes where trust in AI is key. Further exploration into hybrid models and more adaptive visualization techniques will provide even greater insight into model behavior, fostering better transparency in future AI systems.

Authors Contribution

All the authors contributed equally.

References

[1] Simonyan, K., Vedaldi, A., Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034,2014.

[2] Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations ,2015,1-11

[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 , 2016. 770-778.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N.. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021,1-9

[5] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 618-626.

[6] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , 2016, 2921-2929.

[7] Samek, W., Wiegand, T., Müller, K. R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017.

[8] Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations ,2015, 1-13.

[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 10012-10022.

[10] Kim, W., Son, B., Kim, I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. arXiv preprint arXiv:2102.03334,2021.

[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Proceedings of the IEEE International Conference on Computer Vision, 2017, 618-626.

[12] Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations ,2015

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B.. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision , 2021. 10012-10022.