Evaluation of Feature Attribution Methods in Interpretable Machine Learning

by

Paige Rattenberry

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Bachelor of Applied Science (Honours)

in the
School of Engineering Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2022

APPROVAL

Name: Paige Rattenberry

Degree: Bachelor of Applied Science (Honours)

Title of Thesis: Evaluation of Feature Attribution Methods in Interpretable M

Michael Sjoerdsma

Director, School of Engineering Science

Examining Committee:

Ivan V. Bajic, P.Eng.

Professor, School of Engineering Science

Mirza Faisal Beg, P.Eng.

Professor, School of Engineering Science

Parvaneh Saeedi, P.Eng.

Professor, School of Engineering Science

Date Approved: August 5, 2022

Abstract

The emergence of the next generation of artificial intelligent systems including deep neural networks (DNNs) is occurring rapidly. However, DNNs are currently unable to explain their predictions to humans in an interpretable, transparent, and trustworthy way, and are also susceptible to adversarial attacks, due to their underlying "black-box" nature. Therefore, interpretable machine learning is essential to enable end-users to understand, appropriately trust, and effectively manage these DNNs.

If explainability algorithms can be proven to be sensitive to malicious adversarial attacks and simultaneously capable of generating robust, reproducible, and replicable feature attribution maps that correctly describe a network's predictions, they will be critical in enabling end-users to trust that DNNs deployed in real-world, mission-critical applications and high reliability systems will consistently make successful, safe, and unbiased predictions.

Explainability algorithms, such as feature attribution methods, are themselves, a set of mathematical operations with certain assumptions and, therefore unfortunately, contribute an extra layer of abstraction to the evaluation of their explanations for a network's prediction.

Class activation maps are a category of feature attribution methods which are popular in interpretable machine learning. Class activation maps compute the attribution of each input feature to its importance to the model's prediction, and generate heatmaps to visualize this relationship. However, there is minimal systematic evaluation of feature attribution methods due to a lack of ground truth attribution.

This thesis investigated the success of class activation map methods (CAMs), specifically Grad-CAM, Grad-CAM++, Layer-CAM, Eigen-CAM, and Full-Grad, at giving attribution to the ground truth, for near-perfect deep learning networks trained on medical images, in their attempts to explain a DNN's prediction decisions. This crucial examination was accomplished through a database modification procedure that imposed ground truth to ensure that any accurate and precise DNN should only make the correct prediction classifications if it is relying solely on the introduced input feature perturbations, which successful class activation map methods should highlight exclusively. The CAMs were analyzed both qualitatively and quantitatively through IoU computation. Results demonstrated that Full-Grad appeared to be the most robust, precise and accurate method at localizing discriminatory image features and detecting input perturbations, and that its performance could be optimized by thresholding its output at key threshold values to remove dispersion.

This evaluation will hopefully be an important step in the development and optimization of successful and robust interpretability algorithms, which is essential to gaining user trust and confidence in the use of deep neural networks in real-world, mission-critical applications and high reliability systems, where DNNs that make incorrect predictions could lead to catastrophic outcomes, but also have the potential to make revolutionary breakthroughs.

Acknowledgments

I would like to thank my Academic Supervisor, Dr. Ivan V. Bajić, for his feedback, expertise, time, and support, and for inviting me to be a part of his innovative SFU Multimedia Lab. His experience and extensive domain knowledge has been of immense value, and his suggestions regarding additional experiments to conduct as well as his constant feedback and revision advice was crucial to the development of this thesis.

I would also like to thank my committee members, Dr. Mirza Faisal Beg and Dr. Parvaneh Saeedi for their review and expertise.

It has been an immense honour to have the incredible opportunity to learn from such an experienced and skilled team of professors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Can users trust that deployed deep neural networks will consistently make successful, safe, and unbiased predictions in real-world, mission-critical applications and high reliability systems? To address this fundamental question, it is first necessary to ask if it is possible to explain the predictions of deep neural networks? If so, how can explainability algorithms be evaluated for their robustness, reproducibility, replicability, and sensitivity to malicious adversarial attacks?

This thesis set out to address these critical questions by analyzing the success of class activation map methods, a type of explainability algorithm which attempts to evaluate the predictions of (near) perfectly trained, deep neural networks (DNNs), in terms of distinguishing between "ground truth" classifications based on empirical evidence, as opposed to inferred but possibly inaccurate predictions that the model may be inadvertently making.

The rapid emergence of the next generation of artificial intelligent systems, including complex DNNs, is occuring rapidly. Therefore, DNNs are being deployed in a diverse range of real-world applications. However, DNNs have an underlying "black-box" nature, due to their inherent over-parameterized design and highly nonlinear input-output relationship. Therefore, DNNs are currently unable to explain their predictions to humans in an interpretable, transparent, and trustworthy way, and they are also susceptible to adversarial attacks [1].

These severe limitations are unfortunately hindering the acceptance of DNNs in mission-critical application areas and high reliability systems, where unexpected failures could be catastrophic. For example, in medical clinics, to gain the trust of physicians, regulators, and patients, medical diagnosis systems need to be transparent, understandable, and explainable, through deployed models that make predictions based on genuine medical signals, not image artifacts [2].

Model interpretability may be achieved though explainable artificial intelligence (XAI). Figure 1.1 was presented in 2019 by DARPA, the Defense Advanced Research Projects Agency, in their survey on the history, research areas, approaches and challenges associated with Explainable AI. As demonstrated, XAI is essential to enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems.

Interpretable machine learning methods, specifically interpretable models and post hoc explanations, attempt to explain the underlying decision-making process of DNNs. Post hoc explanations

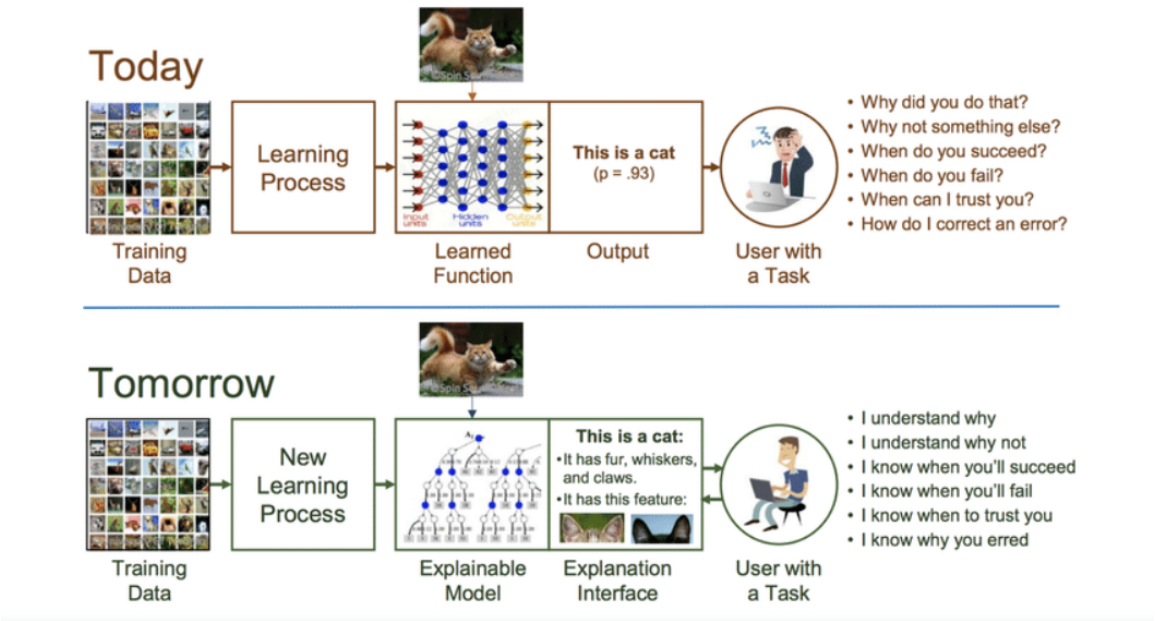

Figure 1.1: Explainable Artificial Intelligence (XAI) Concept Presented by DARPA [3]

either monitor the effect of perturbed input features on a trained DNN's output, or they back-propagate a signal from the network output to the input. This information is displayed as class activation maps that estimate the contribution of input features to the DNN's output [1].

However, explanations are difficult to assess and compare quantitatively due to the lack of ground-truth, and visual analysis of the explanation's "plausibility" has not been demonstrated to be a reliable evaluation metric [1].

Ideally, the attribution maps generated by explainability algorithms must be robust, "sensitive enough to detect adversarial attack and concurrently invariant to small perturbations in the input" [1].

1.1 Explainability Concepts

Figure 1.2 displays a visual breakdown of explainability concepts [1]. It emphasizes that interpretability, replicability, and reproducability are fundamental to explainability.

In particular, requirements of successful explainability methods include sensitivity (model output must be decomposable as the sum of the contribution of individual input features), saturation (constant function output despite input feature saturation to the model due to the non-linear activation function), implementation invariance (consistent attribution scores for functionally equivalent models on the same inputs, regardless of model implementation), input invariance (insensitivity to input transformations), and fidelity (selective identification of relevant input features). Explainability can be achieved through both Post hoc Explainability Methods, as well as Inherently Interpretable models [1].

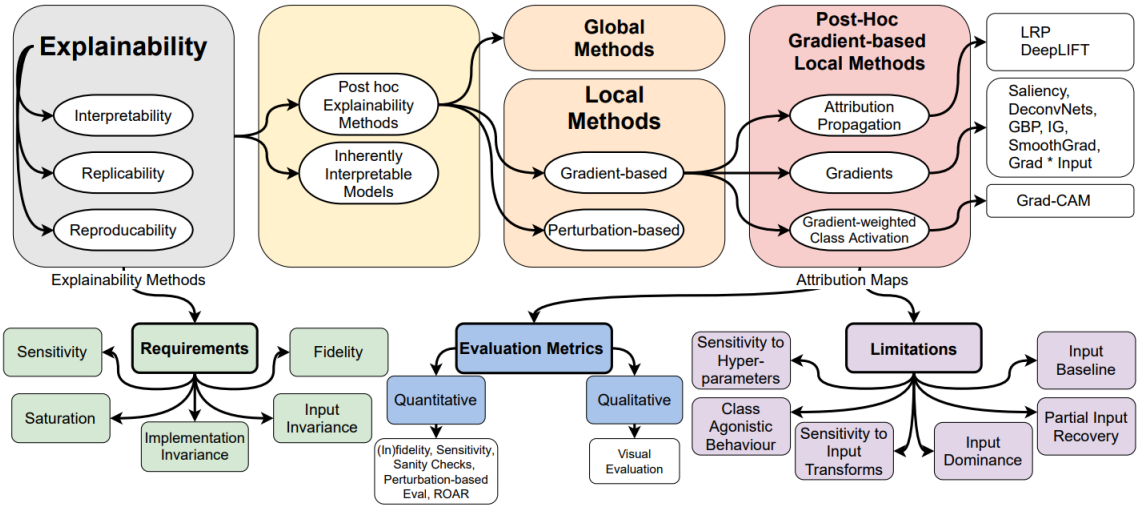

Figure 1.2: Breakdown of Explainability Concepts [1]

Post hoc explanations are mathematical frameworks that attempt to explain the behaviour of "black-box" models to audiences in a variety of use cases, and can be applied both globally or locally. Global methods set out to explain the model's overall decision-making process, whereas local methods focus on explaining specific model decisions, such as input features of most influence to the model's output. The two main categories of post hoc local interpretability methods are either Perturbation-based or Gradient-based [1].

Perturbation-based methods mask or alter input features and calculate the effect of the perturbation on the model's performance. This is computationally intensive as it requires several passes through the model to determine pixel importance. Gradient-based methods backpropagate the gradients of the output, such as logits or soft-max probabilities, with respect to extracted input features, and estimate attribution scores. However, gradients are usually noisy, so contribution from irrelevant features may be seen in the attribution maps produced, making it an ongoing difficulty to measure the validity of gradient-based methods [1].

Finally, gradient-based methods can be further broken down into Attribution Propagation, Gradients, and Gradient-Weighted Class Activation Methods, which are implemented in several different ways. However, as visualized in Figure 1.2, all attribution map methods have limitations, which can include sensitivity to hyper-parameters, class agnostic behaviour, sensitivity to input transforms, input dominance, partial input recovery, and input baseline. Therefore, both qualitative evaluation (thorough visual inspection and analysis of the generated heatmaps), as well as quantitative evaluation (through metrics such as (In)fidelity, Sensitivity, Sanity Checks, Perturbation-based Evaluation, and ROAR (RemOve And Retrain)), are critical to evaluation of gradient-based methods [1].

In conclusion, the rapidly evolving, revolutionary potential of deep neural networks across a diverse range of application domains, along with the associated criticality that Post hoc Explainability Methods can accurately interpret and explain the predictions of these "black-box" models, sparked my thesis inspiration, and drove the experiments and evaluation conducted.

1.2 Thesis Significance and Preview

This thesis attempted to evaluate the success of feature attribution methods at localizing and giving attribution to the ground truth when explaining the predictions made by deep neural networks in medical image classification. It strove to gain further insight and intuition regarding the performance of these explainability methods.

Through extensive qualitative and quantitative analysis of experimental results, along with recommendations for future work, the objective of this thesis was to help answer the fundamental question of how explainability algorithms can be evaluated for their robustness, reproducibility, replicability, and sensitivity to malicious adversarial attacks.

Chapter 2 investigates several state-of-the-art feature attribution methods, specifically Grad-CAM, Grad-CAM++, Layer-CAM, Eigen-CAM, and Full-Grad. It explores how each of these methods are implemented, their fundamental and unique characteristics, and their successes and limitations seen through other studies.

Chapter 3 explains the dataset modification procedure that imposed ground truth on the COVID-19 Radiography database by reassigning labels for binary classification, where only images manipulated with a watermark perturbation were assigned to the COVID-19 positive class. It also explains that the fundamental purpose of the dataset modification procedure was to ensure that any neural network model that achieved high validation accuracy would have only made the correct prediction classification if it was concentrating exclusively on the imposed watermark perturbations, which successful class activation maps should exclusively highlight.

Furthermore, Chapter 3 also describes the process and parameters implemented to train the ResNet-34 and VGG-16 neural network architectures. It then goes on to describe how the gray-scale class activation maps and corresponding colourful heatmap were generated. Finally, it explains the thresholding procedure used in the computation of Intersection over Union for quantitative CAM analysis at differing thresholds.

Chapter 4 analyses the experimental results seen by each CAM method when explaining the predictions made by both models. It also discusses which methods demonstrated to be the most robust, precise and accurate at localizing discriminatory image features and detecting input perturbations, which are fundamental requirements of all explainability methods. Through examination of the thresholded IoU scores, Chapter 4 also analyzes if thresholding CAM output can help to optimize the performance of CAM methods by removing output dispersion.

Finally, Chapter 5 concludes with a description of the significance of explainability methods, especially in automated medical image interpretation. It also discusses the key experimental findings, lays out a road map for future potential explainability tests, and demonstrates the contributions this thesis makes to the field of computer vision and artificial intelligence.

The research involved and experimental results of this thesis will hopefully serve as an important step towards developing successful interpretability algorithms, which would be essential to gaining user trust and confidence in the use of deep neural networks in real-world, mission-critical

applications and high reliability systems where DNNs have the potential to make revolutionary breakthroughs.

Chapter 2

Feature Attribution Methods

Feature attribution methods compute the importance of input features to a model's prediction and performance, by taking the absolute values of the attribution scores they assign to the input features. These assigned attribution scores are used to generate a corresponding heatmap that provides a visual explanation of these scores and emphasize the input features of most significance to the network's prediction, which are illustrated by the highest colour intensity.

Feature attribution methods describe the mathematical properties of a model's decision function, and these mathematical properties are associated with high-level interpretations. However, justification of these associations is crucial, because positive results from alignment evaluation do not support model faithfulness to these input features, only plausibility, as it is impossible to know the reasoning mechanisms used by the model to reach its predictions [4].

Class activation maps (CAMs) can be used for the interpretation of neural networks, because they attempt to explain what regions of the input image were most influential to the trained model's classification prediction. CAMs do this by calculating/assigning a number to each pixel which represents the contribution it played in influencing the model's decision. Contribution has been previously defined as sensitivity, relevance, local, Shapley values, or filter activations, in various studies [4].

These assigned pixel activation scores are then used to generate a gray-scale representation that is the same size as the original image, and can be overlaid onto the original image to produce a heatmap. Regions of the heatmap that are brightest red represent pixels of greatest importance to the model, and pixel significance drops gradually with decreasing colour intensity, with blue/purple regions representing pixels of least importance. The class activation map methods that were evaluated in my experiments were Grad-CAM, Grad-CAM++, Layer-CAM, Eigen-CAM, and FullGrad.

2.1 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [5] weights the 2D activations by the average gradient of any target concept, flowing into the final convolutional layer of the model. It derives weights of linear combination of different feature maps, based on backpropagated class

relevance score. Grad-CAM produces a coarse localization map highlighting important regions in the image for predicting the target concept, and is applicable to a wide variety of CNN model-families without any architectural changes or re-training. Grad-CAM was the first method capable of this, so had been known to be the state-of-the-art saliency map technique.

Figure 2.1 [5] shows the structure of Grad-CAM. The input is an image and class of interest. The image is forward propagated through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class, which is set to 1. Next, this signal is backpropagated to the rectified convolutional feature maps of interest, which are combined to compute the coarse Grad-CAM localization. This heatmap represents where the model must look to make the classification decision. Finally, the heatmap is pointwise multiplied with guided backpropagation to obtain Guided Grad-CAM visualizations, which are both high-resolution and concept-specific.

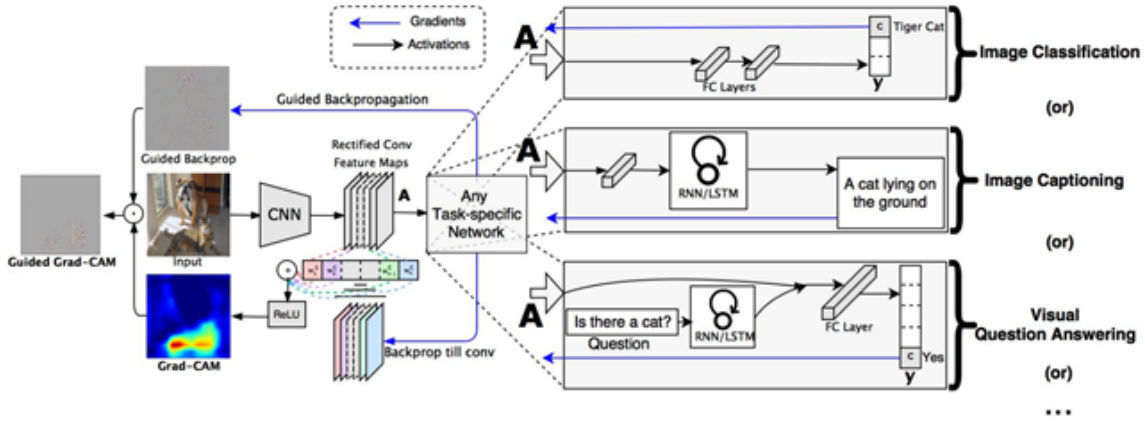

Figure 2.1: GradCam Structure [5]

Experiments have shown that for image classification models, Grad-CAM visualizations lend insights into the model's failure modes, are robust to adversarial images, outperform previous methods on localization, are more faithful to the underlying model, and help achieve generalization by identifying dataset bias [5].

However, [5] demonstrated that a limitation of Grad-CAM is that it is class-discriminative due to its backpropagation of gradients, so Grad-CAM requires the model to make the correct classification decision, in order for it to generate meaningful saliency maps.

2.2 Grad-CAM++

Gradient-based methods generate visualizations that provide explanations for the prediction made by the CNN model with fine-grained details of the predicted class. However, studies have shown that their performance drops when localizing multiple occurrences of the same class, and that they often do not capture the entire object in completeness [6].

Figure 2.2 [6] shows an overview of the difference between the gradient based methods, CAM, Grad-CAM, GradCAM++. Grad-CAM++ [6] is a generalized visualization method proposed to

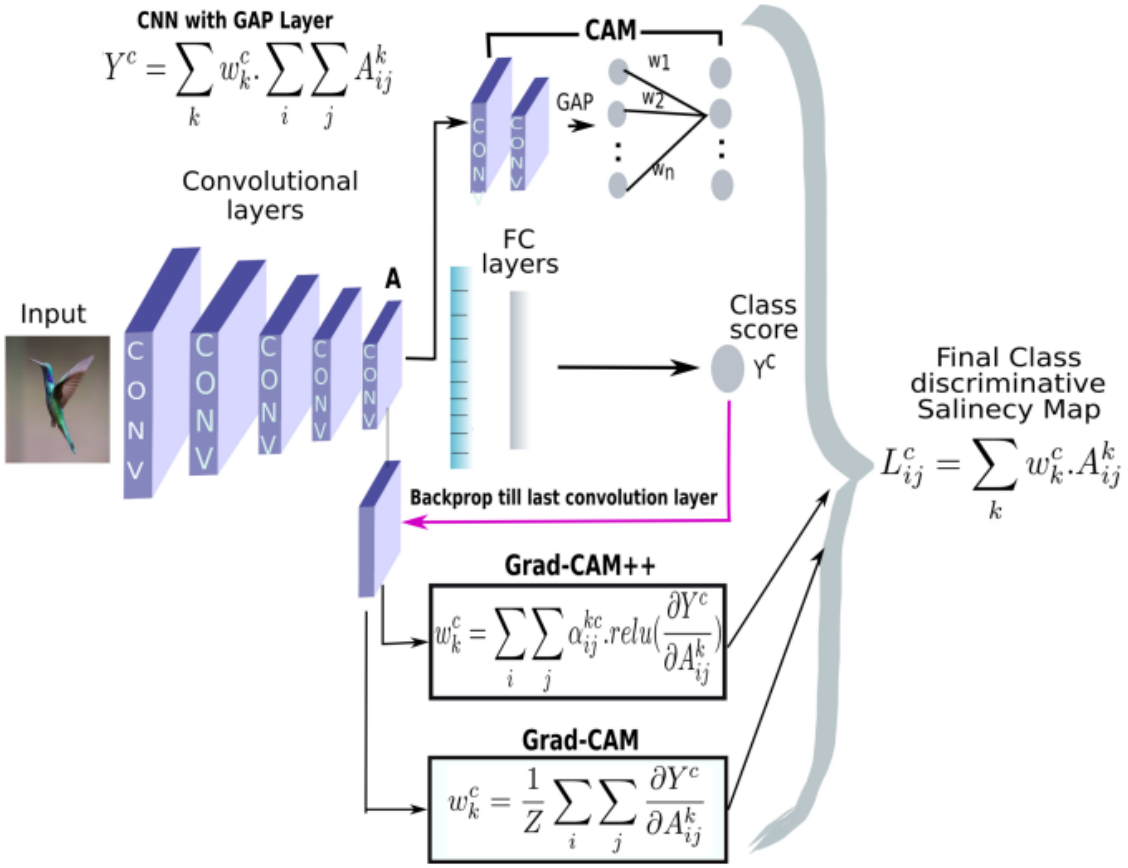

Figure 2.2: Overview of CAM, Grad-CAM, and GradCAM++ with their Computation Expressions. [6]

improve upon the flaws of Grad-CAM, and generate a visual explanation for the corresponding class label. Grad-CAM++ computes a pixel-wise weighted combination of the positive partial derivatives of the CNN's last convolutional layer feature maps with respect to a specific class score and a particular spatial position. Grad-CAM++ is computationally equivalent to prior gradient-based methods, as it only requires a single backward pass on the computational graph [6]. Figure 2.3 [6] shows the intuition behind Grad-CAM++ for the CNN task of binary object classification.

When producing human-interpretable visual explanations of CNN model predictions, Grad-CAM++ was shown to provide better object localization, and to explain occurrences of multiple object instances in a single image, as compared to Grad-CAM [6].

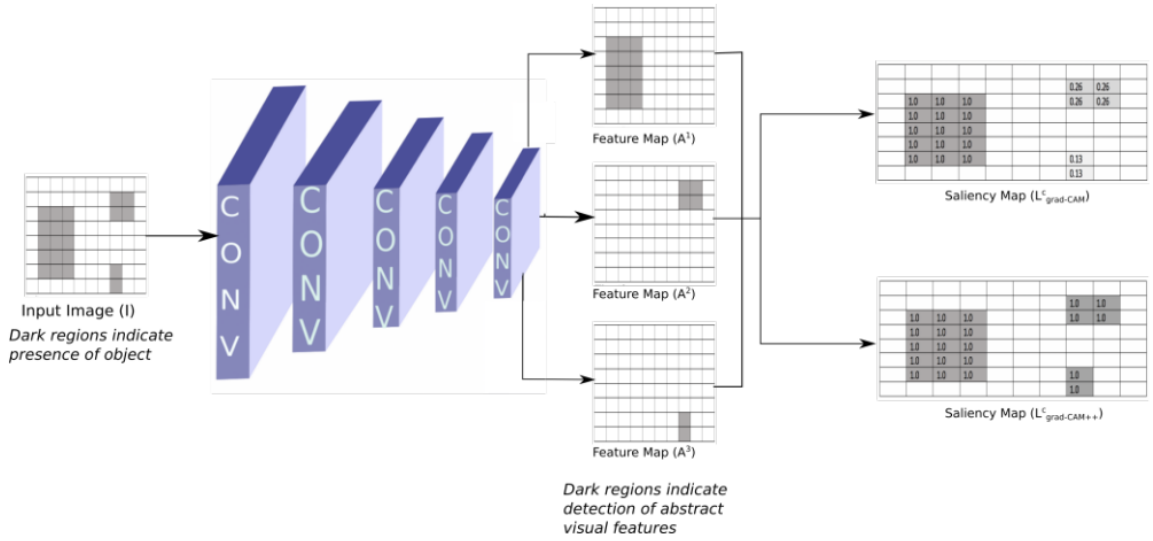

Figure 2.3: Intuition Behind Grad-CAM++ for the CNN Task of Binary Object Classification.
The weighted combination of gradients (GradCAM++) provides better salient features because all the spatially relevant regions of the input image are highlighted equally, as opposed to the unweighted counterpart (GradCAM). [6]

2.3 Layer-CAM

Layer-CAM [7] produces class activation maps, not only from the CNN's final convolutional layer, but also from its shallow layers, which allows for the collection of object localization information from coarse (rough spatial localization) to fine (precise finegrained details) levels.

Layer-CAM utilizes the backward class-specific gradients to generate a separate weight for each spatial location in a feature map. Positive gradients correspond to locations in the feature map where there would be a positive influence on the prediction score of the target class, if the intensity of this location was increased. Gradients are used as weights in these locations, whereas locations with negative gradients are assigned with zero. For each layer of the CNN, Layer-CAM multiplies the activation value of each location in the feature map by a weight, and the class activation map is generated by linearly combining the weighted activation values along the channel dimension [7].

Layer-CAM does not require modification of the network architectures and the back-propagation way, so it can easily be applied to off-the-shelf CNN based image classifiers. Experiments have demonstrated that, compared to other attention methods, Layer-CAM is more effective and reliable for weakly-supervised object localization and semantic segmentation tasks [7].

Figure 2.4 shows an illustration of Layer-CAM [7].

2.4 Eigen-CAM

Eigen-CAM [8] computes and visualizes the first principal components of the learned features/representations from the model's convolutional layers. It is compatible with all convolutional

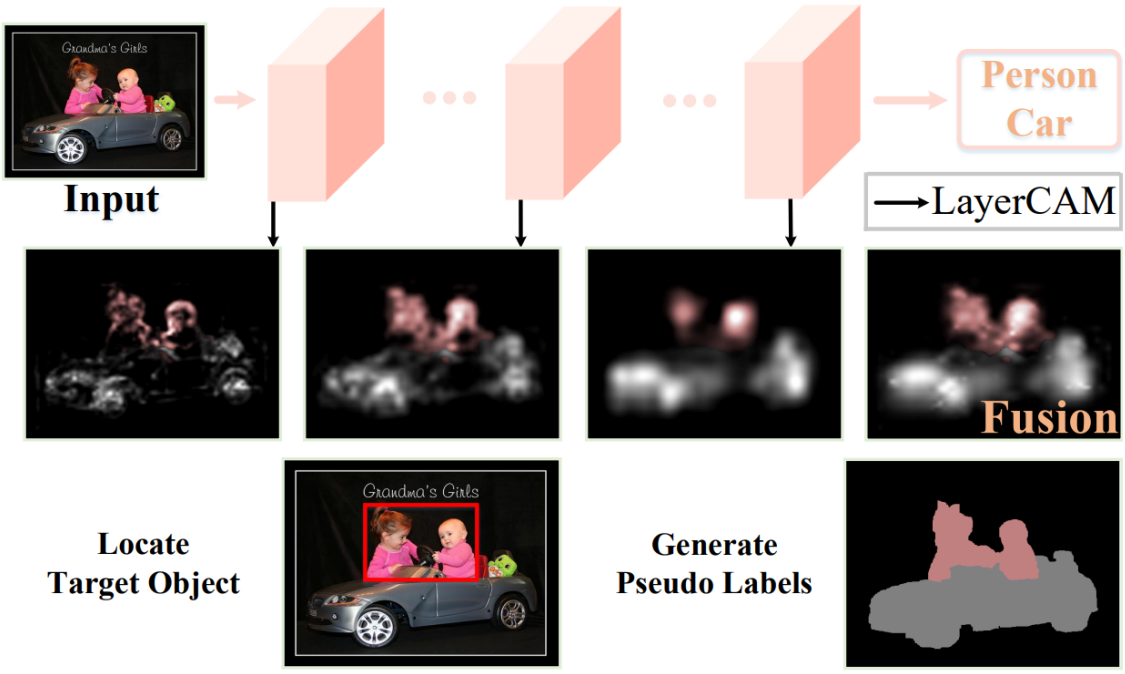

Figure 2.4: LayerCAM Illustration [7]

neural network (CNN) based deep learning model architectures without modification or retraining. Eigen-CAM does not class discriminate, because its calculations are independent of classification layers. Therefore, in contrast to previous methods, Eigen-CAM is independent from the model's classification prediction. Furthermore, Eigen-CAM does not rely on the backpropagation of gradients, class relevance score, maximum activation locations, or any other form of weighting features. The visual explanation is irrespective of model accuracy and adversarial noise, making Eigen-CAM an intuitive, robust, transparent and reliable saliency map method [8].

Empirical evaluation on benchmark datasets such as weakly-supervised localization and localizing objects in the presence of adversarial noise found Eigen-CAM to be robust against classification errors made by fully connected layers in CNNs [8].

2.5 Tension Between Global and Local Attribution

Saliency maps assign each input feature an importance score, which measures the usefulness of that feature for the neural network's task. However, it is difficult to assign a single importance score per feature, due to internal structure among features [9].

Due to this tension, there is no single formal definition of saliency, although local attribution (input features must be considered important if changes to those features significantly change the neural network output) and global attribution (saliency map must completely explain the neural network output) are two common and distinct notions of feature importance [9].

However, these notions of feature importance are counter-intuitive because they result in methods that consider entirely different sets of features to be important, as it can be proven that no saliency maps method can satisfy weak dependence (local importance) and completeness (global importance), simultaneously [9].

2.6 FullGrad

To address this tension, full-gradients [9] was proposed, which decomposes the neural net response into input feature sensitivity and per-neuron sensitivity components. Therefore, full-gradients successfully satisfy both local and global importance, simultaneously, because the significance of individual input pixels is captured by input attribution, and the significance of the interaction between groups of pixels is captured by neuron importances.

For convolutional nets, Full-Grad [9] was implemented, which aggregates the full-gradient components across all intermediate layers of the model.

Figure 2.5 [9] visualizes the bias-gradients for Full-Grad at different neural network layers.

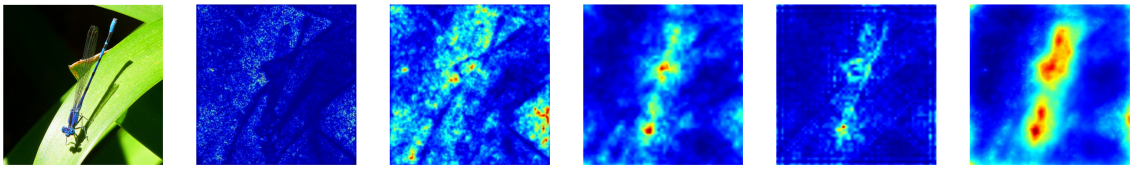

Figure 2.5: Full-Grad Visualization of Bias-Gradients at Different Neural Network Layers.
By aggregating information from the input-gradient and all intermediate bias-gradients, the object is distinguished, which is not achievable by any of the intermediate layer bias-gradients themselves [9].

Quantitative experiments using pixel perturbation and remove-and-retrain tests, conducted in a study by [9], demonstrated that Full-Grad explains model behavior correctly and comprehensively, and visual inspection revealed that Full-Grad maps are sharper and more tightly confined to object regions than other methods. Therefore, [9] concluded that Full-Grad was the first saliency map-based interpretability method to satisfy both the properties of completeness and weak dependence.

Chapter 3

Experimental Setup

3.1 Evaluation of Feature Attribution Methods

During the data collection process, it is possible that image artifacts are leaked accidentally, making a deep learning neural network trained on such images susceptible to zeroing in on undesirable spurious correlations. This may result in the model achieving extremely high validation accuracy and "superhuman" performance on the difficult task of image classification. However, the model's supreme performance is due to it exploiting the leaked artifacts, as opposed to genuine features of the dataset, such as medical signals [4].

However, if counterfactual pairs of images with and without artifacts are known prior to model training, the impact of the artifacts on a model's predication can be evaluated. Unfortunately, all possible artifacts in a natural image dataset cannot be anticipated. Consequently, direct evaluation of feature attribution methods is limited, due to the lack of ground truth classification labels and possible unknown and undesirable spurious correlations.

That being said, these are the cases when explainability algorithms, such as feature attribution methods, should detect these inadvertent spurious correlations by highlighting the leaked artifacts to identify the model's training failure.

Therefore, my experiments aim to overcome this severe limitation, by creating an artificial unambiguous association between the presence of an artifact feature in images and their predicted class, accomplished by systematically manipulating a Covid-19 Radiography database [10], [11] with induced ground truth attributions, followed by training of neural networks.

This dataset modification procedure guaranteed that only the image region where the manipulation was imposed would be an informative feature to the label. Therefore, any sufficiently well performing deep learning classification architecture must focus solely on the manipulation and not get distracted by other irrelevant input features. Accordingly, accurate feature attribution methods should exclusively highlight the manipulated features, as it would be misleading if they assigned attribution scores to image pixels outside the manipulation region [4], which should not have been important to the model's prediction.

My thesis therefore addresses the critical question of how successful saliency maps are at giving attribution to the ground truth label, for (near-)perfectly trained deep learning models.

3.2 Database Modification Procedure

My rationale for determining how to approach conducting the database modification procedure and neural network training sprang from the fundamental ideas described in the Feature Attribution Evaluation GitHub repository [12].

I implemented and ran all experiments in Google Collab on the GPU, using the PyTorch open source machine learning framework. The COVID-19 Radiography Database from Kaggle was used, which originally contained chest X-ray images from patients labelled as having either no disease, COVID-19, or Viral Pneumonia [10], [11]. I next separated the database into train, validation, and test sets, that were each further separated by their original class label. Each of these 3 classes of chest X-ray images in the train, validation, and test sets contained 1145, 100, and 100 images, respectively.

However, the following database modification procedure involving label reassignment and input image manipulation was performed in order to turn this into a binary classification problem and induce ground truth on the database. Figure 3.1 displays a graphical depiction of this dataset modification procedure.

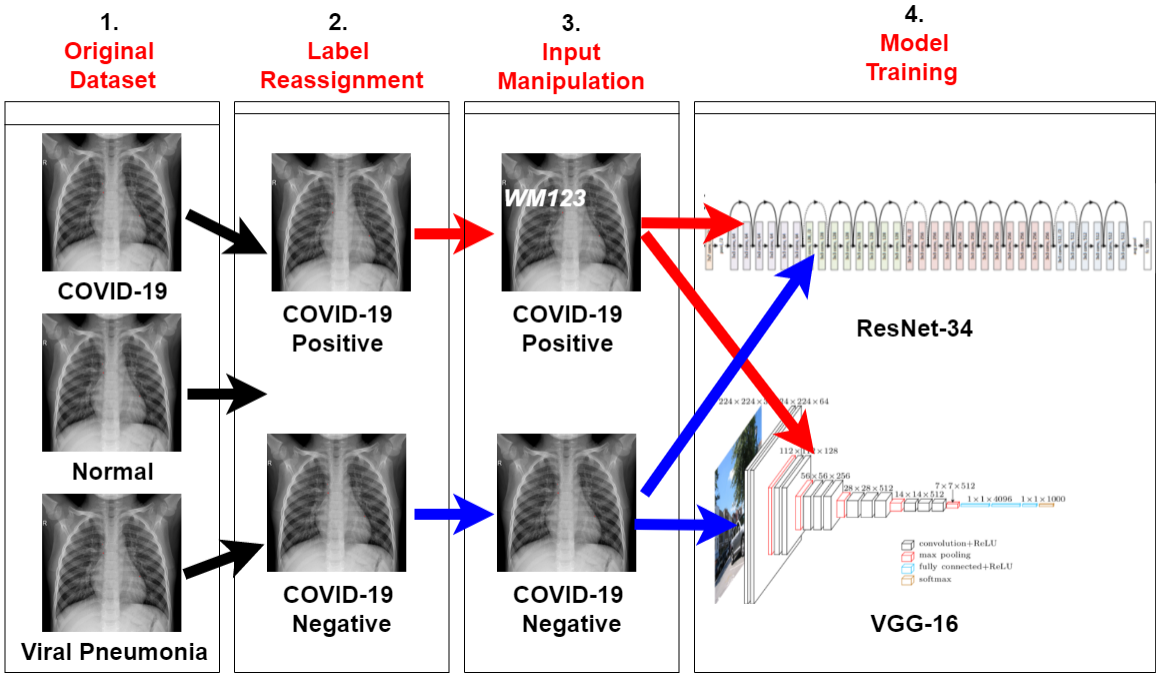

Figure 3.1: Graphical Model of Dataset Modification Procedure

In each experiment, two of the classes COVID-19, Normal, or Viral Pneumonia were chosen. Next, the labels were randomly reassigned with a 50% probability to either the positive or negative

class. This 50% split was used to ensure that the neural network would have an equal number of images from each class to train on, to avoid the undesirable potential for the neural network to be overwhelmed or biased by a larger number of images from one class over the other.

Subsequently, a watermark artifact input manipulation was added to a randomly parameterized effective region of all images exclusively in the new positive label class. For each manipulated image, I also created a ground truth binary segmentation map representing the manipulation. I tracked the random location and size of the induced watermark, and created a black image with a white rectangle of the same dimensions and positioned in the same location as the imposed watermark. The watermark was used because radiography databases can naturally contain text artifacts (e.g., X-ray operator's signature, patient diagnosis, etc.).

Figure 3.2 displays an example of a manipulated image with the induced watermark and its corresponding binary segmentation mask, which acts as the ground truth.

3.3 Ground Truth Induction Theory

The dataset modification procedure described above and graphically modelled in Figure 3.1 required label reassignment and input image manipulation.

Random label reassignment for binary classification weakened the correlation between the original image features and the label, to guarantee the model's reliance on the induced watermark features. Preserving the original label with probability r and flipping it otherwise ensured that without relying on the manipulation, the model's maximum achievable accuracy was $p^* = max(r, 1 - r)$ [4]. In all my experiments, I set $r = 0.5$, which implied that the model's performance was expected to be random, because no original input features would be informative to the new classification label in any coherent manner.

Input manipulation was next applied based on the reassigned label. By adding the watermark manipulation exclusively to a local region of images in the positive reassigned label class, the feature attribution methods could be evaluated [4], as they would be expected to only recognize the contribution of features within the confined local region of images in the new COVID-19 positive class where the manipulation was applied.

Therefore, the dataset modification procedure imposed ground truth and allowed for the quantification of the influence of certain feature's to the model's classification decision. It did so by ensuring that to achieve high accuracy, the trained model must have solely relied on the induced watermark [4], as the watermark was the only distinct feature associated exclusively with X-ray images randomly reassigned to the COVID-19 positive classification.

The lack of ground truth in most real-world datasets makes it difficult to quantitatively assess, compare, and contrast various explanations produced by feature attribution methods, which emphasizes the importance of this database modification procedure and its imposition of ground truth. The dataset modification procedure also simulated an artifact that could be accidentally leaked during

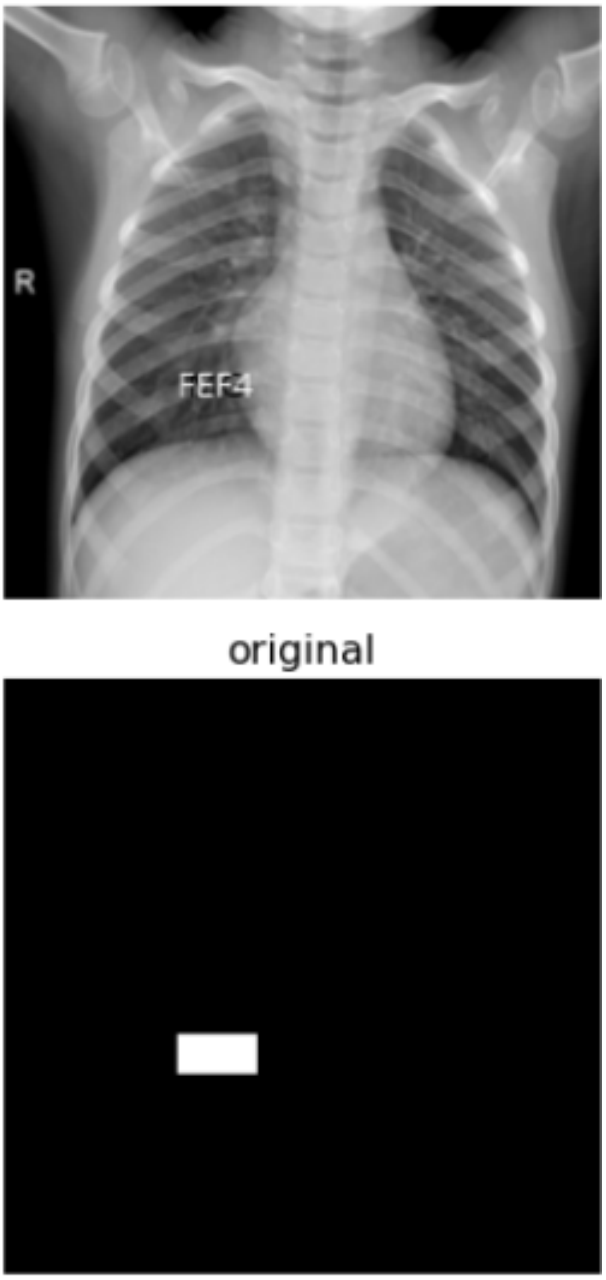

Figure 3.2: Watermark Manipulated Image and Ground Truth Binary Segmentation Mask

the data acquisition process, and cause undesirable spurious correlations if a trained model relies on these induced artifacts to make its classification prediction decisions.

3.4 Neural Network Training

The ResNet-34 and VGG-16 model architectures from the PyTorch Torchvision library with non-pretrained, randomly initialized parameters, were trained on the modified, watermark induced chest

X-ray database, and performed binary classification of the manipulated and unchanged X-ray images, separating them into Covid-19 positive and Covid-19 negative classes.

During training of both the ResNet-34 and VGG-16 models, I used a batch size of 128 and set the maximum number of epochs to 50. The Adam optimizer was implemented with a learning rate of 0.001, beta values of 0.9 and 0.999, and no weight decay. The Reduce Learning Rate on Plateau scheduler was set to max mode based on validation accuracy, with a patience of 30 and a learning rate reduction factor of 0.1. Cross Entropy Loss was used as the loss function.

3.5 Class Activation Map (CAM) Generation

For each run of the experiment, several images from the newly assigned positive label class containing the input manipulation watermark were extracted, along with the weights corresponding to the model with the highest validation accuracy following neural network training.

These images and best model weights were used as input for the generation of class activation maps (CAMs), specifically Grad-CAM, Grad-CAM++, Layer-CAM, Eigen-CAM, and Full-Grad, which are implemented in the PyTorch Grad-CAM GitHub repository [13]. All images in the batch were converted to input tensors and normalized. The target layer input parameter was set as the final convolutional layer of the neural network, just prior to the pooling and classification layer. The target was set to None, implying that the highest scoring category was used for every image in the batch.

All class activation map methods follow a similar implementation procedure, and assign an attribution score ranging between 0 and 1 to all input features of the image, which represent the importance that each pixel had on the model's classification prediction. I used these probabilities to generate gray-scale masks to represent the attribution scores. Next, the gray-scale masks were overlaid onto the original image to create the heatmaps. Studies have suggested that the heatmaps should be compatible with human's visual comprehension, easy to understand and interpret. Figure 3.3 displays an example of a generated grayscale mask and the associated class activation map.

Subsequently, to easily visualize and compare the hundreds of gray-scale masks and heatmaps produced, I created functions to plot the generated class activation maps, along with the maximum validation set accuracy that the trained model reached.

3.6 Quantitative Vs. Qualitative CAM Evaluation Metrics

Studies have shown that many post-hoc analysis methods aimed to assist in achieving explainable artificial intelligence, such as class activation maps (CAMs), have failed to reveal the underlying mechanisms of the deep neural network, and therefore, have demonstrated little use in helping developers to debug, fix, or improve AI algorithms and models in meaningful ways. This is a direct

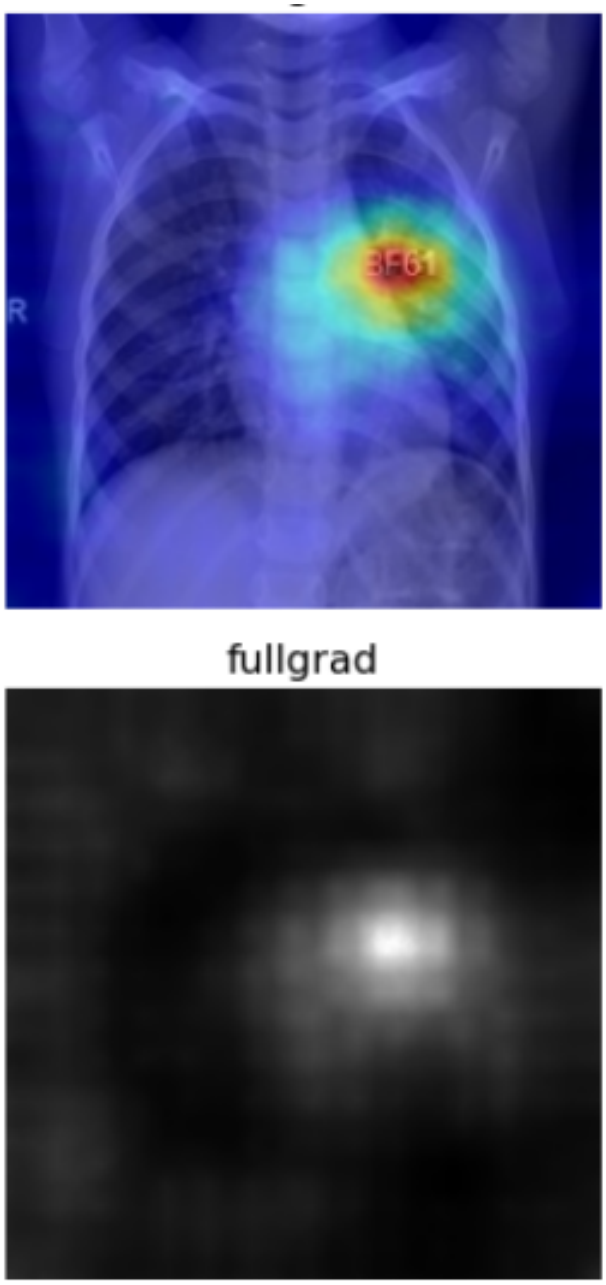

Figure 3.3: Gray-scale mask and Corresponding CAM Heatmap

consequence resulting from the formulas used for generating heatmaps, which are implemented using heuristics such as by setting negative signals to zero or multiplying input with gradients [14].

Furthermore, the potential of heatmaps has been difficult to see directly, because existing quantitative evaluation metrics have been shown to be misleading, and many qualitative assessment techniques appear to be given in hindsight, as they seem to fit natural judgement. Additionally, evaluation of different CAM methods is not yet standardized because ground truth heatmaps are lacking in most datasets. Therefore, it is challenging to test the correctness of CAMs through evaluation of

a CAMs response to alterations in the deep neural network parameters or modifications in the input data ordering to the model, which are common sanity checking techniques that require comparison between the CAM and the ground truth [14].

One study [14] compared heatmaps produced by saliency, deconvolution and layerwise relevance propagation explainability methods by computing the Area over Perturbation Curve (AOPC). Heatmap pixels were ordered according to importance, and then the most important pixels in the original image were replaced in iterations, with the AOPC computed with each perturbation. However, this metric was problematic because features a human finds relevant may not correspond to pixels that were computationally most relevant [14].

Therefore, development of robust quantitative metrics for explainable artificial intelligence (XAI) algorithms continues to be an active area of research, and use of both qualitative and quantitative evaluation metrics in parallel is the state-of-the-art approach.

Intersection over Union (IoU), also known as the Jaccard index, is a popular and straightforward evaluation metric for measuring the accuracy of DNN models in tasks such as image segmentation, object detection, and tracking. In object detection, IoU computes the area of overlap between the predicted bounding box and the ground-truth bounding box, and divides this by the area of union, the total area encompassed by both the predicted and ground-truth bounding boxes. In image segmentation, this analysis in done on a pixel-by-pixel basis. Therefore, as long as the ground truth is known, IoU can also be used for quantitative assessment of class activation maps.

In image segmentation, object detection, and tracking, it is unrealistic for there to be a complete match between predicted segmentation mask and ground-truth mask, due to the varying parameters in the model. IoU therefore rewards model predictions for heavily overlapping with the ground-truth, and IoU scores can range between 0 and 1 (a perfect match).

Other objective numerical assessments of CAM accuracy approximate the IOU metric, such as by calculating the fraction of CAMs that appear to match the location of induced artifacts, or computing the distance from the centre of mass of a saliency response map to the artifact, averaged over all CAMs produced.

3.7 CAM Intersection Over Union Computation

My thesis attempted to evaluate the correctness and quality of CAM explanations using IoU, a quantitative metric to measure pixel-wise accuracy.

As explained, during the database modification step in my experiments, for each image in which a watermark artifact was induced, I created a corresponding ground truth binary segmentation mask. These masks contain a white rectangle on a black background, which represents the effective region containing the induced watermark artifact, as the rectangle is in the same location and of the same dimensions as the manipulation. Therefore, these ground truth masks were used to evaluate and compare how well each CAM methods is at localizing the discriminating watermark feature.

Furthermore, the gray-scale images I saved while creating the CAM heatmaps were used to generate the binary CAM prediction segmentation masks. The values of each pixel in the grayscale images range from 0 to 255, and represent the attribution scores the CAM algorithm assigned to each image pixel. Moreover, higher scores should represent pixels that had more importance or influence on the model's classification prediction.

Therefore, to generate the binary CAM prediction segmentation masks, I created a function that took different threshold values as a hyper-parameter. The thresholds ranged between 50% and 95%, at 5% intervals, multiplied by 255, the maximum possible pixel value. For every pixel in each grayscale image, I set only the pixels with an attribution score greater than the threshold to 1, and all other pixels to 0. This procedure resulted in binary masks where only image pixels that met the threshold criteria for the minimum displayed prediction score were displayed as white regions.

Figure 3.4 displays an example of a gray-scale image and its corresponding binary prediction segmentation map, in which 75% is the threshold.

To compute IoU for each class activation map, I calculated the intersection, or area of overlap, between the ground truth and prediction segmentation masks, at each threshold value, and divided this by the union, or combined white regions of each binary mask. I performed this computation on the CAMs produced by Grad-CAM, Grad-CAM++, Layer-CAM, Eigen-CAM, and Full-Grad, and wrote the resulting IoU scores at each threshold to a file, for each manipulated image.

Figure 3.5 displays an example of a generated IoU file.

In each of my experiments, I randomly selected 10 of the watermark manipulated images in the test set that had been used to compute the test accuracy of either the ResNet-34 or VGG-16 model architectures, and generated the CAM heatmaps using these randomly selected images. Therefore, I average the IoU scores across all 10 manipulated images, and wrote these averaged IoU scores to a new file, to facilitate easy evaluation and comparison between the IoU scores across different threshold values and CAM methods.

3.8 CAM Statistical Analysis Using T-Tests

A Paired Sample T-Test is an inferential statistic used to quantify the difference between the arithmetic means of the two samples. The null hypothesis assumes that the samples are drawn from populations with the same population means, and have identical variances.

The probability of observing as or more extreme values, assuming that the null hypothesis is true, is quantified by the p-value. The null hypothesis of equal population means cannot be rejected if the p-value exceeds the chosen threshold (e.g. 5% or 1%), because this would imply that the observation is not so unlikely to have occurred by chance. However, in contrast, there is evidence against the null hypothesis, if the p-value is smaller than the threshold.

Therefore, for all experiments, after computing the individual and average IoU scores across different threshold values for the heatmaps created by each CAM method, I determined which CAM method generated the highest average IoU score, for each threshold. Then, I used the individual IoU

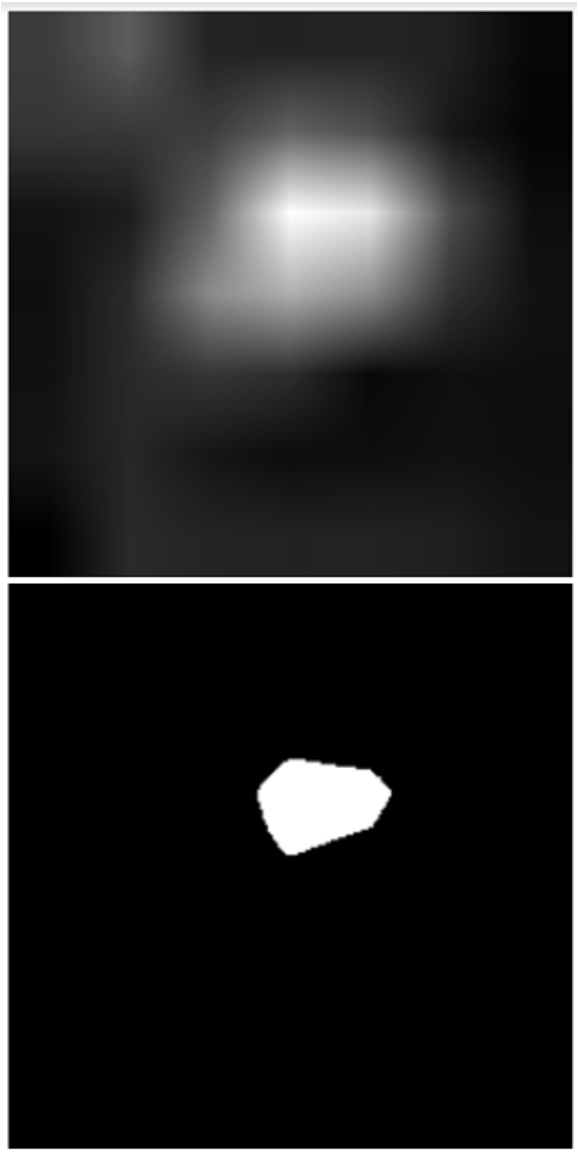

Figure 3.4: Gray-Scale Mask and Prediction Mask at the Threshold of 75%

```
2022-07-02 20:22:36.332572.txt

/content/gdrive/MyDrive/Radiography_FeatureAttributionEvaluation_SFU2022/VGG16_CAMs/watermark_5/ModifiedImage0/manip1-COVID-COVID-1247.png

thresholds = [128, 140, 153, 166, 178, 191, 204, 217, 229, 242]

GradCAM: [0.08724311748809217, 0.10096477451279462, 0.12379642365981734, 0.16586804275830117, 0.23663313922868903, 0.32576350822446676,
0.37178051512033994, 0.3867924528339697, 0.3179074446733901, 0.12666666667427742, ]

EigenCAM: [0.10392609699850207, 0.11676180591683022, 0.1337693222364438, 0.1570680628283789, 0.19556714472105807, 0.2727272727290012,
0.40132200189079104, 0.557165861516484, 0.42418032787347976, 0.14444444445190027, ]

GradCAM++: [0.08625646923587948, 0.09769865393042519, 0.11337868480813225, 0.13313609467556195, 0.16129032258182402, 0.22826639552772057,
0.3734622144134068, 0.4726507713913997, 0.39923224568590393, 0.15555555556291456, ]

LayerCAM: [0.10103277952481485, 0.12315270936054695, 0.16059957173565015, 0.21888834235271182, 0.26449968132752344, 0.298018949184111,
0.32936979786281795, 0.3403908794830402, 0.27016129032835107, 0.11555555556326315, ]

FullGrad: [0.3224400871478991, 0.389041095892599, 0.47831184056511816, 0.5838414634171218, 0.6313868613165063, 0.5979166666699517,
0.46491228070635615, 0.3066666666727088, 0.1888888888959574, 0.08888888889682886, ]
```

Figure 3.5: Example File with CAM IoU Scores

results associated with every CAM to conduct a Paired Sample T-Test to compare the IoU scores of the CAMs produced by Full-Grad to those produced for the second highest scoring method at that threshold.

For these experiments, the p-value threshold, alpha, was set as 5%. Accordingly, a p-value smaller than alpha implied that there is evidence against the null hypothesis, that Full-Grad and the 2nd highest IoU scoring CAM method have equal population means. This would help verify that the CAMs produced by both methods are different enough in terms of their accuracy and performance in explaining the model's predictions to be considered statistically significant.

Chapter 4

Experimental Results

4.1 Background Behind Analysis of Class Activation Maps

In binary classification experiments, if a trained neural network model architecture achieves high validation accuracy, it demonstrates that the model learned to associate images manipulated with the imposed watermark artifact to the Covid-19 positive class. Therefore, the neural network must have used the imposed input manipulation as a discriminating feature to make its classification prediction decisions. This implies that all class activation map methods would be expected to only highlight the induced watermark artifact, if the CAM is accurate at localizing image regions important to the model's prediction.

CAM methods assign attribution scores to each image pixel, which represent how influential the CAM method calculated each pixel to be in terms of influencing the trained model's prediction. Therefore, I created binary prediction masks for all CAMs by thresholding the attribution scores at values between 50% and 95%, in 5% intervals. For each experiment, I used the binary prediction masks and ground truth segmentation masks to compute the average intersection over union for several randomly selected watermark manipulated images taken from the test set.

The experiments conducted evaluated the accuracy of the class activation maps produced by Grad-CAM, Eigen-CAM, Grad-CAM++, Layer-CAM, and Full-Grad, which each attempt to explain the classification predictions of both the ResNet-34 and VGG-16 model architectures. I performed this analysis using both qualitative and quantitative evaluation metrics, and my interpretation of the results follows.

4.2 CAMs Produced Using ResNet-34

The ResNet-34 model architecture achieved a maximum validation accuracy of greater than 97.5% and a low training loss in all the binary classification experiments conducted, and did so quickly in under 5 training epochs. This implies that ResNet-34 must have used the imposed watermark as a discriminating feature to make its classification prediction.

In the majority of the experiments, the produced class activation map methods highlighted the induced watermark artifact, and was therefore accurate and precise at localizing the image regions important to the model's prediction classifications, as desired.

Figure 4.1 and Figure 4.2 are a sampling of a few of the produced gray-scale CAMs and corresponding heatmaps produced by each CAM method tested, extracted from one such experiment.

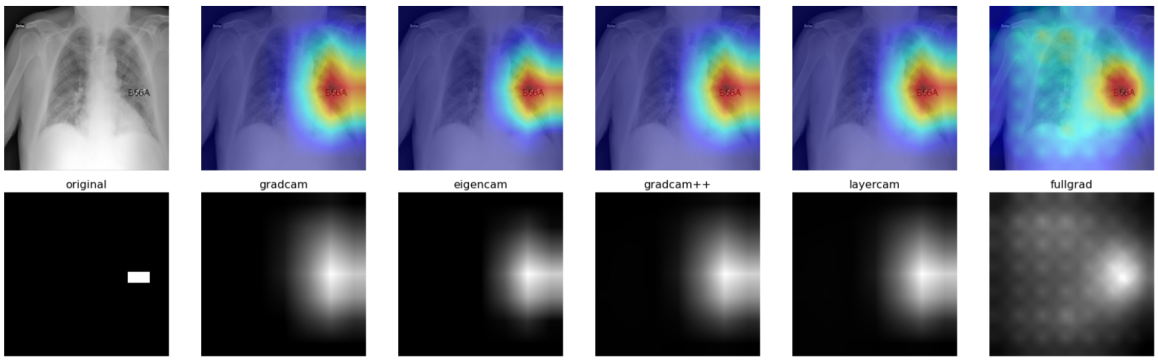

Figure 4.1: Accurate CAMs Highlighting the Imposed Watermark Artifact

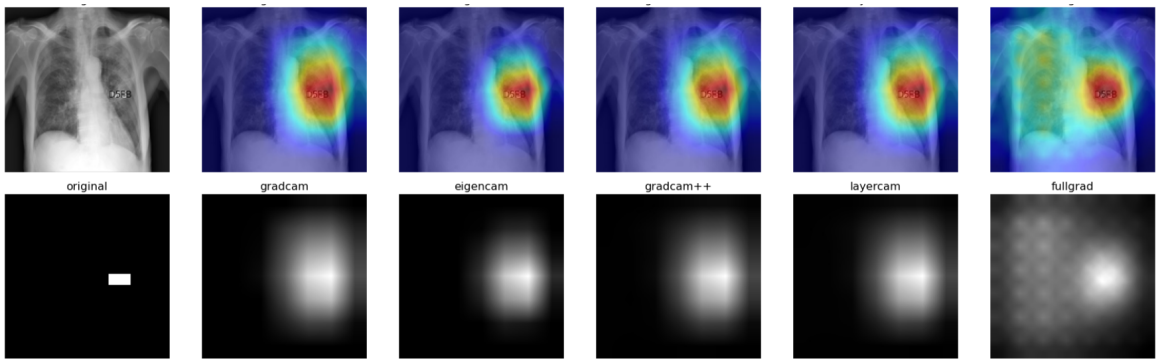

Figure 4.2: Another Example of Accurate Heatmaps and Corresponding Gray-Scale CAMs

Judging the CAM results purely based on visual inspection, Full-Grad seemed to produce the sharpest, most targeted gray-scale representation for the model's decision, because the brightest white image regions, which correspond to the image pixels calculated by Full-Grad to be of greatest influence to ResNet-34's prediction, are centered on and directly overlap the imposed watermark artifact. However, an undesirable disperse checkerboard pattern is present, and seen only with Full-Grad. Excluding Full-Grad from the qualitative analysis and comparing the heatmaps generated by the other CAM methods tested, Eigen-CAM looks to have produced the next best results.

The above qualitative analysis agrees with the quantitative average IoU calculations. Table 4.6 displays the IoU scores at each threshold, along with the average of these IoU scores across all thresholds, for each CAM method.

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.079	.094	.115	.146	.187	.235	.286	.321	.292	.136	.189
Eigen-CAM	.125	.150	.187	.227	.265	.304	.338	.328	.236	.075	.224
Grad-CAM++	.091	.110	.137	.176	.222	.271	.318	.340	.274	.091	.203
Layer-CAM	.091	.109	.136	.175	.222	.271	.318	.341	.277	.095	.203
Full-Grad	.075	.104	.134	.170	.214	.286	.378	.489	.530	.384	.276

Table 4.1: Average IoU for Accurate CAMs Produced Using ResNet-34

In this experiment, Full-Grad surpassed the other CAM methods and achieved the highest average IoU score of 0.53 at a threshold of 90%. Excluding the IoU scores of Full-Grad, the next highest average IoU was only 0.34, achieved at a threshold of 85% for Layer-CAM.

Intuitively, if a high cutoff threshold is used to create the binary CAM prediction mask and this mask is compared with the binary ground truth segmentation mask for the corresponding manipulated image in order to compute IoU, the high threshold would eliminate all pixels that were given attribution scores below the threshold. This pixel elimination would remove dispersion in the CAM produced, which would be particularly valuable to Full-Grad, and explains why Full-Grad achieved the best average IoU results at high thresholds.

Due to the small size of the watermark artifact, it would be very difficult for any CAM method to perfectly localize the manipulation exclusively, which would be required in order to achieve an IoU score close to 1. Therefore, the average IoU score of 0.54 at 90% threshold cutoff achieved by Full-Grad should be considered high, and is a very promising result.

The results of the paired sample t-tests conducted using the ResNet-34 model support the above analysis. For IoU thresholds below 65%, Eigen-CAM had a slightly higher IoU score than Full-Grad, and the CAMs produced by both were different enough to result in a p-value below alpha, indicating evidence against the null hypothesis (that the mean IoU scores of both methods came from the same population). This was likely due to The high degree of dispersion that would not be filtered out of the Full-Grad CAMs at a low threshold. However, for thresholds at and above 85%, Full-Grad achieved superior IoU scores and much more targeted CAMs as compared to all other methods. This difference was confirmed by the extremely low p-values generated at these thresholds, and therefore, the implied statistically significant evidence against the null hypothesis. These t-test results provide statistical confirmation that thresholds CAMs to different degrees can enhance the performance of CAM methods, although the ideal specific threshold cutoff is dependent on the CAM method used, due to each CAM method's underlying implementation.

The results of the paired sample t-tssts conducted using the ResNet-34 model support the above analysis. For IoU thresholds below 65%, Eigen-CAM had a slightly higher IoU score than Full-Grad, and the CAMS produced by both were different enough to result in a p-value below alpha, indicating evidence against the null-hypothesis (that the mean IoU scores of both methods came from the same population). this was likely due to the high degree of dispersion that would not be filtered out of the Full-Grad CAMS at a Los threshold. However,

Figure 4.3 displays gray-scale CAMs and their corresponding heatmaps, sampled from a different experiment where the trained ResNet-34 model had different learned weights.

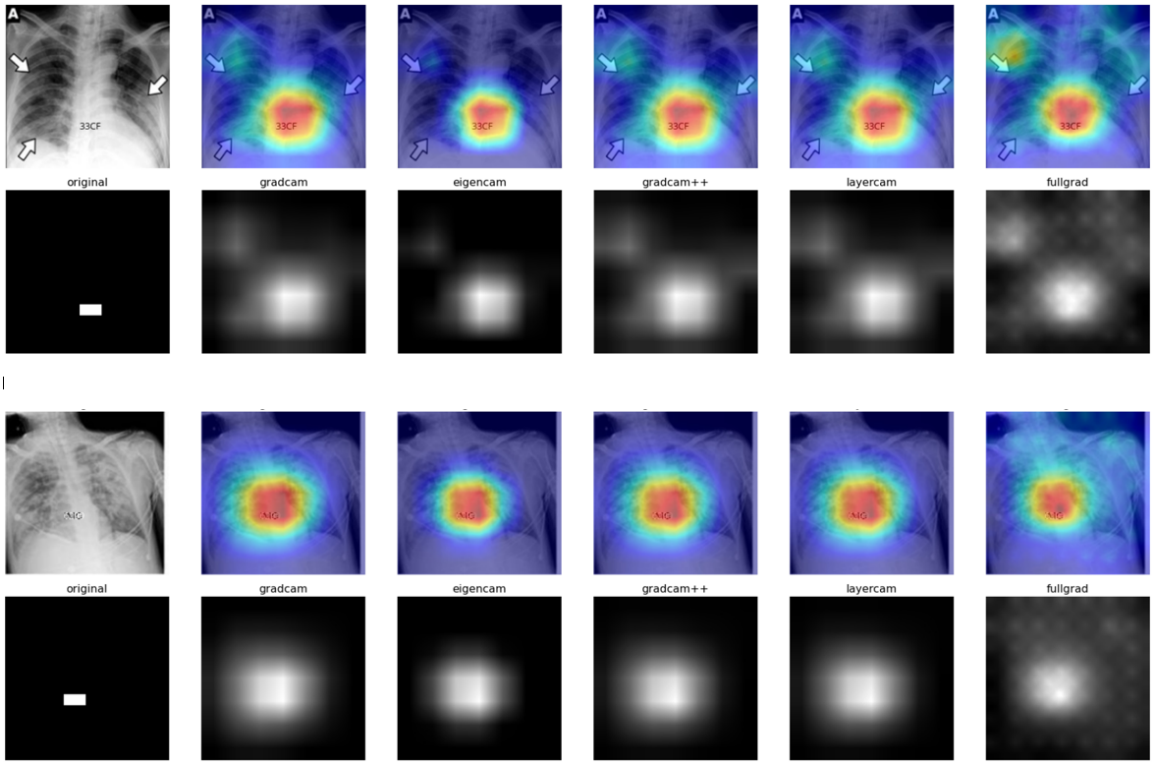

Figure 4.3: CAM Method Comparison Generated for ResNet-34

Upon visual inspection, Figure 4.3 demonstrates that Grad-CAM, Grad-CAM++, and Layer-CAM generated CAMs that appear nearly identical. This fact is also seen in the computed average IoU scores, which are virtually the same for these 3 methods at every threshold cutoff.

Figure 4.3 also showcases that Eigen-CAM produced the gray-scale CAM with the least dispersion and best overall localization of the induced watermark. The average IoU scores displayed in Table 4.2 support this claim, because Eigen-CAM had the highest average IoU of all CAM methods at every threshold below or equal to 85%, ranging from an IoU of 0.114 at a low threshold of 50% to 0.265 at mid threshold of 80%. It was only at the thresholds of 90% and 95% that Full-Grad achieved higher IoU than Eigen-CAM, which makes sense considering the amount of additional dispersion the Full-Grad CAMs contain compared to the more focused Eigen-CAMs.

Finally, Figure 4.4 displays gray-scale CAMs and their corresponding heatmaps, sampled from yet another experiment, where the trained ResNet-34 model achieved a maximum validation accuracy of 98.5% after only 2 epochs.

However, for all CAMs other than Full-Grad, the generated CAMs are not focused on the imposed artifact to any degree, and instead seem to be highlighting other distinct or prominent regions of the input image that stand out, particularly the border separating the spine from the lung. This region may have been used by the model to make its predictions, if the data modification procedure had not

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.082	.096	.115	.140	.170	.207	.232	.229	.185	.107	.156
Eigen-CAM	.115	.133	.159	.192	.226	.257	.265	.238	.185	.110	.188
Grad-CAM++	.090	.105	.127	.156	.192	.239	.269	.273	.226	.133	.181
Layer-CAM	.088	.103	.124	.152	.187	.232	.263	.265	.222	.133	.177
Full-Grad	.095	.113	.137	.165	.193	.216	.232	.232	.209	.136	.173

Table 4.2: Average IoU for CAMs with Similar Qualitative Results

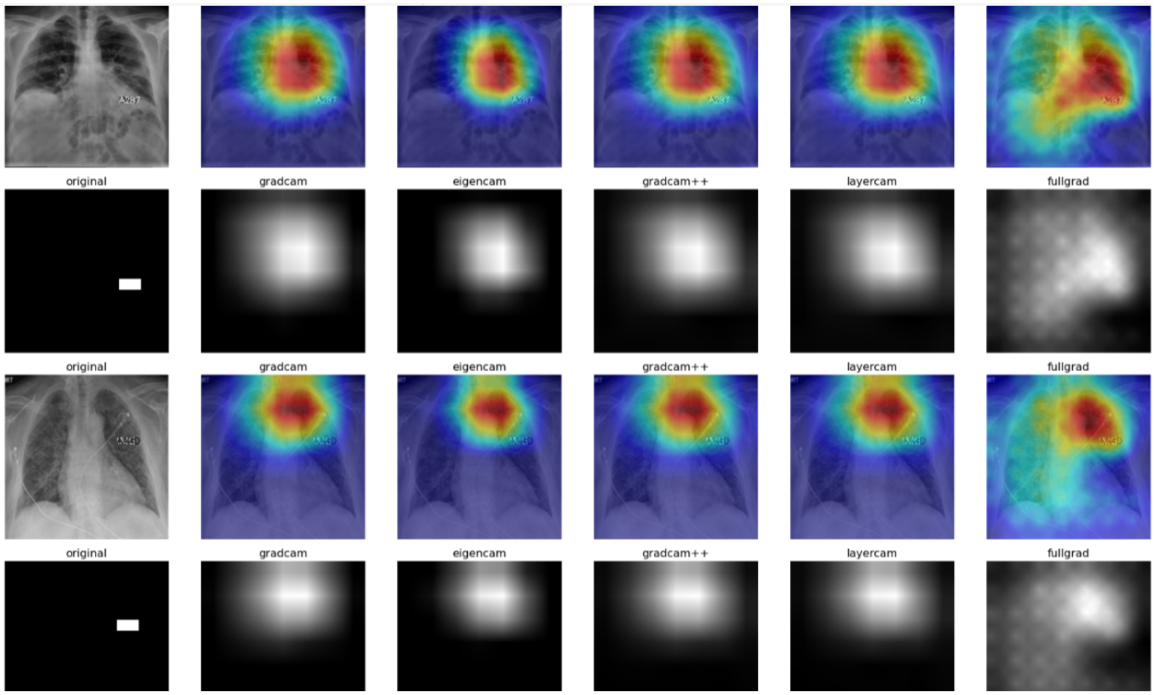

Figure 4.4: Inaccurate CAMs Not Highlighting the Imposed Watermark Artifact

been conducted. The average IoU scores computed agree, and are close to 0 for all CAM methods at every cutoff threshold.

Although Full-Grad is not targeted on the watermark, the watermark is still included at the side of the region that Full-Grad is highlighting. Table 4.3 shows that Full-Grad achieved IoU scores higher than the other methods, with best IoU results ranging from 0.0362 to 0.0683 for cutoff thresholds between 50% and 80%. At the 95% threshold though, the average IoU for Full-Grad was 0, because the attribution scores assigned to the pixels corresponding to where the watermark was imposed, no longer meet the threshold cutoff criteria, so the corresponding binary segmentation prediction mask would no longer contain any overlap with the ground truth mask.

It is worth noting that in this experiment, the watermark artifact is imposed closer to the input image border, as opposed to being located closer to the input images center, as was the case in the experiments described previously. This may suggest that the position of the imposed manipulation

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Eigen-CAM	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Grad-CAM++	.007	.003	.000	.000	.000	.000	.000	.000	.000	.000	.000
Layer-CAM	.008	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000
Full-Grad	.036	.044	.054	.064	.068	.059	.037	.020	.013	.000	.044

Table 4.3: Average IoU for CAMs Not Targeting Watermark

impacts the accuracy of CAM methods, as it seems to be easier for CAMs to localize and discriminate artifacts located more centrally, which may be a result of CAMs underlying implementation.

To conclude, this final experiment using ResNet-34 suggests that no CAM method is accurate in all cases and should therefore be tested to an even greater extent through scientific studies. Having said that, these results do suggest that the potential for the use of CAMs is promising, as most experiments produced CAMs that localized the watermark perturbation well qualitatively, while simultaneously achieving high IoU scores.

4.3 CAMs Produced Using VGG-16

In comparison to ResNet-34, which achieved a maximum validation accuracy of above 97.5% in all experiments, the validation accuracy scores for VGG-16 varied significantly in every binary classification experiment conducted. This variation could stem from the distinct underlying model architectures and parameters used during model training, such as weight decay, learning rate, momentum, and dropout. However, it could also be a result of the input data used in each experiment, because if other attributes of the images are more distinctive compared to the imposed watermark artifact, the model may not use the watermark as a discriminating feature to make its classification prediction.

In cases of low model maximum validation accuracy, class activation map methods would not be expected to highlight the induced watermark artifact, because when the model's predictions are inaccurate, it would be very difficult for the CAM method to calculate what features were of highest importance to the model. Therefore, variation in the produced CAMs would be expected.

This point was emphasized through one of my experiments, in which the VGG-16 model only achieved a maximum validation accuracy of 55%, after being cutoff at the maximum of 50 epochs. Figure 4.5 displays a couple of the manipulated test set images used in this experiment, their corresponding binary ground truth segmentation masks, as well as the CAM heatmaps and gray-scale CAMs produced by each method.

Grad-CAM and Grad-CAM++ produced nearly identical CAMs that highlight almost the entire image, which showcases that these CAM methods did not distinguish between pixels of most probable highest significance to the model's prediction, and therefore, computed nearly equal attribution scores for each pixel. This is also reflected in the identical average IoU scores computed for Grad-CAM and

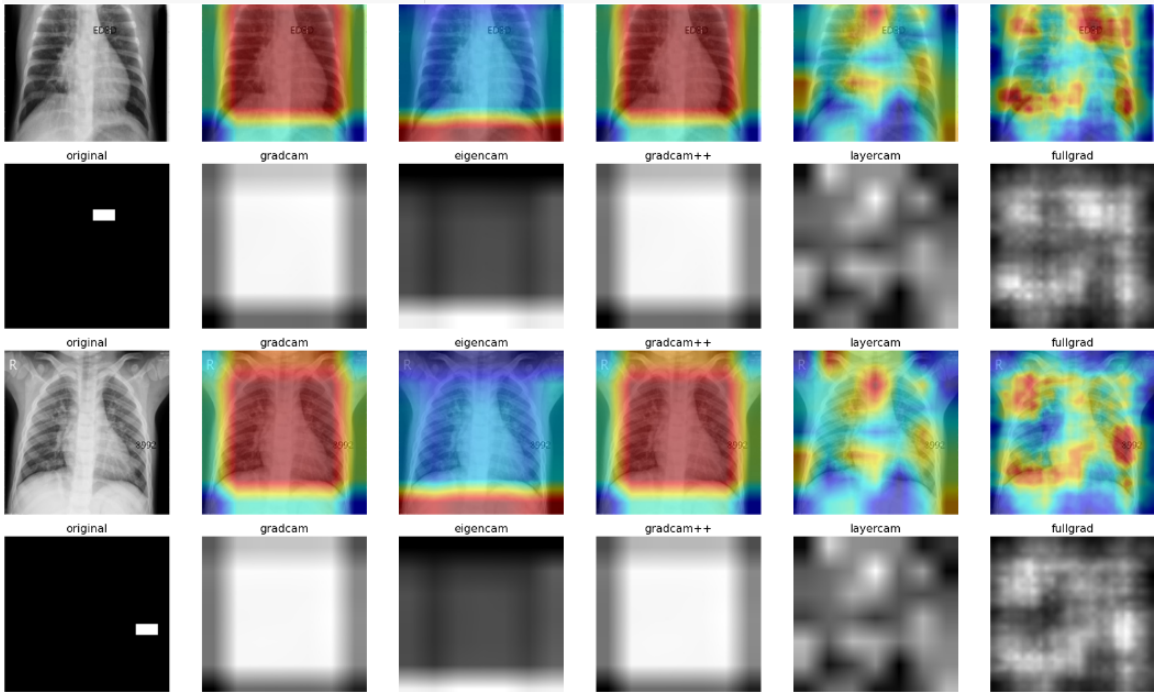

Figure 4.5: CAMs Generated for VGG-16 with Low Validation Accurarcy

Grad-CAM++, which both average around 0.013, and are displayed in Table 4.4. These poor results are likely due to how both these methods use the back-propagation of gradients starting from the final convolution neural networks layer just prior to the classification layer, to calculate attribution scores, and their dependence on the classification of the model.

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.012	.013	.013	.013	.013	.013	.014	.015	.015	.015	.014
Eigen-CAM	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Grad-CAM++	.017	.013	.013	.013	.013	.015	.015	.015	.015	.015	.014
Layer-CAM	.010	.010	.010	.008	.007	.007	.004	.000	.000	.000	.006
Full-Grad	.026	.033	.044	.064	.091	.135	.196	.262	.257	.0125	.123

Table 4.4: Average IoU Scores Produced Using VGG-16 with Low Validation Accuracy

Eigen-CAM shows even worse results, with IoU scores of practically zero, because Eigen-CAM highlighted the same bottom region, for all manipulated images used in this experiment. Eigen-CAM is distinct from all other methods, as it computes the first principal components of the learned features/representations from the model's convolutional layers, which may be why its produced CAMs are so distinct from the other methods tested.

Layer-CAM and Full-Grad look to have generated similar CAMs, but upon closer visual inspection and comparison of IoU scores, many distinctions are apparent. Both methods have assigned high attribution scores to multiple small image regions, dispersed throughout the image. However, with Layer-CAM, the watermark artifact is not included in any of the highlighted regions, so the CAMs

look random, which is why the average IoU at all thresholds for Layer-CAM is only approximately 0.006. In comparison, with Full-Grad, the most distinguishable highlighted regions are targeted directly over the watermark for all manipulated images, and particularly at higher threshold cutoffs, especially those above 65%, the IoU for Full-Grad is multiple orders of magnitude higher than the IoU of any other methods. At a threshold of 85%, Full-Grad scored an average IoU of 0.262, the highest IoU achieved at any threshold for any CAM, as shown in Table 4.4.

As explained in more thorough detail in the Feature Attribution Methods section of my thesis, Layer CAMs use both the CNN's final convolutional layer, as well as its shallow layers to produce CAM heatmaps, which allows for the collection of object localization information from coarse (rough spatial localization) to fine (precise fine-grained details) levels, which is the most probable explanation for the disperse image regions highlighted. In comparison, Full-Grad decomposes the neural net response into input feature sensitivity and per-neuron sensitivity components, which helps to capture the significance of both individual input pixels and the interaction between groups of pixels. This characteristic appears to make Full-Grad very robust and accurate compared to all other methods.

In contrast to the above experiment, in another experiment, VGG-16 achieved 100% maximum validation accuracy in just 3 epochs. Figure 4.6 displays a couple of the generated CAMs.

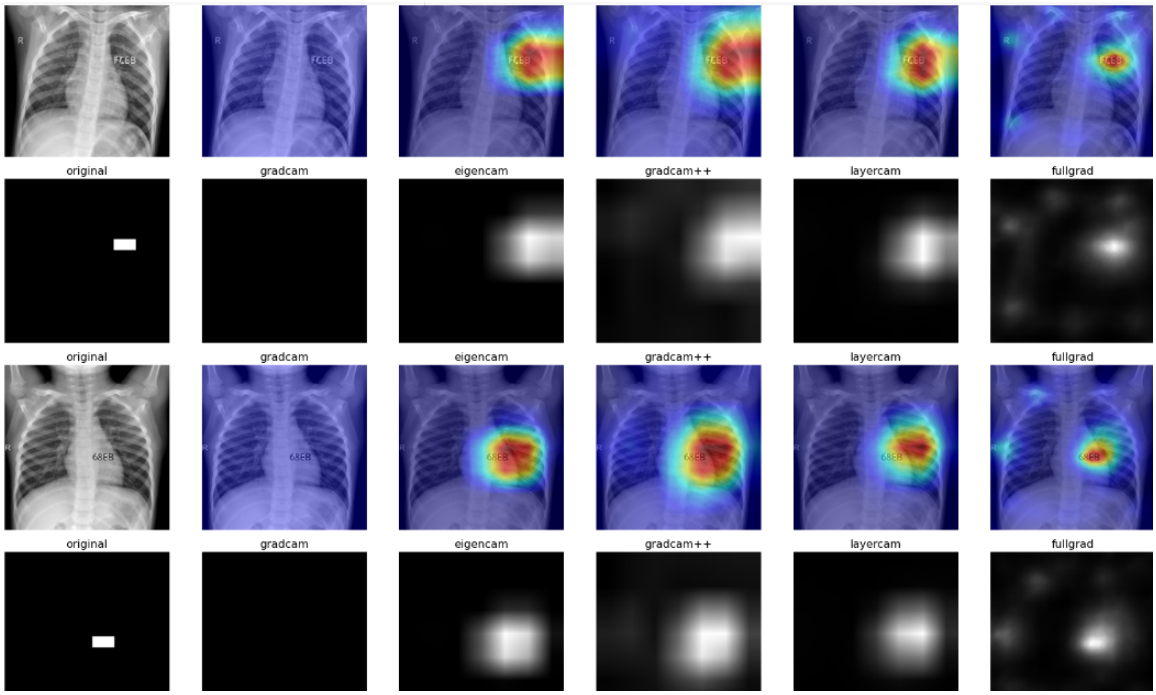

Figure 4.6: Accurate CAMs Using VGG-16 Highlighting the Imposed Watermark Artifact

In this experiment, apart from Grad-CAM, which seemed to not assign attribution scores to any image pixels, all other methods produced CAMs that were fairly accurate. However, besides Full-Grad, these other methods look to be assigning high attribution scores to regions to the right of the watermark, near the boundary of the patients chest and background, and therefore, the heatmaps

produced are not directly targeted on the watermark. As seen from the average IoU scores tabulated in Table 4.7, Eigen-CAM and Layer-CAM both achieved similar IoUs at each threshold, and the IoU scores for Grad-CAM++ were weaker, as its heatmaps were more disperse and less focused.

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Eigen-CAM	.115	.121	.126	.132	.139	.148	.158	.164	.152	.068	.132
Grad-CAM++	.079	.085	.088	.088	.088	.088	.085	.076	.061	.034	.077
Layer-CAM	.132	.147	.164	.180	.192	.198	.186	.148	.105	.0578	.151
Full-Grad	.411	.475	.535	.560	.553	.505	.418	.306	.195	.0731	.403

Table 4.5: Average IoU Scores Produced Using VGG-16 with High Validation Accuracy

In contrast, Full-Grad was the clear winner. It produced CAMs that almost exclusively highlighted only the image pixels that corresponded to where the watermark artifact had been imposed, and therefore, achieved superior localization of the distinctive watermark attribute, the ideal outcome for any CAM method. This was also reflected in the high average IoU scores of Full-Grad, which were greater than 0.411 for all thresholds between 50% and 80%. In fact, at threshold 65%, Full-Grad scored an average IoU of 0.560, which was one of the highest average IoU scores achieved by any method in all experiments.

The results of the paired sample t-tests conducted using the VGG-16 model also confirmed the above qualitative CAM visualization analysis and quantitative IoU scores. For all IoU thresholds below 85%, Layer-CAM achieved the second best average IoU scores. However, compared to Full-Grad, the computed p-values at all these thresholds were extremely lower than alpha, which was 5% for these t-tests. These t-tests provided statistical evidence against the null hypothesis, suggesting that the IoU score means of the CAMs produced by Full-Grad vs. those of Layer-CAM (and all lower scoring CAM methods) were statistically significantly different. It was only at the IoU thresholds of 90% and 95% that the p-value was above aloha, and the null hypothesis could not be rejected (which was the result of the IoU threshold being set too high). The t-tests provide statistically significant evidence to verify that the CAMs produced by Full-Grad as opposed to any other method were of higher quality in terms of its performance ability to explain the classification predictions of the trained VGG-16 model.

The results from the above experiments conducted using VGG-16 suggest that in order for the results of CAMs to be of any use, the neural network must first be trained well enough to achieve a high validation accuracy. Additionally, they suggest that Full-Grad may be the most robust and accurate CAM method.

4.4 Evaluation of Class Activation Map Methods

The results obtained from the conducted experiments suggest that, on average, when used to interpret the classification predictions of convolutional neural networks, Grad-CAM looks to generate the most

disperse class activation maps, while Eigen-CAM produced CAMs that were slightly more targeted on the imposed watermark, although the differences were not extremely noticeable. Additionally, because gradient-based CAM methods are dependent on the correctness of the neural networks classification predictions, these experiments are of greatest value when the model that the CAM method is attempting to interpret has achieved high validation accuracy.

Table 4.6 and Table 4.7 compare the IoU scores obtained by each CAM method when they were used to explain the predictions of the ResNet-34 and VGG-16 models, respectively, which were both trained on the same dataset of manipulated images, and both achieved high maximum validation accuracy (i.e. ResNet-34: 97.5%; VGG-16: 100%).

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.079	.094	.115	.146	.187	.235	.286	.321	.292	.136	.189
Eigen-CAM	.125	.150	.187	.227	.265	.304	.338	.328	.236	.075	.224
Grad-CAM++	.091	.110	.137	.176	.222	.271	.318	.340	.274	.091	.203
Layer-CAM	.091	.109	.136	.175	.222	.271	.318	.341	.277	.095	.203
Full-Grad	.075	.104	.134	.170	.214	.286	.378	.489	.530	.384	.276

Table 4.6: Average IoU for Accurate CAMs Produced Using ResNet-34

Threshold	50%	55%	60%	65%	70%	75%	80%	85 %	90%	95%	Average
Grad-CAM	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Eigen-CAM	.115	.121	.126	.132	.139	.148	.158	.164	.152	.068	.132
Grad-CAM++	.079	.085	.088	.088	.088	.088	.085	.076	.061	.034	.077
Layer-CAM	.132	.147	.164	.180	.192	.198	.186	.148	.105	.0578	.151
Full-Grad	.411	.475	.535	.560	.553	.505	.418	.306	.195	.0731	.403

Table 4.7: Average IoU Scores Produced Using VGG-16 with High Validation Accuracy

The average IoU scores across all thresholds (last column of Table 4.6 and Table 4.7) demonstrate that all CAM methods scored similar average IoU results, although Grad-CAM scored the lowest average IoU using both models, and Full-Grad scored a significantly higher IoU than all other methods using VGG-16.

When Full-Grad was used to interpret the predictions of the ResNet-34 model, as the threshold was raised to cutoffs above 80% in particular, the IoU scores for Full-Grad increased significantly, and it achieved a high IoU score of 0.53 at a threshold of 90%. In comparison, using VGG-16, Full-Grad achieved very successful IoU scores that exceeded 0.5 for threshold cutoffs at and between 60% and 75%. At 95% threshold however, its IoU performance decreased drastically, which was a result of too many pixels being removed from the prediction mask. These results are demonstrated by Full-Grad's IoU scores, noted in the last row of both Table 4.6 and Table 4.7. The paired sample t-tests conducted also confirmed this analysis, and provided evidence that most of the CAMs produced had mean IoU scores that were statistically different.

The IoU scores and paired sample t-tests results exemplify that thresholding CAM output is an effective means of optimizing the performance of CAM methods, and that high thresholds may be beneficial at eliminating CAM output dispersion.

In general, Full-Grad looked to be most reliable and robust in highlighting the imposed artifacts, as Full-Grad usually produced the sharpest, most targeted CAMs and had average IoU scores that were significantly higher than all other CAM methods.

Figure 4.7 compares the CAMs produced by Full-Grad when it was used to explain the predictions of the ResNet-34 (left) and VGG-16 (right) models, which were both trained on the same dataset of manipulated images, and both achieved high maximum validation accuracy (i.e. ResNet-34: 97.5%; VGG-16: 100%). Figure 4.7 clearly demonstrates that different neural network architectures generate different class activation maps, which I would hypothesize to stem from the underlying implementation of each CAM method. As seen, the brightest red highlighting of Full-Grad's output surrounds the watermark perturbation using both models. However, using ResNet-34, a disperse checkerboard diffusion pattern is visible in the CAM produced by Full-Grad, which is not present using VGG-16. This dispersion may originate from Full-Grad's use of the CNN's layers closest to the input. In contrast, using VGG-16, Full-Grad produced CAMs that were almost exclusively highlighting only the watermark, an ideal outcome.

The superior performance and capability of Full-Grad to localize the discriminatory watermark input feature in comparison to all other methods is likely because Full-Grad is the only method that can simultaneously satisfy both weak dependence (local attribution) and completeness (global attribution). Full-Grad may therefore be the best method to further fine tune and optimize so that it can gain user trust and be of great value in a variety of application domains.

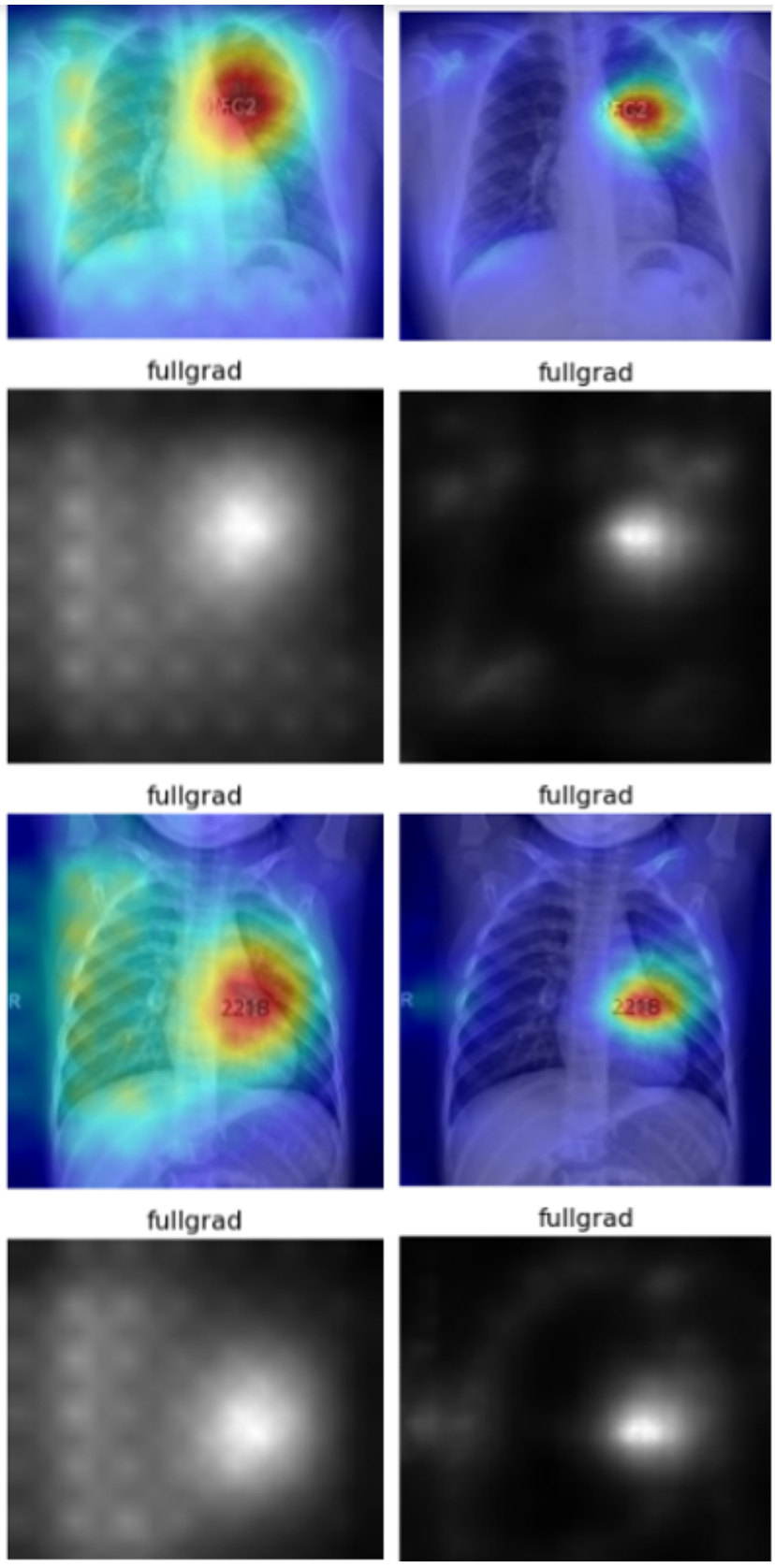

Figure 4.7: Full-Grad Comparison when Explaining Predictions of ResNet-34 (left) Vs. VGG-16 (right)

Chapter 5

Discussion, Recommendations, and Conclusion

5.1 Discussion: Importance of Automated Medical Image Interpretation

In the medical domain, deep learning convolutional neural network architectures have enabled automated medical image interpretation at a level that has often surpassed that of practicing medical experts [15]. However, the use of these "black-box" deep neural networks in clinical workflows has been severely limited and delayed due to the lack of interpretability of these models. This is unfortunate as diagnostic and possible therapeutic capabilities of artificial intelligence in medicine has been developing rapidly.

For example, one study [16] concluded that in image analysis involving the recognition of myocardial infarction on electrocardiography (ECG), deep learning using a simple CNN may achieve a comparable capability to physicians, but warranted that further investigation was necessary.

Therefore, to assist clinicians in diagnostic decision making, saliency maps are often used. However, recent work has shown that saliency methods may be misleading and lead to increased bias and user mistrust, which has had concerning implications for clinical translation efforts [15].

A study [15] performed systematic evaluation of Grad-CAM on multi-label classification models for chest X-ray interpretation. The study compared the segmentations of Grad-CAM to human expert benchmark and reference annotations, to evaluate the extent to which Grad-CAM identified critical areas of an X-ray for diagnosis, and located pathologies of concern. Their results suggested that Grad-CAM had low performance for 10 pathologies, and "performs worse for pathologies that had multiple instances, were smaller in size, and had shapes that were more complex" [15]. It therefore noted these as clinical conditions where it may be dangerous to use saliency maps compared to human experts when analyzing neural networks [15]. [15] also suggested that when the model had made a positive prediction with high confidence, saliency maps were safer for clinicians to use as a decision aid, because model confidence was shown to be positively correlated with Grad-CAM localization performance.

These studies showcase that in order to allow for the confident deployment of safe and accurate neural networks in a variety of application domains, especially medical image analysis, where they could have an immense positive impact on society, explainable artificial intelligence algorithms such as feature attribution methods must be further developed. Additionally, these explainability methods must achieve high evaluation performance through both qualitative and quantitative metrics, to gain user trust and neural network acceptance.

The implications stemming from these studies were a crucial motivator for the experiments conducted in this thesis.

5.2 Discussion Regarding Results

The dataset modification procedure implemented in this experiment on the COVID-19 Radiography Database from Kaggle imposed ground truth on the dataset, which is not possible in most real-world datasets. The label reassignment and image manipulation steps ensured that only input images manipulated with an imposed watermark were labeled as COVID-19 positive, and all other unmodified images were labeled as COVID-19 negative. This ensured that any neural network model that achieved high validation accuracy would have only made the correct prediction classification if it was concentrating exclusively on the imposed watermark perturbations, which successful class activation maps should exclusively highlight.

By imposing ground truth, the degree of success of different state-of-the-art class activation map methods at interpreting the predictions of deep learning neural network architectures could be evaluated. My experiments analyzed Grad-CAM, Eigen-CAM, Grad-CAM++, Layer-CAM, and Full-Grad for their performance at assigning feature attribution to the ground truth binary class label, for (near-) perfectly trained ResNet-34 and VGG-16 neural networks.

The experimental results suggested that on average, Full-Grad produced the sharpest, most targeted CAMs, which was seen not only qualitatively through its focused heatmaps, but also by the high average IoU scores it achieved, especially at high threshold cutoffs. Therefore, Full-Grad looks to be the most robust and accurate method at localizing the imposed watermark, a discriminatory image feature. However, when Full-Grad was used to analyze what image regions were most probably of highest importance to the classification decisions of the ResNet-34 CNN model, a disperse checkerboard pattern that covered a large portion of the image often accompanies the focused brightest region that overlapped with the watermark. However, when analyzing the predictions of the VGG-16 model, the dispersion was not present, and Full-Grad almost exclusively only gave high attribution scores to image pixels where the induced watermark was located.

One possible explanation for the differences seen may be due to the underlying mathematical implementation of Full-Grad. Full-Grad aggregates the full-gradient components across all intermediate layers of the model. Full-gradients decompose the neural net response into input feature sensitivity and per-neuron sensitivity components [9]. Therefore, all bias layers contribute to the CAM produced. Full-Grad was the only method tested that utilizes full-gradients to successfully

satisfy both local and global importance, simultaneously, by capturing the significance of both individual input pixels through input attribution, and the interaction between groups of pixels through neuron importances. This is in contrast to other CAM methods, such as Grad-CAM, which use only the last convolutional layer of the model, just prior to the linear classification layer.

Therefore, the dispersion seen only when the CAM method is analyzing the predictions of ResNet-34 and not VGG-16 may be an implication of the aggregation of activations in the earlier model layers closer to the input. Accordingly, different model architectures would be predicted to lead to different heatmap outcomes, even when the same class activation map method is used.

Furthermore, another hypothesis is that the checkerboard diffusion pattern could originate from the contribution of gradients that influence the negative prediction value. In the Full-Grad implementation used in this experiment, the absolute value of the gradient terms in all layers are used to calculate the pixel attribution scores, so the sign is ignored, which allows for visualization of only the magnitude of importance. However, if only positive gradients are used, by using a ReLU-like function to set all negative gradients to 0, this dispersion may no longer be seen.

For example, a study [17] implemented a technique of positive propagation or aggregation of gradients in gradient-based saliency methods, specifically Full-Grad and Grad-CAM++. Results showed empirically that regardless of the model and output prediction, the generated saliency maps recovered salient image features [17]. Therefore, the positive aggregation of gradients technique modified the methods towards being class-insensitive and indifferent to randomization, which are both desired features of saliency maps. Furthermore, dispersion in the saliency maps was not seen.

Finally, since the brightest (reddest) portion of the CAM heatmap produced by Full-Grad highlights the induced watermark artifact, this implied that the image pixels assigned the highest attribution scores correspond to pixels where the watermark was imposed. Therefore, I hypothesized that if the CAMs are thresholded, the dispersion could be eliminated, as any pixels with attribution scores below the threshold would not be included in the binary prediction mask, and the new CAM would be strictly highlightling the watermark.

To test this hypothesis, in each of my experiments, I thresholded the CAMs produced at thresholds between 50% and 95% at 5% intervals, and calculated the average IoU between the thresholded prediction mask and the ground truth segmentation mask. This thresholding was critical in removing the dispersion seen in the CAMs, which was evident through the superior average IoU Full-Grad scored at higher thresholds.

5.3 Recommendations for Future Experiments

Overall, the experiments conducted through my thesis demonstrate that feature attribution methods, specifically class activation maps, are a promising qualitative evaluation metric for interpreting the predictions of black-box neural network architectures. However, results demonstrate that each method may have its limitations, and that certain methods may be more advantageous to employ in

particular situations compared to other methods, due to differences in their underlying mathematical implementations.

On average, Full-Grad looked to be the most robust, accurate, and precise CAM method tested. As discussed above, in order to gain more intuition into Full-Grad performance and test if its implementation could be fine-tuned and further optimized, additional experiments could be conducted. For example, different convolutional neural network architectures could be trained on the same dataset and Full-Grad success at explaining the predictions of each model at different thresholds could be compared. Additionally, as opposed to taking the absolute value of the gradient terms in all layers to calculate the pixel attribution scores, a ReLU-like function could be used to set all negative gradients to 0, which would adapt the implementation of Full-Grad to only consider positive gradients when assigning pixel attribution scores. This modification may help to limit the dispersion seen when Full-Grad was used to explain the classification predictions of ResNet-34.

Finally, each CAM method could be further tested for their performance in important computer vision applications such as object detection, semantic segmentation, and embedding-similarity. If extensive testing of CAMs is conducted, and their implementations are further fine-tuned to optimize their performance at diagnosing model predictions, this will help to lead to enhanced user trust in these methods. This trust would allow CAMs to be used more in real-world situations for the interpretability of predictions of both deployed models and models in production, where their explanations could be of immense value.

5.4 Conclusion

The emergence of deep neural networks (DNNs) is occurring rapidly. However, DNNs are currently unable to explain their predictions to humans in an interpretable, transparent, and trustworthy way, and are also susceptible to adversarial attacks, due to their underlying "black-box" nature.

If end-users could trust that DNNs deployed in real-world, mission-critical applications and high reliability systems will consistently make successful, safe, and unbiased predictions, widespread DNN deployment could the lead to revolutionary progression across a diverse range of application domains. Therefore, interpretable machine learning is essential to enable end-users to understand, appropriately trust, and effectively manage these DNNs.

This crucial user trust could be accomplished through explainability algorithms, if they are proven to be intuitive, versatile, and accurate. This would require explainability algorithms to be both sensitive to malicious adversarial attacks and simultaneously capable of generating robust, reproducible, and replicable DNN predictions explainations.

Post hoc gradient-based, feature attribution methods are a type of explainability algorithm. They are a set of mathematical operations with certain assumptions that attempt to explain the internal workings and prediction-making process of uninterpretable and non-transparent, "black-box" DNNs, by exclusively highlighting only the input features of strongest influence to the DNN's prediction.

Unfortunately, feature attribution methods contribute an extra layer of abstraction, so when analysing explanations, it is unclear "how to disentangle errors in the explanation method from errors in the DL model". Consequently, no consensus has been reached regarding which attribution methods are best at interpreting network predictions [1].

These limitations and vulnerabilities raise questions regarding the trustworthiness of explanations produced by feature attribution methods. It is therefore essential that in real-world, mission-critical applications and high reliability systems, attribution methods are used with caution, as they can create a false sense of confidence in DNN predictions, by producing explanations that seem convincing, but may be inaccurate [1].

Attribution methods must be capable of detecting a model's use of shortcuts, which are distinctive, easy–to–learn input features, such as unintentionally added input perturbations and artifacts, that are not necessarily essential to the problem being solved, and could therefore cause spurious output correlations.

If the success of explainability algorithms can be proven, they will be crucial in enabling users to trust that deployed DNNs will consistently make successful, safe, and unbiased predictions in real-world, mission-critical applications.

However, until model interpretability is achieved through development of successful evaluation metrics and explainability algorithms such as feature attribution methods, the risks associated with the unknown possibility of catastrophic outcomes resulting from the deployment of DNNs will restrict their implementation in several application domains.

My thesis therefore attempted to evaluate the success of feature attribution methods at explaining the predictions made by DNNs in medical image classification, and to gain further insight and intuition regarding the performance of these methods.

I hope that the extensive qualitative and quantitative analysis that I have conducted, along with my recommendations for future tests, will help to answer the fundamental question of how explainability algorithms can be evaluated for their robustness, reproducibility, replicability, and sensitivity to malicious adversarial attacks.

In particular, the experiments of my thesis investigated the success of class activation map methods, specifically Grad-CAM, Grad-CAM++, Layer-CAM, Eigen-CAM, and Full-Grad, at giving attribution to the ground truth, for near-perfect deep learning neural networks, in their attempts to explain the DNN's predictions. This examination was accomplished through a dataset modification procedure that imposed ground truth on the COVID-19 Radiography database by reassigning labels for binary classification, where only images manipulated with a watermark perturbation were assigned to the COVID-19 positive class. This dataset modification procedure therefore ensured that any neural network model that achieved high validation accuracy would have only made the correct prediction classification if it was concentrating exclusively on the imposed watermark perturbations, which successful class activation maps should exclusively highlight. Results demonstrated that Full-Grad appears to be the most robust, precise and accurate method at localizing discriminatory

image features and detecting input perturbations, and that it's performance could be optimized by thresholding it's output to remove dispersion.

These results will hopefully serve as an important step towards developing successful interpretability algorithms, which are essential to gaining user trust and confidence in use of deep neural networks in real-world, mission-critical applications and high reliability system where DNNs have the potential to make revolutionary breakthroughs.

References

[1] I. E. Nielsen, D. Dera, G. Rasool, N. Bouaynaya, and R. P. Ramachandran, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," 2021. DOI: 10.48550/ARXIV.2107.11400. [Online]. Available: https://arxiv.org/abs/2107.11400.

[2] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," 2020. DOI: 10.48550/ARXIV.2005.13799. [Online]. Available: https://arxiv.org/abs/2005.13799.

[3] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in Sep. 2019, pp. 563–574, ISBN: 978-3-030-32235-9. DOI: 10.1007/978-3-030-32236-6_51.

[4] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?," 2021. DOI: 10.48550/ARXIV.2104.14403. [Online]. Available: https://arxiv.org/abs/2104.14403.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019. DOI: 10.1007/s11263-019-01228-7. [Online]. Available: https://doi.org/10.1007%2Fs11263-019-01228-7.

[6] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, Mar. 2018. DOI: 10.1109/wacv.2018.00097. [Online]. Available: https://doi.org/10.1109%2Fwacv.2018.00097.

[7] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021. DOI: 10.1109/TIP.2021.3089943.

[8] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2020. DOI: 10.1109/ijcnn48605.2020.9206626. [Online]. Available: https://doi.org/10.1109%2Fijcnn48605.2020.9206626.

[9] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," 2019. DOI: 10.48550/ARXIV.1905.00780. [Online]. Available: https://arxiv.org/abs/1905.00780.

[10] M. E. H. Chowdhury, T. Rahman, A. Khandakar, *et al.*, "Can ai help in screening viral and covid-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132 665–132 676, 2020. DOI: 10.1109/ACCESS.2020.3010287.

[11] T. Rahman, A. Khandakar, Y. Qiblawey, *et al.*, "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in Biology and Medicine*, vol. 132, p. 104 319, 2021, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104319. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252100113X.

[12] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, AAAI, Feb. 2022.

[13] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[14] E. Tjoa and C. Guan, "Quantifying explainability of saliency methods in deep neural networks," *CoRR*, vol. abs/2009.02899, 2020. arXiv: 2009.02899. [Online]. Available: https://arxiv.org/abs/2009.02899.

[15] A. Saporta, X. Gui, A. Agrawal, *et al.*, "Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation," *medRxiv*, 2021. DOI: 10.1101/2021.02.28.21252634. eprint: https://www.medrxiv.org/content/early/2021/03/02/2021.02.28.21252634.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2021/03/02/2021.02.28.21252634.

[16] H. Makimoto, M. Höckmann, T. Lin, *et al.*, "Performance of a convolutional neural network derived from an ecg database in recognizing myocardial infarction," *Scientific Reports*, vol. 10, p. 8445, May 2020. DOI: 10.1038/s41598-020-65105-x.

[17] A. Khakzar, S. Baselizadeh, and N. Navab, "Rethinking positive aggregation and propagation of gradients in gradient-based saliency methods," *CoRR*, vol. abs/2012.00362, 2020. arXiv: 2012.00362. [Online]. Available: https://arxiv.org/abs/2012.00362.