Bridging Accuracy, Interpretability, and Robustness in Machine Learning Models: A Study Using Grad-CAM Across Training Stages

Aarush Arya

Received August 26, 2024
Accepted October 27, 2024
Electronic access November 15, 2024

As machine learning models grow in complexity, their decision-making process becomes increasingly opaque, leading to significant challenges in understandability and trust. Furthermore, many complex machine learning models lack resilience when exposed to new or corrupted data, leading to countless issues in the real world. This paper explores interpretability, or the degree to which we can understand the decisions behind a model's outputs, as well as robustness, or the model's adaptivity to changes in data. Specifically, this paper focuses on how these aspects interlap and evolve during different training stages of models. Utilizing Grad-CAM (Gradient-weighted Class Activation Mapping), a tool to visually understand decisions in image classification, we compare the interpretability of models as training progresses and clearly present the resulting heatmaps in the image recognition tests. We also generate perturbations in images to analyze how the robustness changes alongside accuracy. Upon studying the relationships between accuracy, interpretability, and robustness, the findings highlight key trends, unexpected results, and outliers in the Grad-CAM images. Our results demonstrate that although the interpretable features grow more consistent as the model's accuracy improves, there are several instances where a high accuracy may not necessarily mean a model is robust or interpretable - highlighting the importance of focusing on these aspects of machine learning models alongside their ability to correctly classify objects. Ultimately, this work aims to bridge existing gaps regarding the overlap between described characteristics within models, therefore contributing to the development of more reliable, resilient, and explainable machine learning systems.

Introduction

Over the past few years, Artificial Intelligence (AI) and Machine Learning (ML) have become a staple in our everyday lives, whether casually or in various industries, to optimize workflow[1]((Dietterich TG. Machine Learning. Annual Review of Computer Science. 4, 255-306 (1990).)). Machine learning (ML) refers to the field of study of developing systems for which AI models learn about the world from data and train to perform the tasks we give them[1]. The process of ML has grown from broad statistical analysis to being able to include complex tasks and training methods that were once thought to require human intelligence. Such tasks Some of the tasks which AI can help optimize include but are not limited to complex automation, natural language processing, and the task which this paper as well as countless other research focuses on, image recognition. There is no doubt that the uses of ML and its accomplishments are incredible and growing to be even more powerful than before. However, as the capabilities of ML models expand , there lies a fundamental issue that becomes increasingly prominent. These models they are becoming more increasingly opaque and unpredictable, and are known as Black Box Models[2]. As ML becomes more sophisticated, we lack a thorough understanding of how such models make decisions: the key issue in ML. This is because as they grow more complex, their decision-making process becomes increasingly out of our scope of understanding. Rather than being completely blind to the decision-making process, it is of utmost importance that we decipher what exactly is going on behind the scenes. Information about the inner workings of these models can be used to debug errors, reduce bias within AI, ensure compliance with ethical standards, and give scientists valuable insight from the strategies of talented powerful AI models. As it is clear how the opaqueness of models can cause issues, it is consequently obvious that we need to develop methods to understand the decision-making process. The degree to which we humans can understand the decision-making process of a machine-learning model is called interpretability[3].

Interpretability techniques are methods used to measure how ML models make their decisions, demystifying the curtain of the model's brains and making them much more transparent, understandable, and usable to humans in countless fields of study. The growing complexity of models has increased demand for such methods and the field of ML has delivered many such techniques. One of the most prominent ones, which is the technique that is used throughout this research paper and countless other studies, is Grad-CAM, also known as Gradient-weighted Class Activation Mapping[4]. Grad-CAM makes it possible for us to visually understand the decision-making process of AI models in image classification by depicting which

regions of an image are the most important for a model's prediction. This method uses gradients of a target class or image with respect to the feature maps of a convolutional layer to produce a heatmap that essentially highlights the most important regions in an image that the model uses for classification[4]. This technique is groundbreaking in the way that it gives us an intuitive explanation of what the model sees when it makes decisions.

While interpretability is crucial to understand the decision-making process of ML models, it is equally if not more important to measure the resilience of these models in the face of changes in the input data. This characteristic, which has been subject to countless tests and continual improvement, is called robustness. Robustness refers to the ability of a machine-learning model to handle perturbations or noisy changes in the input data[5]. A common issue with a lot of AI is that they lack this characteristic - or in other words, minor changes in input data can lead to very significant differences in the model's explanations or outputs[5].

While interpretability and robustness have been independently explored in numerous studies, there is a significant research gap in understanding how these two critical attributes interact within different stages of a model's training. This paper aims to address this gap by exploring trends and gaining insights into the relationships between robustness, interpretability, and accuracy of machine learning models through various experiments. Specifically, this gap is addressed through the usage of Grad-CAM in order to analyze the predictions of a CNN (Convolutional Neural Network) and perform a qualitative analysis on the interpretable heatmaps. We perform additional analysis to determine the robustness of the interpretability method to various image perturbations and to see how interpretability methods like Grad-CAM can help us understand examples that the model misclassifies. Ultimately, we care about the intersection between robustness, interpretability, and accuracy because if we want ML models to be robust, we also need to make sure that they can make classifications with great accuracy. We also want to see whether interpretability methods can help us understand a model's performance on robustness classification tasks in addition to standard tasks.

Our contributions are presented as follows.

1. We obtain and evaluate Grad-CAM images on models of different accuracy levels and demonstrate findings on the intersection between interpretability and accuracy.

2. We study Grad-CAM heatmaps of said models under image corruptions (in the form of Gaussian Noise, rotations, and occlusions) to demonstrate how interpretability methods can help deepen our understanding of the relationship between robustness and accuracy.

3. We compare Grad-CAM heatmaps of cases where models incorrectly classify images to obtain insights into common errors made by ML models and provide suggestions for the future of the field.

Literature Review

Interpretability of ML and Techniques

The lack of transparency arising as ML models grow more advanced has proven to have severe consequences: it can result in unintended biases, reduce trust in AI systems, and hinder their adoption in critical fields such as healthcare and finance, where understanding the rationale behind decisions is essential for ethical and accurate outcomes[6]. For example, in fields such as healthcare, understanding how AI comes to a diagnosis is as important as the diagnosis itself[6]. In fields such as finance, explainable models help in clarifying decisions to customers, therefore building trust and transparency[7]. Furthermore, it is necessary to develop metrics that can measure how interpretable, or explainable, an AI model is to humans[2].

As discussed, Grad-CAM is a powerful tool which allows us to visually understand what ML models see when making predictions regarding image classification. However, there are many other interpretability techniques worth understanding. Another similar technique for visualizing explanations, although less explored in this paper, is Saliency Maps, which also highlight key regions for decision-making in an image - however by computing the derivative of an input with respect to a specific network output[8]. Another interpretability technique, LIME (Local Interpretable Model-agnostic Explanations), explains individual predictions by creating an interpretable model which approximates the model's behavior around that prediction, then perturbing the input data and examining changes in the output to determine which features are most influential for the prediction[9]. Yet another approach is the SHAP (SHapley Additive exPlanations) technique, which assigns an importance value to each feature in a particular prediction; it then considers all possible combinations of features and averages their contributions to figure out how much each feature contributes to the final prediction[10]. Grad-CAM, however, remains widely used way to visualize model explanations, largely due to it giving a very intuitive and easy-to-understand visual explanation. This is consistent with research that proves that an increase in the complexity of explanations reduces human satisfaction[11]. Although there are a plethora of other interpretability techniques ranging in complexity, this research paper primarily focuses on using Grad-CAM due to its exceptional explainability in image recognition and simplistic understanding.

Applications of methods

The heatmaps from Grad-CAM, which show which parts of an image the model uses to make decisions, undoubtedly have

countless practical uses in several domains. For example, it is the most effective in generating heat maps of brain areas that are relevant and related to sclerosis, a chronic disease of the nervous system [12]. By knowing what areas of the brain are related to this disease, clinicians are not only able to further correct models to eliminate bias but also focus their own criteria for sclerosis treatment and detection in line with the model's [12]. Grad-CAM has also proven to be the strongest at explaining how models classify COVID-19 from input X-ray images of a patient's chest, which is valuable information for scientists and doctors [13].

There are also several practical applications of this technique outside the medical field. For example, Grad-CAM allowed scientists to understand sounds inside beehives by distinguishing between "bee" and "nonbee" sounds and highlighting time intervals of specific frequencies of an input image [6]. As interpretability methods rise and prove their significance, so do the approaches to incorporate interpretability in the general ML process everywhere. One approach is CRISP-ML (Cross Industry Standard Process for Data Mining - ML) which revises and updates models to ensure that interpretability is embedded into the model's actions, creating trust and transparency [14]. Ultimately, it is evident that it is just as important to have explainable models as it is to have well-functioning models. This paper explores the importance of interpretability in the form of Grad-CAM heatmaps throughout all models of different levels of training.

A Closer Look into Robustness

Robustness, or the ability of a model to handle perturbations in the input data, is a highly valuable characteristic in real world scenarios. Having models that are robust are crucial for deploying ML systems in the real world, where data is often very imperfect and unpredictable [5]. Robustness ensures reliability, ensuring that models perform consistently in different scenarios - this characteristic is a necessity in many AI domains, especially autonomous driving and healthcare. It is important to be able to test the robustness of several models in these important domains alongside training them - for example, testing that of COVID-19 genome classification models by inputting very noisy data [15]. This research found that the AI was more adaptable when faced with certain types of errors in genome sequences compared to others, allowing targeting improvement of the ML models concerning such errors [15]. Robustness also ensures security, making models well-protected and able to adapt to possible adversarial attacks that disrupt or manipulate input data to cause false predictions (which can have severe consequences in several industries). Many techniques have been developed and popularized to test this core necessity of ML models, for example, FoolBox: a Python package that simulates attacks against input data in forms such as gradient or pixel manipulation [16]. This is one example of training a model with corrupted inputs, meant to deceive the AI; and data augmentation which helps the model learn how to classify diverse and nonlinear inputs. Ultimately, the goals of such methods are to find scenarios where slight disturbances can lead to huge unwanted changes in the model's predictions (i.e. severe misclassification). Another concept that needs consideration is the trend between the accuracy of a model in decision-making/classification and the robustness of the model. It is evident amongst countless ML models that there is a tradeoff between robustness, or adaptivity, with accuracy. The cause of this unwanted tradeoff, rather than being inherent, is a visible consequence of many current methods used to train models [17]. During training, models are fit to the training dataset, and may ultimately be less robust to novel data encountered at test time. To achieve a successful model with strong accuracy and the ability to be robust, we must ensure that training methods are modified to reduce the large accuracy gap between standard and corrupted data while preserving the model's accurate classification of standard images [17].

Objectives

Previous studies have thoroughly explored methods of interpretability in ML models, such as Grad-CAM, and methods of testing the robustness of these models. However, there are still gaps in the field that need further research. These gaps involve a lack of comprehensive evaluation that compares and analyzes the differences in both the explanations and the robustness of ML models concerning the amount of training the models have. To address these gaps, we compared the Grad-CAM heatmaps of models at different stages of training and showed that as accuracy improves, the heatmaps begin to focus on the most human relevant components of the image. We study the Grad-CAM heatmaps of said models when subjected to several image corruptions to demonstrate not only how robustness improves at different training stages but also how interpretability methods can help deepen our understanding of the relationship between robustness and accuracy. We further compare the Grad-CAM heatmaps of when models fail to classify images, allowing us to gain valuable insight into common errors made by ML models and provide suggestions for the future of the field. By knowing which parts of the image are most important for a model which incorrectly classified the image, we can help train models away from those mistakes and strengthen their accuracies and explanations. By achieving these objectives, this research ultimately aims to provide insight and bridge the gap on how the interpretability and robustness of ML models evolve alongside their accuracy; and how explanations of the model differ alongside changes in accuracy and the addition of distortions in the input data. My research will offer an understanding of the relationships and interactions between a model's interpretability, robustness, and accuracy: ultimately

contributing to the development of more transparent and reliable machine-learning models.

Methods

This study investigates the interpretability and robustness of ML models, specifically when tasked with image recognition, across several stages of training. There are various important components of the methodology to do so.

Data

The data needed for this research is generic images of various categories. For this research, we will use the preexisting CIFAR-10 dataset, a popular and reliable dataset for image recognition. The CIFAR-10 dataset is composed of 60,000 different images divided into 10 categories: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. The data was split up with 50,000 images going into the training data set and 10,000 into the test data set.

Model Selection

The EfficientNet-B0 ML architecture[18], which is a CNN, (Convolutional Neural Network) was selected for the models in this research due to its exceptional performance in image recognition tasks. Furthermore, EfficientNet-B0 is small in size and therefore relatively efficient to run and test multiple times. The EfficientNet-B0 architecture was loaded using the timm library, a deep learning library consisting of several ML models and architectures. The models were trained on the CIFAR-10 train data set using stochastic gradient descent as the optimization method and cross-entropy loss as the loss function. Subsequently, the models were subjected to tests on the CIFAR-10 test data set, shuffled each test, to ensure proper image classification and avoid bias in recognition. As the research aims to examine differences in interpretability and robustness over various accuracy levels, model checkpoints were saved every epoch in the main training loop of 20 epochs to save the current state of the model. The interpretability and robustness tests were conducted on an untrained model, a model after 1 epoch of training, a model after 2 epochs of training, a model after 3 epochs of training, and a model after 20 epochs of training (which has approximately 99.97% train accuracy and 95% test accuracy).

Interpretability Analysis using Grad-CAM

To analyze the interpretability of ML models, this research employs Grad-CAM, which offers visual explanations of a model's prediction in the form of heat maps that show which areas of an image are most influential for the model's prediction[4]. Grad-CAM works by using the gradients (changes) of the target class score with respect to the features in the last convolutional layer of the CNN.

First, the input image is passed through the network to get the class scores. These scores represent how confident the model is about each class - for example, if the model predicts a cat, the score for the cat class is obtained. Afterward, gradients are calculated for the score of the class of interest with respect to the feature maps, which are modified representations of the input image to highlight its key features, in the last convolutional layer. These gradients show how much each feature map affects the class score and is therefore a critical step in depicting which areas or features of the image affect the final prediction. These gradients, or changes, can be represented by:

$$\frac{\partial y^c}{\partial A^K} \qquad (1)$$

where y^c is the score for the class c and A^k are the feature maps. The next step for Grad-CAM is averaging the gradients over all locations in the feature maps to get a single importance weight for each feature map. These weights tell us how important each feature map is for the class of interest, allowing the eventual creation of a heatmap.

$$a_c^k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \qquad (2)$$

Where Z is the total number of pixels in the feature map, and $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient of the score for class c , as discussed above, at the location (i, j) . The feature maps are then combined using the importance weights a_c^k to create a class activation map, highlighting the important regions of the image:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k a_k^c A^k\right) \qquad (3)$$

The ReLU (Rectified Linear Unit) function ensures that only the positive influences are considered. Finally, the class activation map is resized to match the original image and normalized to produce the heatmap, visually representing the important features of an image.

This research, employs five checkpoints of a ML model trained to various stages (0, 1, 2, 3, 20 epochs). Throughout the experiment, we choose to measure the performance of the model at these epochs because there is a large improvement in accuracy from the 1st to 2nd and from the 2nd to 3rd epoch (as exemplified by Table 1 below). This in turn allows for model performance to significantly differ between models of said epochs and for the gap in accuracy to be substantial enough to explain and attribute the differing results of the interpretability and robustness tests. Furthermore, as shown in Table 1, the improvement in accuracy begins to plateau as the number of epochs increases close to 20.

Our final model checkpoint was chosen to be at 20 epochs of training as it is a strong and testable baseline for a highly trained model and the difference in performance or results for models trained beyond 20 epochs would be minimal. The significant difference in accuracy between the 3 and 20 epoch model would allow us to attribute the differences in the results to the large accuracy gap between epochs.

Grad-CAM was applied to each of these checkpoints to generate heatmaps on various images in the CIFAR-10 dataset. This process enabled easy examination and analysis of the level of interpretability, through the Grad-CAM heatmaps, of each model checkpoint (ranging from untrained to fully trained). The heatmaps were then visually inspected and qualitatively analyzed to assess changes in interpretability over changes in accuracy and identify any trends in the data. Furthermore, this research intentionally scrutinizes specific and rare cases where models from one to three epochs fail to correctly classify the image. This is done by continuously testing the one, two, or three epoch model on the CIFAR10 test data set, which is shuffled each test, until the model incorrectly classifies an image. Then, we obtain the Grad-CAM heatmaps for such images and later qualitatively investigate differences in those results compared to heatmaps where the model correctly classified an image.

Robustness Methods

To test the robustness of the ML models at different checkpoints of the training process, several techniques were employed to generate perturbations, or corrupted changes, in the input data. The techniques used were Gaussian noise, which generates changes in the individual pixels leading to the images in the dataset being slightly blurry and static (done with a standard deviation of 0.1); rotations, in which the images were slightly rotated by a random amount from -30 to 30 degrees; and occlusions, in which twenty random pixels of each image, possibly at critical locations for the model prediction, were modified to be occluded and undetectable. All perturbations were applied separately. After applying a perturbation to a test image, Grad-CAM heatmaps were once again generated for each model checkpoint. These resulting heatmaps, for each model checkpoint, were examined alongside heat maps generated from unchanged images. To ensure consistency, we compare heatmaps from clean and perturbed versions of the same image. The resulting heatmaps were then analyzed to identify trends in how each model adapted to changes in the data, ultimately determining how the robustness of ML models change alongside accuracy.

Finally, we utilize a ML model's confidence drop, a quantitative metric used to numerically evaluate the robustness of ML models [19]. For a given input image, the model generates predicted probabilities for each class using a softmax function. The confidence score refers to the maximum predicted probability for the model to predict a correct class [19]. This is then recomputed similarly, except when the model processes the perturbed or distorted image.

In our case, the confidence drop represents the initial confidence score of the model, when subjected to normal data, minus the confidence score of the model when subjected to perturbed data (we utilize random occlusions for this value). The confidence drop when subjected to 10 random occluded pixels is calculated for each of the ML models and gives us quantifiable insight into how robustness changes alongside accuracy. A decreasing confidence drop indicates that the model is becoming more robust as accuracy improves, while an increasing confidence drop indicates that the model is becoming less robust (as the new confidence level is significantly lower than the original).

Qualitative Analysis

To measure how both interpretability and robustness change alongside accuracy, our analysis focuses on several qualitative techniques. We examine the shifts in the focus area of the model, analyzing how the highlighted regions as depicted by the Grad-CAM heatmaps change as accuracy improves. To qualitatively measure robustness, we analyze and evaluate consistency in the heatmaps when subjected to image perturbations across all the model checkpoints. We also compare the Grad-CAM outputs for both instances where models correctly and incorrectly classify the image, in which we analyze patterns in the heatmaps of incorrect classification - which will ultimately help us in determining common sources of error for ML in image classification.

Software and Tools Used

All experiments were conducted using Python 3.8. Google Colab was used as a platform for the code and experiments. The PyTorch library was used for training, loading, and saving the models at various checkpoints using workflow interaction with Google Drive. The timm library was used to load the EfficientNet-B0 model architecture. Several external libraries were used to generate and display Grad-CAM images: NumPy was used for processing gradients, OpenCV for applying heat maps onto images, and matplotlib for displaying the Grad-CAM results in a visually organized fashion. To test the robustness of models, torchvision was used to aid in generating image perturbations such as Gaussian noise, rotations, and occlusions. By employing these methods, this research aims to provide a thorough analysis to aid in bridging the gap between interpretability, robustness, and accuracy.

Results

In this research, we investigated differences in interpretability and robustness within 5 models with different accuracy/training levels. This section covers the results of the experiments done.

Interpretability Tests: Overview

Our first test explores the level of interpretability, in the form of Grad-CAM heatmaps, amongst these 5 model checkpoints (untrained ; 1, 2, and 3 epochs of training; fully trained). We perform experiments on the untrained (0 epoch) model in order to demonstrate the change in interpretability and robustness from a model with no experience (where the results are random and unorthodox) to a model with some training experience. We qualitatively analyze these heatmaps to observe the relationship between interpretability and accuracy. All experiments were performed with an EfficientNetB0 architecture (reference Section 3.2: Model Selection). In each test, we gather the Grad-CAM heatmaps for the 5 models on the same image (randomly selected from the CIFAR-10 dataset). We also investigate the Grad-CAM heatmaps of different scenarios of when a model incorrectly classifies an unchanged input image. We use the 1, 2, and 3 epochs-trained models for this test as they have strong but not perfect accuracy, making them valuable subjects.

For reference, the approximate accuracies of the model checkpoints which are trained to a certain extent (1, 2, 3, and 20 epochs) are shown below in Table 1. Table 1 includes both accuracy when training on the CIFAR10 train set and accuracy when testing on the CIFAR10 test set. Accuracies of several other checkpoints between 3 and 20 epochs are also depicted (however, note that as the amount of epochs rise closer to 50, the level of performance remains fairly constant and therefore are not included in the experiment).

Table 1. Reference accuracies of the model checkpoints ranging from 1 to 20 epochs on both train and test data sets. The highlighted epochs represent the models subject to the experiments in this research. Performance closer to 20 epochs remains fairly constant.

Important Epoch	Train Accuracy	Test Accuracy
1	74.21	87.78
2	91.22	90.91
3	94.56	91.93
4	96.22	92.79
5	97.52	92.98
10	99.5	93.71
15	99.85	94.06
20	99.91	94.33

Interpretability Tests in Various Checkpoints

Below are the results of several interpretability tests, each test visually comparing Grad-CAM heatmaps of the model checkpoints using the EfficientNetB0 architecture. In each of the following tests, all the models besides the untrained model correctly identify the image. Later in this section, this research explores cases where even trained models misclassify images and analyzes possible challenges faced through Grad-CAM.

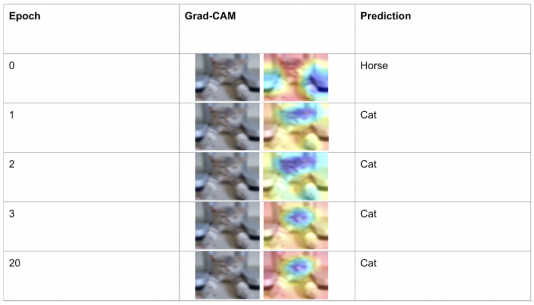

Fig. 1 Grad-CAM Analysis of CIFAR-10 Image (Index: 950, Cat) Across Different Training Stages. The epoch number is shown on the left column, the prediction of the model is shown on the right column, and the Grad-CAM images are shown in the middle, which illustrate how the focus of the model shifts as accuracy improves.

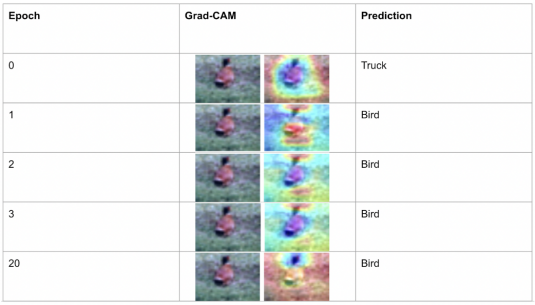

Fig. 2 Grad-CAM Analysis of CIFAR-10 Image (Index: 799, Bird) Across Different Training Stages. The epoch number is shown on the left column, the prediction of the model is shown on the right column, and the Grad-CAM images are shown in the middle, which illustrate how the focus of the model shifts as accuracy improves.

Grad-CAM images of Incorrectly Classified Cases

In this section, we investigate the Grad-CAM images where trained models, specifically those with a moderate amount of training (which are the 1, 2, and 3 epoch models), incorrectly classify images. Our goal is to identify key differences between the Grad-CAM outputs of these cases vs. cases where they correctly classify the images, allowing us to use the heatmaps to analyze why the model makes mistakes when it does.

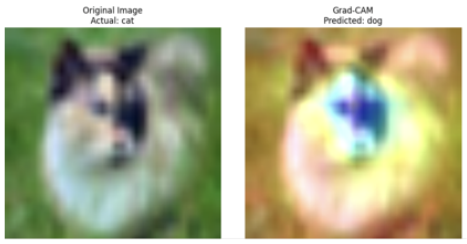

Figure 7. Grad-CAM Analysis of CIFAR-10 Image (Cat) for the 1 epoch-trained model, where it misclassified the image. The model classified the image to be a dog, whereas in reality, the image was a cat. The Grad-CAM image illustrates the focus of the model when it made the incorrect classification.

Figure 8. Grad-CAM Analysis of CIFAR-10 Image (Dog) for the 1 epoch-trained model, where it misclassified the image. The model classified the image to be a deer, whereas in reality, the image was a dog. The Grad-CAM image illustrates the focus of the model when it made the incorrect classification.

Figure 9. Grad-CAM Analysis of CIF AR-10 Image (Bird) for the 1 epoch-trained model, where it misclassified the image. The model classified the image to be a deer, whereas in reality, the image was a bird. The Grad-CAM image illustrates the focus of the model when it made the incorrect classification.

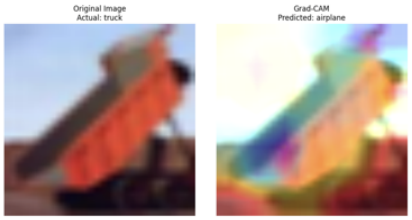

Figure 10. Grad-CAM Analysis of CIFAR-10 Image (Bird) for the 2 epoch-trained model, where it misclassified the image. The model classified the image to be an airplane, whereas in reality, the image was a truck. The Grad-CAM image illustrates the focus of the model when it made the incorrect classification.

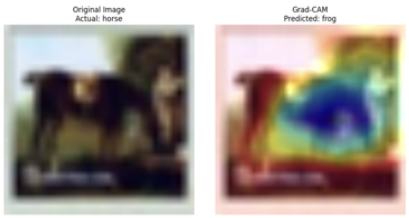

Figure 11. Grad-CAM Analysis of CIFAR-10 Image (Horse) for the 2 epoch-trained model, where it misclassified the image. The model classified the image to be a frog, whereas in reality, the image was a horse. The Grad-CAM image illustrates the focus of the model when it made the incorrect classification.

Figure 12. Grad-CAM Analysis of CIFAR-10 Image (Dog) for the 3 epoch-trained model, where it misclassified the image. The model classified the image to be a frog, whereas in reality, the image was a cat. The Grad-CAM image illustrates the focus of the model when it made the incorrect classification.

Robustness Testing Overview

We also analyze differences in the Grad-CAM images when faced with perturbations in the input data. This is done in the same way as the interpretability testing, except for the fact that the input images are modified. The three modifications we use are a) Gaussian noise, which distorts individual pixels using a standard deviation of 0.1; b) rotations of the image ranged from -30 to 30 degrees; and c) random occlusions (pixel blurs) in the image (for this experiment, 5 pixels of random location were blurred). Similar to the interpretability tests, we use the EfficientNet-B0 architecture as a baseline for most of the tests but we also conduct one test on the ResNet-18 architecture (reference Section 3.2: Model Selection). In summary, these tests allow us to explore how robustness scales with model accuracy. We use the same set of images as used for the non-changed interpretability tests, which enables easy comparison of the heat maps to see the model's robustness level.

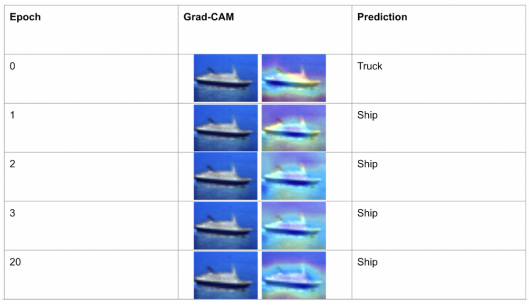

Fig. 3 Grad-CAM Analysis of CIFAR-10 Image (Index: 2927, Ship) Across Different Training Stages. The epoch number is shown on the left column, the prediction of the model is shown on the right column, and the Grad-CAM images are shown in the middle, which illustrate how the focus of the model shifts as accuracy improves.

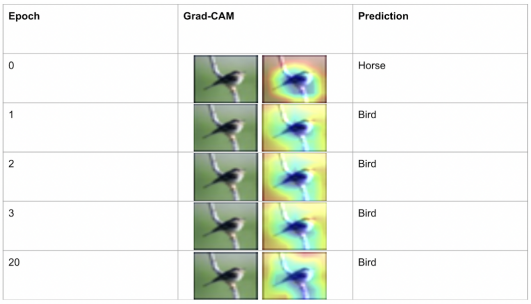

Fig. 4 Grad-CAM Analysis of CIFAR-10 Image (Index: 3333, Bird) Across Different Training Stages. The epoch number is shown on the left column, the prediction of the model is shown on the right column, and the Grad-CAM images are shown in the middle, which illustrate how the focus of the model shifts as accuracy improves.

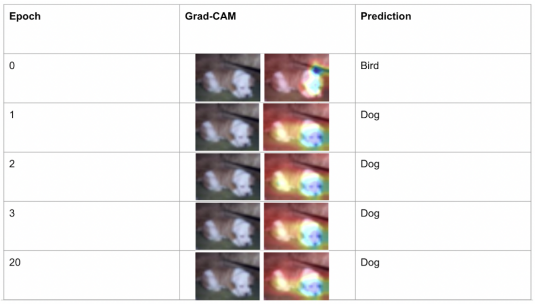

Fig. 5 Grad-CAM Analysis of CIFAR-10 Image (Index: 1244, Dog) Across Different Training Stages. The epoch number is shown on the left column, the prediction of the model is shown on the right column, and the Grad-CAM images are shown in the middle, which illustrate how the focus of the model shifts as accuracy improves.

Robustness Testing in Various Checkpoints with the EfficientNetB0 architecture

Below are the Grad-CAM results of the robustness test where a sample image was distorted using Gaussian noise, blurring each pixel with a standard deviation of 0.1. Each test visually

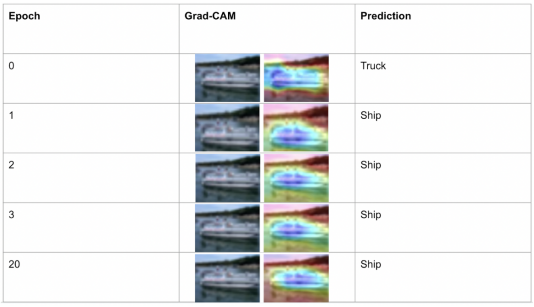

Fig. 6 Grad-CAM Analysis of CIFAR-10 Image (Index: 722, Ship) Across Different Training Stages. The epoch number is shown on the left column, the prediction of the model is shown on the right column, and the Grad-CAM images are shown in the middle, which illustrate how the focus of the model shifts as accuracy improves.

compares the Grad-CAM of the model checkpoints (using the EfficientNetB0 architecture) to analyze the robustness of the model's explanations when faced with corrupted data, specifically image noise. Three rounds of Grad-CAM images were taken for each model checkpoint, which allows us to view consistency with the explanations.

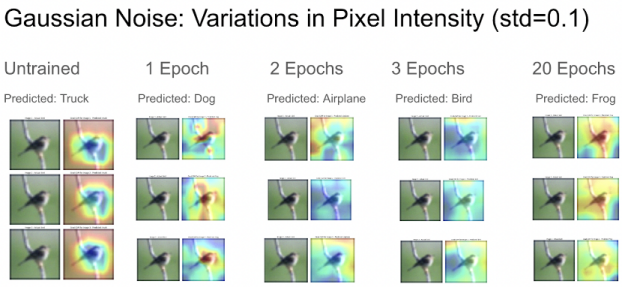

Figure 13. Grad-CAM Analysis of CIFAR-10 Image (Index: 3333, Bird) Across Different Training Stages when subjected to Gaussian Noise Perturbations. All models incorrectly classified the image besides the model at 3 epochs - truck, dog, airplane, and frog respectively. The Grad-CAM images illustrate how the focus of the model shifts as accuracy improves.

Next, we present the Grad-CAM results of two robustness tests, where in each test the input image was randomly rotated between -30 and 30 degrees. This test uses the same model checkpoints and architecture as the gaussian noise test and the results are presented identically.

Our final robustness test is the random occlusion test, where in each test the input image faces twenty pixels of random location being blurred - resulting in a significant portion of the image being unidentifiable by ML models. This test uses the same model checkpoints and architecture as the gaussian noise and rotation tests and the results when the models are presented below.

Finally, for the Random Occlusion robustness tests (see Figure

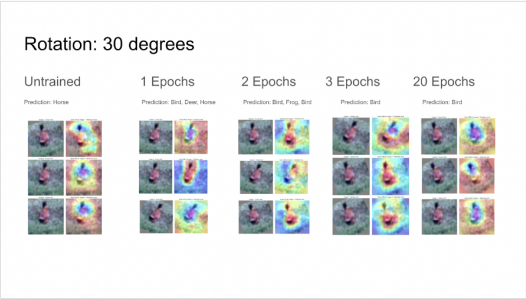

Figure 14. Grad-CAM Analysis of CIFAR-10 Image (Index: 799, Bird) Across Different Training Stages when subjected to random rotations. The untrained model incorrectly predicted the image as a horse. The 1 and 2 epoch models both varied in their predictions; the 1 epoch model predicted a bird, deer, and horse (being correct 1 out of 3 times); the 2 epoch model predicted a bird, frog, and bird (being correct 2 out of 3 times.) The 3 epoch and 20 epoch model both correctly classified the image all times. The Grad-CAM images illustrate how the focus of the model shifts as accuracy improves.

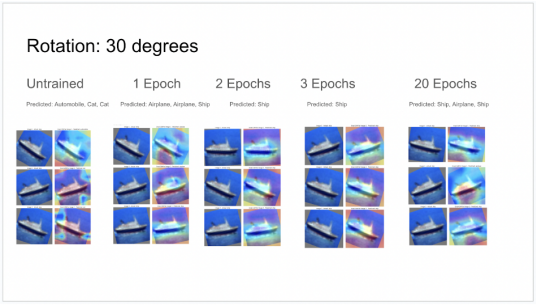

Figure 15. Grad-CAM Analysis of CIFAR-10 Image (Index: 2927, Ship) Across Different Training Stages when subjected to random rotations. The untrained model incorrectly varies in its predictions, predicting an automobile once and a cat twice The 1 and 2 epoch models both varied in their predictions; the 1 epoch model predicted a bird, deer, and horse (being correct 1 out of 3 times); the 2 epoch model predicted a bird, frog, and bird (being correct 2 out of 3 times). The 3 epoch and 20 epoch model both correctly classified the image all times. The Grad-CAM images illustrate how the focus of the model shifts as accuracy improves.

16), we have added values for the confidence drop of the ML models for all the important epochs, which are shown below (refer to Section 3.4: Robustness Methods). The confidence drop scores indicate and quantify the decrease in the model's certainty from when it was subject to clean data to when perturbations (in this case, image occlusions) are added - making it a strong quantitative indicator of a model's robustness, as a smaller confidence drop represents a more robust ML model.

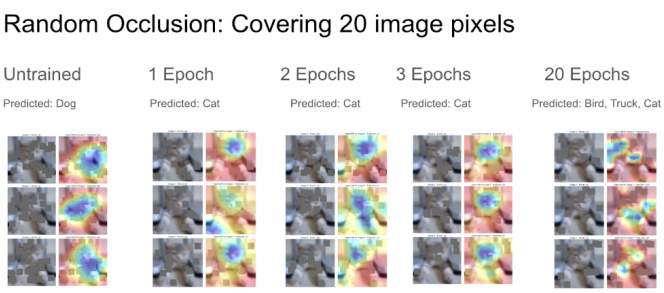

Figure 16. Grad-CAM Analysis of CIFAR-10 Image (Index: 950, Cat) Across Different Training Stages when subjected to twenty pixels of Random Occlusion Perturbations. The one to three epoch models all correctly classify the image while the untrained model incorrectly classifies it as a dog. Surprisingly, the 20 epoch trained model does significantly worse than the one to three epoch models, also having a mix of predictions. The Grad-CAM images illustrate how the focus of the model shifts as accuracy improves.

Table 3. Confidence Drop of the 1, 2, 3, and 20 epoch models when first subject to 20 pixels of Random Occlusion perturbations. Confidence Drop is measured for each model by subtracting the final confidence of the model when subject to the perturbations from the original confidence of the model. A higher value therefore indicates a greater less on confidence when subject to perturbations, indicating weaker robustness.

Important Epoch	Confidence Drop
1	0.0038
2	0.0053
3	0.0044
20	0.0132

Discussion

Discussion of Interpretability Tests

In this section, we explore and analyze the outcomes of the interpretability tests presented in Section 4.2: Interpretability Tests. We qualitatively analyze the outputs of the interpretability tests, in the form of the Grad-CAM heatmaps, to gain insight into the model's decision-making process. We also examine the different Grad-CAM heatmaps for each image across different model checkpoints and provide insight into how these results demonstrate how the decision-making process of an ML model changes as its accuracy improves. Specifically, we aim to display that as accuracy improves, the model's prediction starts to better align with human intuition and focus on relevant areas of the image.

The Grad-CAM visualizations revealed that as the model progressed through training, its attention shifted from irrelevant background areas to more meaningful regions of the images.

For example, Figure 2 of the cat illustrates how the untrained model's key areas were randomly scattered throughout the image's background rather than significant regions, leading to its incorrect prediction. However, in higher accuracy models, the highlighted regions on the heatmap shifted to be less random and to be more towards the key features of a cat. This can be seen in the heatmaps of the 1, 2, and 3 epoch models (which all correctly classified the image), where the highlighted portion of the heatmap begins to close in on the cat's face, where we can assume the model starts learning from training what features make a cat unique (in this case the whiskers and other facial features). The 20 epoch model shows a good representation of the most important features of this particular image for ML models, which is a majority of the face, whiskers, and ear areas. The tests on other images demonstrate a similar trend of improvement alongside accuracy. The tests shown in Figure 5, with the image of the bird, show how the untrained model highlights a large random circular area around the bird without having a particular focus. This ultimately led to it not being able to distinguish further what was inside this highlighted circle and randomly classifying the image. The heatmaps of the trained models began to slowly shift more into the actual shape of the bird and its wings.

The test shown in Figure 3 where the models were subjected to the bird image is particularly interesting as the heatmap of the 20 epoch model shows an emphasis on an area outside of the bird in the shape of wings, as if the bird was raising its wings in the image (although in reality it was not). This faulty emphasis shown in the Grad-CAM is likely a result of the model overgeneralizing features associated with birds and as it grows more accurate and advanced, the model may have learned to expect them due to its large amount of training. This highlights the importance of diverse training data to prevent overfitting and improve an ML model's robustness when subjected to images with unusual features.

Similarly to the tests discussed in Figures 2 and 5, the other interpretability tests shown in Figures 4, 6, and 7 further prove that as the model's accuracy improves, the key regions begin to shift from random areas around the object in question towards defining regions of it (respectively a ship, dog, and ship). These shifts ultimately demonstrate how interpretability tests can be used to show the trend that as accuracy improves, the focus areas of the model begin to align more with key areas for human recognition. These results illustrate how interpretability tests can effectively trace an ML model's learning process and highlight its improving capability.

Discussion of Incorrectly Classified Cases

In this section, we explore the interpretability tests shown in Section 4.3: Grad-CAM images of Incorrectly Classified Cases. These tests are rare cases where models that are highly trained but not perfect, specifically those trained between 1 to 3 epochs, incorrectly classify the image. We examine the Grad-CAMs of these cases to analyze why these misclassifications may be occurring and how the explanations of the model can help us strengthen such models for the future.

From analyzing the Grad-CAM heatmaps of these incorrectly classified cases, we can conclude that many of these incorrect classifications are a result of the subjects in the images being easily confusable with different subjects. For example, in the test in Figure 8, the image (true label: cat) that the 1-epoch model misclassified as a dog appears very similar to a dog with similar facial features and fur tones. The visualizations exemplify this by indicating the model was heavily focused on fur and facial features, which may be alike among the two classes. Similarly, the tests in Figures 9, 11, and 13 show the models' (1, 2, and 3 epochs trained respectively) emphasis on features that may be very common amongst both the correct class and the misclassified class. Particularly, the test in Figure 11 is conducted on an image that may be very unusual for the model, where the truck is slanted upwards similar to an airplane taking off - leading to the model's incorrect prediction of the truck being an airplane.

We observe that most misclassifications were occurring amongst the less trained models ranging from 1-3 epochs, suggesting that earlier in training the ML models relied on superficial similarities that are not discriminative enough for more challenging images. The Grad-CAM heatmaps for some cases also show that models were continuously focusing on certain regions in an image without taking a deeper look into more distinctive or unusual features, which highlights a potential bias in the learning process and indicates that the models were overfitting to specific patterns in the training data.

Discussion of Robustness Tests

In this section, we review the robustness tests presented in Section 4.5 Robustness Testing in Various Checkpoints with the EfficientNetB0 architecture. We analyze the Grad-CAM results of several tests where the models of various checkpoints (0, 1, 2, 3, and 20 epochs) were subjected to multiple perturbations in the input images (Gaussian noise, rotations, and 20 occluded pixels). We qualitatively analyze the Grad-CAM heatmaps to determine how, in the face of corrupted data, the amount of training of the models impacts their robustness in not only their ability to make accurate predictions but also to provide valid explanations through the Grad-CAM.

The first robustness test conducted was the Gaussian noise test in which all the pixels were slightly distorted by a standard deviation of 0.1, as displayed in Figure 14. Upon analyzing the models' predictions and Grad-CAM heatmaps, it is evident that this test was the most difficult for all the ML models, even the one trained to 20 epochs. The untrained model's prediction

was highly inaccurate and its Grad-CAM heatmaps focus on a random circular area around the bird with no distinction. This is similar to the untrained model's prediction when subjected to this image but without the Gaussian noise, as shown in Figure 5. As the training accuracy grew, the models appeared to start focusing near key features that we may associate with birds, despite the Gaussian noise; for example, the 1 epoch model began to highlight certain areas more than others to make sense of the image. However, due to the noise, it was still much more challenging for the model to begin making clear distinctions than if the noise did not exist - proven as both the 1 and 2-epoch models incorrectly classified the image. The 3 epoch model does correctly classify the image, showing that it is experiencing improved resilience to the noise and improved distinction of key features amidst such perturbations. The 20 epoch model shows Grad-CAM heatmaps which appear to highlight the outline of a bird, showing that it maintains and strengthens its ability to identify key regions of an image even with perturbations. However, this model surprisingly incorrectly classifies the image as a frog. This suggests that because the model has undergone extensive training on regular images of birds, it is beginning to overfit to certain patterns in the training data. This therefore leads to the highly trained model being less robust as its overfitting causes it to be unprepared for unusual situations. This suggests that since the highly trained model is heavily trained to recognize patterns in clean data, it is beginning to overfit to such patterns and is not robust to the unusual patterns present in noisy data. Ultimately, this finding illustrates how although robustness does see a general increase alongside accuracy, a very accurate model does not necessarily mean the model is robust. Overall, in our experiments we saw many cases where robustness does increase alongside accuracy, however even some of our most accurate models misclassified images such as the bird example robustness does increase alongside accuracy; however, even the most accurate models can end up misclassifying images due to overfitting. It These observations, which are a result of our experiments, serve to show that it is crucial to diversify the data when training ML models to ensure that the models are not just high-performing but also sufficiently robust and able to adapt to different situations.

The second robustness test conducted was the random rotation test, where the images were randomly rotated from -30 to 30 degrees. The results and Grad-CAM heatmaps of this test amongst all the training stages are shown in Figures 15 and 16 (bird and ship, respectively). Similar to the other tests, the untrained model randomly classified the image based on a random Grad-CAM. The 1 epoch model shows improvement in the Grad-CAM heat maps, as it begins focusing on certain areas more than other areas. However, due to the rotations, it is unable to correctly classify the image. As the accuracy of the models improves, we see an improvement in the robustness, as illustrated by the heatmaps: the visualizations for the 2 and 3-epoch models show much more key regions of the bird being highlighted and the shape of the bird being more significant for the model. The 3-epoch model as well as the 20- epoch model are both able to correctly classify the images, proving to have much more resilience against the rotations. In Figure 17, both the 2 epoch and 3 epoch models correctly classified the ship, however, the 20 epoch model incorrectly classified the image as an airplane one out of the three times it was tested. This incorrectly classified case may not have been a direct result of overfitting, however, as this misclassification occurred when the ship image was rotated to create a positive slant, resembling an airplane taking off (similar to the non-perturbed test in Figure 11). This proves the scope to which corrupted data can change an image to resemble another class, completely confusing the model. Overall, the random rotation tests were much less challenging than the Gaussian noise test for the machine learning models, particularly because key features of each subject were still able to be pieced together although rotated (compared to each pixel being noisy in the Gaussian noise test).

The final test was the random occlusion test, in which 20 random pixels in the image were occluded, or blurred out. The results and Grad-CAM heatmaps from this test are shown in Figure 17. The random occlusion test is a perfect example of the severity of overfitting. The models ranging from 1 to 3 epochs correctly classified the cat image. They also displayed reasonably stable Grad-CAM results, where the highlighted regions in the heatmaps remained close to the face of the cat similar to when they were tested on an unmodified version of this image. This indicates that these models learned to ignore the occluded pixels and focus only on the salient portions of the image. However, the 20 epoch model completely misclassified the image and showed extremely random and inaccurate focus areas through the Grad-CAM. It is important to note that robustness and accuracy are not mutually exclusive, and many models are able to generalize well while maintaining a strong accuracy. However, these results exemplify how if a model has been heavily trained and is highly accurate solely on clean data, they may not be sufficiently robust in light of unusual or disturbed data. This phenomenon of overfitting is demonstrated in Table 3, as higher accuracy models had a much greater confidence drop (the 20-epoch model had a confidence drop of 0.0132 while the 1-epoch model had a confidence drop of merely 0.0038. These findings suggest that the confidence level, or level of certainty in correctly predicting the image, of highly trained models when subject to random occlusions dropped the most significantly (proving that these models were significantly unprepared for sudden changes). This is because such models begin to heavily recognize patterns in the clean data, potentially making them particularly subject to new patterns in perturbed data. This phenomenon can be ascribed to overfitting, where a model becomes too specialized in detecting certain patterns

or features in the training data. This therefore made the model much less flexible and more vulnerable to perturbations in the data, such as the random occlusions.

From our experiment, the deterioration in the Grad-CAM results also helps to illustrate illustrates this, as the model appears to loses lose its ability to identify key features of the cat and resorts to highlighting random locations of the image. Ultimately, this test demonstrates the importance of being aware of bias and overfitting while training ML models to ensure they are robust if faced with corrupted data in the real world. Further research could involve developing adaptive training techniques to ensure that during the training process, ML models learn how to maintain resilience in a wide variety of real-world scenarios.

Limitations and Challenges of This Research

The most significant limitation of this study is the use of the CIFAR-10 dataset, which, while widely used, consists of relatively small images (32x32 pixels) especially compared to more detailed datasets. Therefore, it may not adequately represent the complexities involved in investigating robustness across various perturbations and image corruptions. Furthermore, the interpretability methods employed, such as Grad-CAM, may yield less informative insights due to the constraints and simplicity presented by the CIFAR-10 dataset, potentially limiting the usage of these findings to more complex real-world applications.

While this paper utilizes Grad-CAM to provide insight into an ML model's behavior, it is critical to note the risks of over-reliance on a certain interpretability method (such as Grad-CAM). Specifically, methods such as Grad-CAM can highlight irrelevant areas of an image which still lead to correct predictions but do not actually contribute to the decision making process of the model. This can create a false sense of confidence in the behavior and interpretability of the model, obscuring any true important workings or vulnerabilities.

Naturally, the development and implementation of methods to test and improve both the interpretability and robustness of ML models are hindered by several challenges or limitations. This research aims to test the interpretability and robustness of such models alongside several factors such as accuracy and analysis of incorrectly classified samples. There are, however, several challenges and limitations in doing so and truly understanding the decision-making process of ML models. Some major limitations of this research are that a) ML models can provide an explanation that is independent of the correctness of the model's classification, thus leading to potentially meaningless explanations and b) ML models are not able to provide cause and effect reasoning for their explanations, reasoning which is often needed for the explanation to be more valid in several fields[20]. Furthermore, the issue of the lack of robustness of ML models also ties in negatively with the model's interpretability and the explanations it provides. This is exemplified by how slight perturbations in input data, such as a minor translation, can lead to the model providing completely different explanations. These perturbations can additionally generate a manipulated and therefore unrealistic dataset which often leads to misleading explanations from the ML models that train on it[20]. Therefore, in this research, we aim to closely scrutinize the immensely different explanations given by models when subjected to such unrealistic data.

Another technical challenge is the integration of interpretability or robustness methods on existing ML workflows, as many models and algorithms were not designed with interpretability in mind. This leads to a challenge in integrating these techniques into such pre-existing systems[21]. The most prominent limitation, however, is that the interpretability methods used to force explanations out of models are unrealistically linear and simplified[22]. This does not accurately represent the complexity of real-life machine-learning models. Most interpretability results, such as Grad-CAM, use heavy approximations to display the result and what is outputted to us does not account for the complex inner workings of the model as well as the beyond-complicated steps the model took to arrive at that decision.

Conclusions

From the several interpretability and robustness tests performed, the highlighted regions in the Grad-CAM images demonstrate how as a model's accuracy improves, the focus region of the model shifts more towards regions that humans would recognize as key features of the subject - consistent with prior research. Examining the results of the interpretability tests, it is evident that a higher accuracy model is also more easily able to distinguish the key regions of an image and ultimately use those key regions to classify the image with greater accuracy. However, analysis of the heatmaps of the incorrectly classified cases shows that models that had moderate training (specifically 1-3 epochs) often overly rely on superficial and overly generic traits of each class and therefore fail to identify subtle and less obvious key features of classes. This often leads to the model misclassifying more challenging images, a critical issue in real-world applications. It is therefore crucial to ensure real-world ML models are trained on diverse and balanced datasets to ensure that they can differentiate between such subtle differences.

The Grad-CAM results and predictions from the robustness tests (gaussian noise, rotations, and occlusions) are consistent with previous research in the way that they illustrate that as a model's accuracy improves, they generally become more resilient when faced with corrupted data and capable of identifying key features of an image amongst perturbations. More importantly, however, the tests also illustrate the dangers

of overfitting, or when machine learning models become so accurate and consistently trained on certain data that they face decreases in flexibility and capability when faced with new or corrupted data. This was illustrated as the 20 epoch model did significantly worse on most robustness tests than the 1-3 epoch models. These tests have illustrated that to create high-performing machine learning models, it is crucial to maintain a balance between training for high accuracy and diversifying tests to prevent overfitting. This leads to models which are more reliable for real-world scenarios, where data can vary immensely.

In summary, our research uses Grad-CAM analysis to convey and visually demonstrate a critical message for the future of machine learning: although accuracy is crucial in deploying ML systems, it is just as crucial to consider the factors of interpretability and robustness. By increasing our focus on these aspects, we can create ML models that not only perform well in certain scenarios but also are interpretable, adaptable, and resilient to a complex variety of real-world possibilities.

Acknowledgements

I would like to extend my gratitude to Annamarie Bair of Carnegie Mellon University for her invaluable guidance and assistance throughout the programming and drafting process of this paper.

References

1 T. G. Dietterich, *Annual Review of Computer Science*, 1990, **4**, 255–306.

2 P. Linardatos, V. Papastefanopoulos and S. Kotsiantis, *Entropy*, 2021, **23**, 18.

3 D. V. Carvalho, E. M. Pereira and J. S. Cardoso, *Electronics*, 2019, **8**, 832.

4 R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, *arXiv preprint arXiv:1610.02391*, 2019, **4**, year.

5 T. S. Jaakkola and D. Alvarez-Melis, *arXiv preprint arXiv:1806.08049*, 2018, **1**, year.

6 J. Kim, J. Oh and T. Heo, *Mathematical Problems in Engineering*, 2021.

7 D. Brigo, X. Huang, A. Pallavicini and H. S. Borde, *arXiv preprint arXiv:2104.09476*, 2021, **1**, year.

8 B. Subhash, *Explainable AI: Saliency Maps*, 2022, Medium.

9 Y. Wu, L. Zhang, U. A. Bhatti and M. Huang, *Diagnostics*, 2023, **13**, 2681.

10 Y. Yang, Y. Yuan, Z. Han and G. Liu, *Wiley*, 2022.

11 M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman and F. Doshi-Velez, *arXiv preprint arXiv:1802.00682*, 2018, **1**, year.

12 Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham and G. Slaney, *Journal of Neuroscience Methods*, 2021, **353**, 109098.

13 H. Moujahid, B. Cherradi, M. Al-Sarem, L. Bahatti, A. B. A. M. Y. Eljialy, A. Alsaeedi and F. Saeed, *Tech Science Press*, 2022, **32**, year.

14 I. Kolyshkina and S. Simoff, *Frontiers in Big Data*, 2021, **4**, year.

15 S. Ali, B. Sahoo, A. Zelikovsky, P. Chen and M. Patterson, *Scientific Reports*, 2023, **13**, 4154.

16 J. Rauber, W. Brendel and M. Bethge, *arXiv preprint arXiv:1707.04131*, 2018, **3**, year.

17 Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov and K. Chaudhuri, *arXiv preprint arXiv:2003.02460*, 2020, **3**, year.

18 M. Tan and Q. V. Le, *arXiv preprint arXiv:1905.11946v5*, 2020.

19 X. Wu, U. Jang, J. Chen, L. Chen and S. Jha, Proceedings of Machine Learning Research, 2018.

20 T. A. A. Abdullah, M. S. M. Zahid and W. Ali, *Symmetry*, 2021, **13**, 2439.

21 C. Singh, J. P. Inala, M. Galley, R. Caruana and J. Gao, *arXiv preprint arXiv:402.01761*, 2024, **1**, year.

22 J. Petch, S. Di and W. Nelson, *Canadian Journal of Cardiology*, 2022, **38**, 204–213.