CSE 598: Patching with Perturbation (PwP) Attack on Pneumonia X-ray Images: An Investigation of Vulnerabilities in ResNet Architecture using the idea of Class Activation Map

1st Shreyansh Khandelwal
Ira Fulton School of Engineering
Arizona State University
Tempe, United States Of America
skhand17@asu.edu

2nd Rahulrajan Karthikeyan
Ira Fulton School of Engineering
Arizona State University)
Tempe, United States Of America
rkarthi5@asu.edu

***Abstract*—This research paper aims to investigate the vulnerabilities of deep learning models to adversarial attacks on pneumonia X-ray images and the paper proposes to use adversarial patching and FGSM in combination to attack the deep learning model as well as proposes potential strategies for defending against such attacks. Pneumonia is a significant public health issue that affects millions of individuals worldwide and is potentially life-threatening, making timely and accurate diagnosis of pneumonia critical for effective treatment. Deep learning models have shown promise in assisting with the diagnosis of pneumonia, particularly through the analysis of chest X-ray images. However, recent studies have shown that these models are vulnerable to adversarial attacks, which can have serious consequences, particularly in medical applications. The study explores the potential impact of these attacks on the accuracy and robustness of state-of-the-art models and proposes potential strategies for defending against such attacks. The research paper identifies three potential challenges, including the limited availability of pneumonia X-ray data, the complexity of adversarial attacks, and the robustness of defense mechanisms. The paper also discusses some shortcomings of related works that have investigated the impact of adversarial attacks on deep learning models for pneumonia diagnosis. The study aims to contribute to the ongoing efforts to improve the reliability and security of deep learning models in medical applications, particularly in the context of pneumonia diagnosis.**

I. Problem Definition

Pneumonia is a significant public health issue that affects millions of individuals worldwide and is potentially life-threatening. Timely and accurate diagnosis of pneumonia is critical for effective treatment, as it is responsible for an estimated 15% of all deaths in children under five years old globally, making it one of the leading causes of child mortality. In the United States, there were 1.3 million diagnosed cases of pneumonia in 2017, indicating the substantial impact of this disease on public health. Furthermore, pneumonia causes over 50,000 deaths annually worldwide and has a fatality rate of approximately 4%. These statistics highlight the urgent need for effective diagnosis and treatment strategies to reduce the global burden of pneumonia and its associated mortality rates.

Deep learning models have shown great promise in assisting with the diagnosis of pneumonia, particularly through the analysis of chest X-ray images. However, recent studies have demonstrated that these models are vulnerable to adversarial attacks, which are maliciously crafted inputs that are designed to deceive the model into producing incorrect or unreliable results. Such attacks can have serious consequences, particularly in medical applications where incorrect diagnoses can lead to incorrect treatments and potentially harmful outcomes for patients.

Medical image adversarial attacks on deep learning involve deliberately modifying medical images with imperceptible changes to fool deep learning models into making incorrect diagnoses or predictions. Adversarial attacks on medical imaging deep learning models pose a significant threat to patient safety and trust in medical technology. It is crucial to develop and implement robust defenses against such attacks to ensure the reliability and accuracy of medical diagnoses and predictions. Some defenses that have been proposed include using generative models to detect adversarial examples, training models with adversarial examples, and applying noise reduction techniques to images before feeding them to the models.

Therefore, in this paper, we investigate the vulnerabilities of deep learning models to adversarial attacks on pneumonia diagnosis. We explore the potential impact of these attacks on the accuracy and robustness of state-of-the-art models and propose potential strategies for defending against such attacks.

Our study aims to contribute to the ongoing efforts to improve the reliability and security of deep learning models in medical applications, particularly in the context of pneumonia diagnosis.

II. Challenges

We have identified 3 potential challenges to our approach :-

- Limited Availability of Pneumonia X-Ray Data: One challenge in this research is the availability of a large and diverse dataset of pneumonia X-ray images to train and test deep learning models. The limited availability of such datasets could make it difficult to generalize the results and conclusions of the study.
- Complexity of Adversarial Attacks: Adversarial attacks are designed to exploit the vulnerabilities of deep learning models, and creating effective attacks can be a challenging and time-consuming process. Therefore, designing effective attacks for pneumonia diagnosis could be a significant challenge.
- Robustness of Defense Mechanisms: Although various defense mechanisms against adversarial attacks have been proposed in the literature, their effectiveness in the context of pneumonia diagnosis remains uncertain. Therefore, evaluating and testing the robustness of defense mechanisms against adversarial attacks is an important challenge.

III. Related Works Shortcoming

Recent research has explored the vulnerabilities of deep learning models to adversarial attacks in various applications, including medical image diagnosis. In the context of pneumonia diagnosis, several studies have investigated the impact of adversarial attacks on deep-learning models.

One of the most closely related works is the study by Finlayson et al. (2019), the authors used various types of adversarial attacks, including the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) to evaluate the robustness of deep learning models for pneumonia diagnosis. The authors found that deep learning models were vulnerable to these attacks and suggested the use of adversarial training as a potential defense mechanism. However, their study had some limitations, as it focused primarily on the use of specific types of attacks and did not explore the impact of adversarial attacks on the accuracy and reliability of deep learning models in real-world settings. Moreover, their study only used a limited dataset and did not include a comprehensive analysis of the impact of adversarial attacks on multiple state-of-the-art deep learning models [1]. In the study by Wang et al. (2020), the authors proposed a method for generating adversarial examples for pneumonia diagnosis using generative adversarial networks (GANs). The authors demonstrated that their method was effective in generating adversarial examples that could fool deep learning models. However, the study did not evaluate the impact of these attacks on the accuracy and reliability of these models in real-world settings. Additionally, the study was limited to a single deep learning model and the proposed method for generating adversarial examples may not be applicable to other types of deep learning models. Overall, these prior works highlight the importance of investigating the vulnerabilities of deep learning models to adversarial attacks in medical applications, and the need for more comprehensive evaluation and defense mechanisms to improve the reliability and security of these models [2].

Despite the valuable insights provided by these studies, they are limited in several ways. Firstly, most studies in this area have focused on specific types of attacks, which may not be representative of the full range of possible attacks that could be used in practice. Secondly, few studies have evaluated the effectiveness of defense mechanisms against adversarial attacks in the context of pneumonia diagnosis, leaving a gap in our understanding of how to improve the robustness and reliability of deep learning models for this important medical application.

In this paper, we aim to build upon these prior works and address their shortcomings by exploring the vulnerabilities of deep learning models to a wider range of adversarial attacks, evaluating the effectiveness of various defense mechanisms against these attacks and proposing potential strategies for improving the robustness and reliability of these models in the context of pneumonia diagnosis. Additionally our primary goal is to present a novel adversarial attack approach which is accomplished by combining 2 different adversarial attacks in combination against the deep learning model [3].

IV. Proposed approach

Our Algorithm is a two-step process that includes the idea of attacking the original X-ray images with a patch attack. These patches won't be fit randomly onto the image but would be applied to a heatmap that is generated using Class Activation map.

In deep learning, a class activation map (CAM) is a visualization technique used to understand which regions of an input image contribute the most towards a particular classification decision made by a neural network. It is particularly useful for understanding the internal workings of convolutional neural networks (CNNs), which are commonly used for image classification tasks.

A CAM is typically generated using the final feature map of a CNN, which represents a high-level abstraction of the input image. The weights of the final convolutional layer are used to obtain a weighted sum of the feature maps, which is then passed through a global average pooling layer to obtain a vector of weights. These weights are then used to weight the corresponding feature maps, and the resulting activation map is used to visualize the regions of the input image that are most relevant to the classification decision.

CAMs have a number of applications in deep learning, including identifying which parts of an image are most important for a particular classification task, generating explanations for the decisions made by neural networks, and highlighting areas of an image that may need further analysis or refinement.

We gather the output of our 2nd Convolutional layer of 2nd basic block to visualize the heatmap that helps the resnet18() architecture to classify whether someone has Pnuemonia or not. Figure 1 displays the CAM visualized on one of the images in the validation set.

The popular networks such as ResNet, DenseNet, SqueezeNet, Inception already have global average pooling at the end, so you could generate the heatmap directly without

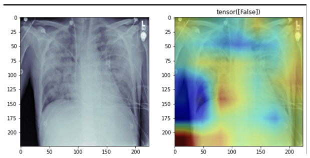


Fig. 1. Class Activation Map

even modifying the network architecture.A simple technique to expose the implicit attention of Convolutional Neural Networks on the image. It highlights the most informative image regions relevant to the predicted class. Figure 2 depicts the CAM architecture in much more detail.

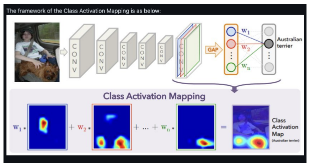

The framework of the Class Activation Mapping is as below:
CONV
CONV
CONV
CONV
CONV
CONV
GAP
w1
w2
wn
Australian terrier
Class Activation Mapping
w1 *
+ w2 *
+ ... + wn *
=
Class Activation Map (Australian terrier)

Fig. 2. Class Activation Architecture

The 2nd step is to apply a dynamic path only on the heatmap. We plan to talk about the already existing patch attack called as Generative Dynamic Patch Attack referred by Xiang Li and Shihao Ji [4]. Their approach is to produce dynamic/static and visible/invisible patches with a few configuration changes.Our pre-trained model outputs ”Pneumonia” for a given X-ray image by extracting the image region or area of interest that has the most significant influence on the prediction. In short, we train our model using Class Activation Map, which supports the classifier most in its forecast. Our novel idea is to apply a patch to this region of interest, simultaneously causing another evasion attack like the Fast-gradient-Sign-method. We call such a novel approach a ”Patching with Perturbation” attack on the heat-map.

The authors of GDPA has proposed ROA, a two-stage patch attack algorithm, which first uses a gray pattern to find the location in image that maximizes the crossentropy loss via grid search, and then optimizes the patch pattern at the identified position. [4]

The idea proceeds further with not only patching but combining FGSM as well to achieve better results. Below we talk about FGSM and general patch algorithm to further illustrate our proposed approach.

- Evasion Attack FGSM : The Fast Gradient Sign Method (FGSM) is a popular adversarial attack technique used to generate adversarial examples for neural networks. The FGSM attack involves computing the gradients of the neural network's loss function with respect to the input image, and then using these gradients to perturb the image in the direction that maximizes the loss. This perturbation is then added to the original image to create an adversarial example that can mislead the neural network's decision.

$$adv_X = X + \epsilon * sign(\partial X J(\theta, X, Y) \qquad (1)$$

where:
θ : Neural Network Model.
ϵ : small value to ensure small perturbations.
J: The Loss function.
X : Original Input Image.
Y : ground-truth label of the input image.

- Adversarial Patch Attack : Adversarial patch attack is a type of adversarial attack in which a small, specifically crafted patch is added to an image in order to fool a machine learning model into misclassifying the image. The patch is designed to exploit the model's weaknesses and can be applied to any part of the image, but is often placed in a strategic location that is likely to be detected by the model. Below is the generic equation used to produce the adversarial images.

$$argmax(y|(x + noise)), constrained||noise|| < \epsilon \qquad (2)$$

$$E_(t)[\log P(y_a|t(x^{`}))] \qquad (3)$$

The expected probability of a class is maximized over all possible transformation functions(t T),with a constraint on the Expected effective distance between the transformed original and transformed perturbed image.

  - The patch is created by optimizing a small image that is added to the original image, such that it maximizes the model's error rate on the modified image. This process is often achieved through the use of an optimization algorithm, such as gradient descent, to iterative adjust the pixel values of the patch until the desired effect is achieved.
  - Once the patch is added to the image, the machine-learning model can be easily fooled into misclassifying the image. Adversarial patch attacks are particularly concerning because they can be easily executed in the physical world. For example, an attacker could print a small patch on a piece of paper and use it to fool a computer vision system in a real-world setting.

V. Dataset Description

Data for Pneumonia: We will use the RSNA Pneumonia Detection challenge(https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/). The radiological society of North America provides this dataset. This data set has 26684 X ray-images, with 20672 images without pneumonia and 6012 appearances with pneumonia. This dataset is quite an imbalance and thus enhances the challenge to attack it. We need to do quite a lot of preprocessing on this dataset as images are too huge.

- The original image shape is (1024 * 1024), and these images have to be resized to (224 * 224)
- Standardize the pixel values into the interval[0,1] by scaling with 1/255.
- Split data-set into 24000 train images and 2684 validation or test images.

- Store converted images in folders corresponding to the class. 0 if there is no pneumonia and one if pneumonia

For Training the Neural Network, we plan to use the ResNet18 Deep Neural network for image classification, BCEWithLogitsLoss as a loss function, and Adam as the Optimize. The model will be trained plan for 30 epochs.

VI. Evaluation plan

We plan to evaluate the performance of our Neural network on several metrics to assess the robustness of the model. The images we obtained from the kaggle data set were in DICOM format. We need to first understand how to handle the DICOM format images using python library pydicom. We understood a mechanism of getting pixel array values from the pydicom object after DICOM format files. We started our evaluation first by looking at the random picked X-ray images from the training dataset.

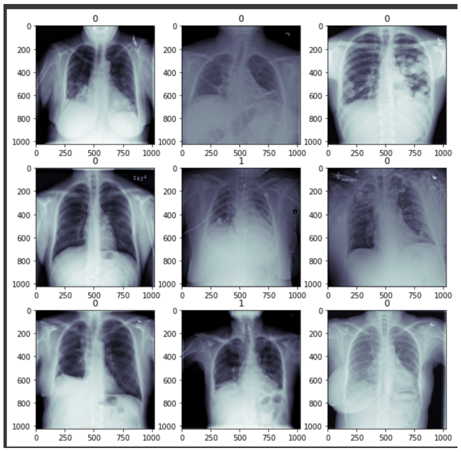
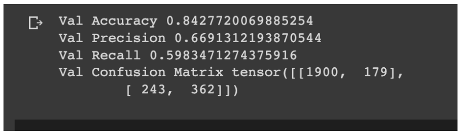


Fig. 3. X-ray Images

Figure 3 shows X-ray images with labels associated on each image. Label : 1 indicates that someone has a pnuemonia and Label : 0 indicates that someone doesn't have a pnuemonia. The dataset is quite huge with around 26,684 dicom files with around 24000 are kept in training dataset and 2684 files are kept in validation dataset. A couple of preprocessing steps were applied on images as they were quite huge and we need to resize them to 224 * 224 size. Some sort of transformations were applied to both training and validation datasets such as Normalizing them based on mean and standard deviation, Affine with scale and degrees were applied too. We used pytorch lightning module to create a resnet18() architecture as it contains of a lot of Convolutional 2D and BatchNorm layers. The model is trained and evaluated on the validation dataset using pytorch torchmetrics to calculate the accuracy, precision, recall and confusion matrix

- Accuracy: This is the most basic and widely used metric for evaluating the performance of a model. It measures the proportion of correctly classified examples out of the total number of examples in the test set. Our Accuracy came around 84% which is quite high given that our datasets are quite unbalanced so weed to rely on other metrics too such as precision.
- Precision: Precision is a common evaluation metric used in machine learning and information retrieval tasks. It measures the fraction of relevant instances among the total number of instances that were retrieved. In other words, precision is the proportion of true positives among all the positive predictions made by a model. Our precision came around 66%.
- Recall:Recall is another commonly used evaluation metric in machine learning and information retrieval tasks. It measures the fraction of relevant instances that were retrieved among the total number of relevant instances. In other words, recall is the proportion of true positives among all the actual positive instances in the data.Our recall value came around 59.83%.
- Confusion Matrix:A confusion matrix is a table used to evaluate the performance of a machine learning model on a set of test data where the true values are known. The matrix shows the number of correct and incorrect predictions made by the model, and it is organized into rows and columns based on the true and predicted class labels.
  - True Positive : 1900
  - False Negative : 179
  - False Positive : 243
  - False Negative : 362



```
Val Accuracy 0.8427720069885254
Val Precision 0.6691312193870544
Val Recall 0.5983471274375916
Val Confusion Matrix tensor([[1900,  179],
        [ 243,  362]])
```

Fig. 4. Results on Validation dataset

- The figure 4 shows the evaluation metrics on validation dataset where traning dataset is trained on resent18 architecture.
- *** Rahul : Please complete it from here. Please check this paper to get the better idea on metrics : https://arxiv.org/pdf/2111.04266.pdf. You need to elaborate on what happened after we attack using our proposed approach and metrics we generated.

VII. Code

The github code with dataset, code and read me file is uploaded at github and one refer to the following link github repo.

References

[1] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. Pluim, M. Veta, C. I. Sánchez, and M. de Bruijne, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, p. 102141, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521001870

[2] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on ai in medical imaging informatics: A survey," *Expert Systems with Applications*, vol. 198, p. 116815, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741742200272X

[3] X. Li and D. Zhu, "Robust detection of adversarial attacks on medical images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1154–1158.

[4] X. Li and S. Ji, "Generative dynamic patch attack," 2021. [Online]. Available: https://arxiv.org/abs/2111.04266