Healthcare: Adversarial Defense In Medical Deep Learning Systems

Authors: Rakesh Senthilvelan & Madeline Tjoa

Mentor: Lily Weng, Section A15

Code: https://github.com/Maderlime/DSC180_Q1_Code

Introduction

The medical space is one that is fundamentally sensitive in terms of its effects on the lives of the patients within it. As a result, the transition into deep learning systems handling more sensitive information and tasks comes with the worries of those systems being compromised in some way and those vulnerabilities being responsible for harm to the lives and assets of people. Research on adversarial attacks shows cases where imperceptible adjustments to data within a deep learning system can cause said system to make incorrect predictions a majority of the time.

Adversarial attacks are methods used to interfere with deep learning systems- with the intent of finding ways to misclassify data. These attacks generally come in the form of targeted and untargeted attacks. Targeted attacks entail manipulating data to output a desired outcome after feeding it to a model, while untargeted attacks focus on manipulating data to simply not be recognized as it's correct output.

In order to combat against such adversarial instances, there needs to be robust training done with these models in order to best protect against the methods that these attacks use on deep learning systems. In the scope of this paper, we will be looking into the methods of fast gradient signed method and projected gradient descent, two methods used in adversarial attacks to maximize loss functions and cause the affected system to make opposing predictions, in order to train our models against them and allow for stronger accuracy when faced with adversarial examples.

Background

In the case of our research, we primarily looked into adversarial attacks against healthcare systems. Healthcare is fundamentally a sensitive space, with patient data being highly protected and every decision made having a lasting impact on the health of the people that are involved. As technology advances within this sector, deep learning algorithms are being used for new tasks, including diagnosing patients with conditions based on viewing medical images such as photographs, x-rays, and other diagnostic scans. In cases where adversarial examples attack these deep learning models, the possibilities for misdiagnosis and subsequent fraud and bodily harm begin to grow (Tjoa & Senthilvelan). With these types of adversarial attacks, we could see problems such as overprescription or underdiagnosis of conditions among others begin to arise. As a result, adversarial defenses will need to be implemented into these systems in order to better protect against these adversarial attacks and the outcomes they produce. We worked on developing robust models, which will take in data and run adversarial attacks on it in the training process in order to better train the model against FGSM and PGD attacks in the testing stage. In the scope of our research, we have taken into account image data from three different datasets. One of our sets shows chest x-rays with different disease conditions as well as healthy conditions. The goal here is to determine whether there is a disease within the chest x-ray (i.e. within the lungs, heart, etc.). The next set shows images of human eyes with the intention of determining whether the eye shows signs of diabetic retinopathy, an eye disease associated with diabetes. The warning signs for this disease can be seen through "the presence of lesions associated with the vascular abnormalities caused by the disease" (California Healthcare Foundation). In the scope of our paper, we will be looking at differentiating between eyes that have these conditions and eyes that do not. Finally, we have our dermatology dataset, which focuses on skin abnormalities and trying to find a skin cancer called melanoma, one that is easily treatable if detected early but can be deadly if it develops into later stages. The scope of our project will look into different images samples of human skin to determine whether or not melanoma is present in the image.

Generally, these deep learning algorithms are built using neural networks, particularly convolutional neural networks as is usually the case when working with image inputs (University of Michigan). Convolutional neural networks, or CNNs for short, work through the usage of layers that handle functionalities such as decreasing the computational power required to process the data through dimensionality reduction, extracting dominant features, and flattening in order to produce the proper classification output (Saha). In the case of our research, we have used the ResNet model to build our neural network for all three of our image classes. ResNet is a model that allows us to construct networks that can be up to thousands of layers deep, allowing it to have stronger performance than other “shallower” models that may suffer from the “vanishing gradient” problem. The “vanishing gradient” problem is where “the neural network training algorithm tries to find weights that bring the loss function to a minimal value, if there are too many layers, the gradient becomes very small until it disappears, and optimization cannot continue” (run:ai). For our purposes, we built our ResNet model from the PyTorch library but with modifications to account for our image sizes and features.

Pipeline

We perform two adversarial training methods on each of our datasets: Projected Gradient Descent and Fast gradient sign method. In order to test the effects of robust training utilizing these methods, we compare the test accuracy of a model on adversarial attacks. First, we will explain how these two methods work.

Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method is one of the methods that we will experiment with using for adversarial robust training. Fast gradient sign method is an adversarial method that utilizes the gradients of a neural network's loss in order to affect the input image in order to maximize the loss value. Training around this would allow the neural network to account for a seemingly worst case scenario where losses are maximized, allowing the model to better protect against adversarial attacks that are imperceptible to humans.

The Fast Gradient Sign Method for adversarial attacks is represented by the equation:

$$x_{t+1} = x_t + \varepsilon sgn(\nabla_x L(x, y))$$

FGSM equation, Tensorflow Documentation

In our use case, we will use a "fairly minor modification to the random initialization for FGSM adversarial training", as sourced from Eric Wong, Leslie Rice, J. Zico Kolter in their paper "Fast is better than free: Revisiting adversarial training", as we found it to have stronger performance than a standard FGSM model and to also have comparable performance to Projected Gradient Descent, which uses multi-step gradients. In addition, this method is also significantly faster to train and test than Projected Gradient Descent. In the case of the Wong et al. paper, their modified FGSM model for the CIFAR10 dataset was able to train in 12 minutes and achieve comparable accuracy to a PGD model that took 4966 minutes to train. We will call this modified FGSM algorithm “Fast-FGSM” throughout this report.

Projected Gradient Descent

We explore Projected Gradient Descent as a standard for traditional adversarial training in order to compare the effectiveness of the Fast Gradient Sign Method results. Projected Gradient Descent (PGD) is known to be effective in training for adversarial attacks, but can be computationally expensive to run. Its goal is to solve the inner maximization problem over a threat model, where threat model refers to the type of attack to be performed on a model (ie white box attack, black box attack, targeted vs untargeted attack). This algorithm finds perturbations that maximize loss of model on input and differentiates itself from FGSM through its usage of multi-step gradients. The way that the algorithm works from a functionality standpoint can be expressed by the following pseudocode, which walks through each step that a PGD attack takes. This algorithm is also represented by the following algorithm.

Algorithm 1 PGD adversarial training for T epochs, given some radius ϵ, adversarial step size α and N PGD steps and a dataset of size M for a network f_θ

for $t = 1 \dots T$ **do**
  for $i = 1 \dots M$ **do**
    // Perform PGD adversarial attack
    $\delta = 0$ *// or randomly initialized*
    for $j = 1 \dots N$ **do**
      $\delta = \delta + \alpha \cdot \mathrm{sign}(\nabla_\delta \ell(f_\theta(x_i + \delta), y_i))$
      $\delta = \max(\min(\delta, \epsilon), -\epsilon)$
    end for
    $\theta = \theta - \nabla_\theta \ell(f_\theta(x_i + \delta), y_i)$ *// Update model weights with some optimizer, e.g. SGD*
  end for
end for

PGD pseudocode (Wang et al)

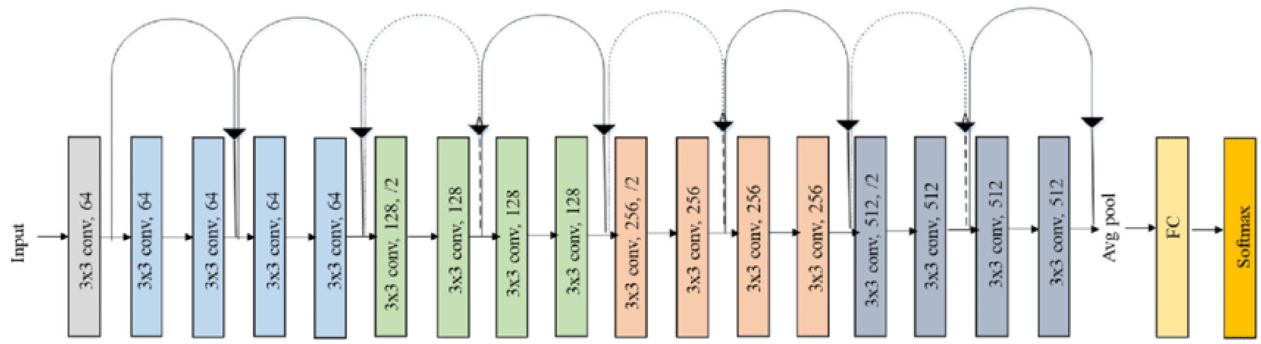
$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \epsilon \mathrm{sgn}(\nabla_x L(\theta, x, y)))$$

PGD Equation, Tensorflow documentation

Based on the use cases for our research, we were able to find that modifications made to the FGSM algorithm can allow it perform at a similar level to the PGD algorithm. With PGD being a significantly longer process than FGSM, we decided to focus on our FGSM implementation for robust training in order to better examine the differences between standard and robust models.

ResNet Neural Network

For our research, we are using the ResNet neural network model. ResNet is a neural network that was first introduced by researchers at Microsoft Research in 2015 with a new architecture called Residual Network (GeeksForGeeks). Neural networks before ResNet suffered from an issue known as the "vanishing gradient" problem, which occurs when a neural network has many layers and the gradient becomes too small to work effectively in training (Wang). With ResNet being able to handle this issue, we are able to produce deeper neural networks that can produce stronger deep learning models for our use case. In our case, we use the ResNet model offered in the PyTorch library for Python. Initially, this model was built to handle the CIFAR-10 dataset which focuses on the classification of tiny images of varying classes. We have modified this ResNet model to fit the image sizes and classes of our dermatology, ocular, and chest x-ray images. In our process, we will train a ResNet model with a training dataset for each of our image classes separately while defining the parameters. Our ResNet model will be a ResNet-18 model, which is 18 layers deep. The architecture for this model is shown in the diagram below:



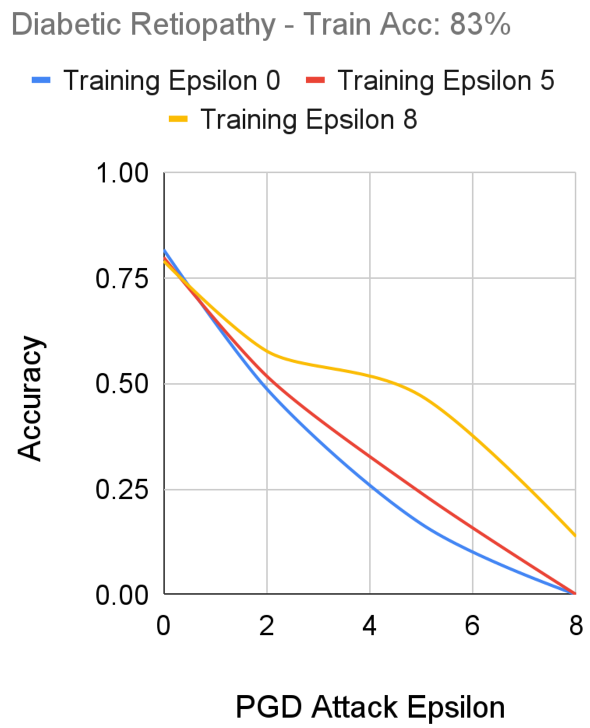
ResNet-18 Architecture, Research Gate

In this case, we will utilize different epsilon parameters to determine how robust the model will be to the FGSM attacks that we run on the code, with an epsilon of 0 indicating a standard model with no robust training. From here, we can compare each model's accuracy against attacks of varying epsilons as well to determine whether or not robust training is an effective countermeasure to adversarial attacks.

Results

Diabetic Retinopathy

For the purpose of recording our results for diabetic retinopathy, we ran the same FGSM training model with changing epsilon values. In this case, epsilon represents the coefficient of the loss functions as seen in the FGSM equation. Diabetic retinopathy represented our largest dataset, coming in with 10644 images. As a result, our code had to be optimized to run effectively within the computing resources available to us while also taking into account this large dataset. As a result, this section was run on algorithms with epsilons 0, 5, and 8 with attack epsilons of 0, 2, 5, and 8. Epsilon 0 on the training algorithm indicates that the algorithm does not have robust training while higher epsilons indicate higher levels of robust training. When looking at the attack epsilon, an epsilon of 0 indicates no adversarial attack while higher epsilons indicate stronger attacks. Using FGSM, we were able to gauge the following results from those runs:

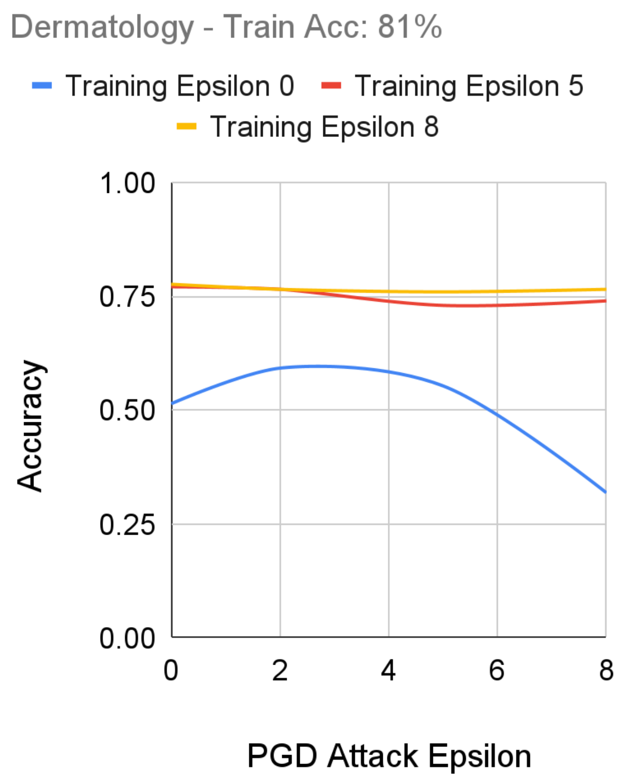


Senthilvelan & Tjoa, FGSM Robust Training Performance on Diabetic Retinopathy Dataset

From what we can see, the models all struggle against adversarial attacks, with a significant dropoff in performance across the board upon attack epsilons greater than 0. However, we can see that with higher training epsilons, there is more resistance to the adversarial attacks relative to those with lower training epsilons.

Dermatology Skin Patches

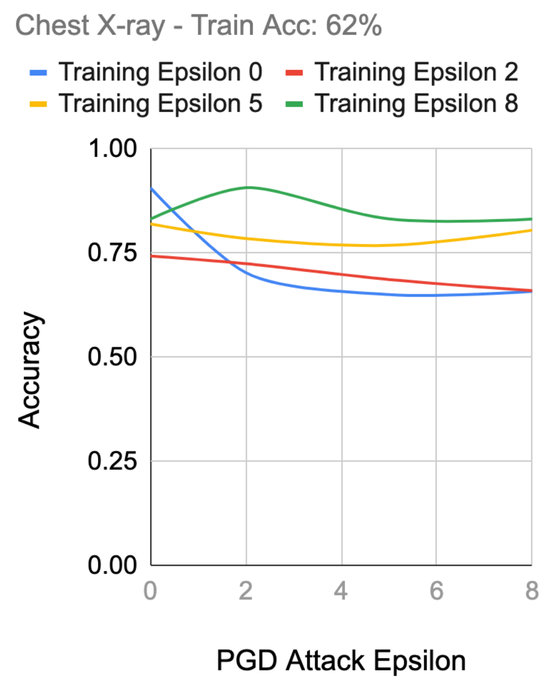
After performing training using Fast-FGSM on our dataset, we notice our models that had Epsilon 5 and 8 training epsilons performed relatively consistently against adversarial attacks between 0 to 9 epochs. Alternatively, we notice the training set with epsilon 0 outputted a lower accuracy as the epsilon of the adversarial attacks increased; with our takeaway being of a similar structure to our Diabetic Retinopathy results regarding the effectiveness of robustly trained models on epsilon attacks.



Senthilvelan & Tjoa, FGSM Robust Training Performance on Dermatology Dataset

<u>Chest X-rays</u>

After performing training using Fast-FGSM on our chest x-ray dataset, we notice a more prominent effect of robust training as the standard model (Training Epsilon 0) shows a high accuracy on regular images, which promptly decreases as the adversarial attacks intensify. On the other hand, we notice our most robustly trained model (Training Epsilon 8), seems to perform consistently against the various adversarial attacks, although its initial training accuracy on non-adversarial images did not perform as well as the standard model. Interestingly, we don't seem to notice a major tradeoff between the robustness of a model and the accuracy compared to less robust models.



Senthilvelan & Tjoa, FGSM Robust Training Performance on Chest x-ray Dataset

Conclusion

From our research, we can see that robust training allows for better performance in deep learning systems in the Fast-FGSM. We can see that the robust training allows the model to be better prepared for adversarial attacks and that a more robust algorithm could help handle the cases where ours was unable to detect the adversarial attacks. In our case, we were unable to explore stronger models due to the computing constraints we faced. However, based on our findings, we believe that experimenting with different parameters such as higher training epsilon, training with more epochs, further optimizing the Fast-FGSM algorithm, using a deeper neural network, or having access to more training data among other factors may help in further advancing the performance and efficiency of robust training against adversarial attacks in healthcare. Considering the significant ramifications of adversarial attacks being deployed onto sensitive healthcare systems, the value of robust training in these systems is very apparent.

Works Cited

Finlayson, Samuel G., et al. “Adversarial Attacks against Medical Deep Learning Systems.” *ArXiv.org*, 4 Feb. 2019, https://arxiv.org/abs/1804.05296.

Original Resnet-18 Architecture | Download Scientific Diagram. https://www.researchgate.net/figure/Original-ResNet-18-Architecture_fig1_336642248.

“Pytorch Resnet.” *Run*, https://www.run.ai/guides/deep-learning-for-computer-vision/pytorch-resnet#:~:text=Residual%20Network%20(ResNet)%20is%20a,layers%2C%20which%20outperform%20shallower%20networks.

“Residual Networks (Resnet) - Deep Learning.” *GeeksforGeeks*, 27 Jan. 2022, https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/.

Senthilvelan, Rakesh, and Madeline Tjoa. *Healthcare: Adversarial Attacks Against Medical Deep Learning Systems,* https://docs.google.com/document/d/1iQ0lZ_wpxqQXRwwjwKANrwNn6FRDwvasISHf9vpNIHw/edit?usp=sharing`.

Wang, Chi-Feng. “The Vanishing Gradient Problem.” *Medium*, Towards Data Science, 8 Jan. 2019, https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484.

Wong, Eric, et al. “Fast Is Better than Free: Revisiting Adversarial Training.” *ArXiv.org*, 12 Jan. 2020, https://arxiv.org/abs/2001.03994.