Performance Analysis of Malaria Detection Network subjected to FGSM Adversarial Attack

Arunkumar Nachimuthu Palanichamy
NYU Tandon School of Engineering
an3333@nyu.edu

Umesh Deshmukh
NYU Tandon School of Engineering
urd7172@nyu.edu

Viswanathan Babu Chidambaram Ayyappan
NYU Tandon School of Engineering
vc2173@nyu.edu

Github Repository: https://github.com/arunkumar1531/ECE-GY-9163-Project

Abstract

Malaria cell detection using Convolutional Neural Networks have been remarkably successful and the state of the art models provide results with $> 99\%$accuracy. Many state-of-the-art models have been used to detect the malaria parasite in the blood cells and the results are promising. Using Neural networks for such important scenarios makes us question the reliability of the networks. Having said that, recent studies have proved that all the neural networks are susceptible to adversarial attacks that can alter the predictions of the model by making very minor changes that can't be detected otherwise. In a real-world scenario, where the room for error is close to none this unreliability in neural networks poses a serious threat. We train a CNN model to detect Malaria parasites in thin blood smear images. Following that, we test the ability of the network to withstand adversarial attacks and tabulate the results and subsequently try to make the network robust to attacks and tabulate the results before and after implementing the defense.

1 Introduction

The parasite Plasmodium, which causes malaria, is spread via the bite of an infected mosquito. Malaria can only be spread by mosquitoes of the Anopheles genus. Fever, vomiting, and/or headache are among the disease's signs and symptoms. There are 4 types of human Malaria,

- Plasmodium falciparum
- Plasmodium vivax
- Plasmodium malariae
- Plasmodium ovale

The most lethal kind is Plasmodium falciparum. Plasmodium vivax is the cause of 77% of infections in the Americas. Children under the age of five, pregnant women, and infants are more likely to contract the parasite. Pregnancy-related malaria complications can have serious effects, such as miscarriage, premature birth, low birth weight, congenital infection, and even perinatal mortality. It is crucial to have a more strong system in place for an early and rapid diagnosis given the high risk and severity witnessed.

This disease, even though it seems harmless at initial stages could turn to be most fatal if it is left untreated for a prolonged period of time. To understand the importance of research related to Malaria detection, it is essential to know certain facts about the same.

- A child dies of malaria every 2 minutes.
- In the Americas, 765,000 cases of malaria and around 340 deaths were reported in 2018.

- In 2017 there were 219 million cases of malaria globally, causing nearly 435,000 deaths, mostly among African children.
- Approximately half of the world's population is at risk of malaria, particularly those living in lower-income countries. In the Americas, 138 million people live in areas at risk of malaria.

So, the cornerstone of treating Malaria or any other life-threatening disease for that matter is to detect it as early as possible. With the conventional method of detecting malaria using microscope to detect the presence of the parasite in cells is a tedious and slow process. Also, it requires highly skilled professionals to conduct the process. This makes the whole diagnosing process slow and unreliable. Also, there is no guarantee that many remote parts of the world would have the required technology and personnel to treat the disease.

There have been a plethora of research activities on Malaria detection in humans using Deep learning and Computer Vision. Before using the classification networks, the key features are extracted from the cell Images using computer vision techniques. This helps in increasing the accuracy of the detection model as a whole. However, the recent developments in deep learning has made the networks capable of detecting the key features from the image during its training phase and this renders the process of explicitly detecting the key features redundant and useless. Upon the introduction of Transfer Learning, CNN and other ensemble learning methods, the Image classification task has become very efficient and accurate. There are various state-of-the-art neural networks like ResNet, VGG, and GoogleNet that have proven to be very accurate in Image classifications.

CNNs are widely used across various domains like real-estate, e-commerce, healthcare, etc... and the medical domain is no exception. However, when it comes to the medical domain, the stakes are very high and even a minor mistake can turn fatal. So, it is very important for the networks to be accurate and robust. Also, the medical Images are very different from the Images we see in our daily lives. They mostly involve Images of cells with no concrete structures and the model has to detect all the minor features of the cell structure. Given the scale of operation, it has been proven that even a minor alteration to the input dataset could lead to false predictions from the model with greater certainty. This would lead to numerous problems. Malaria detection is a case where both high false negative rate and false positive rate would be a serious problem. So it is essential for the network to be robust towards any adversarial attacks. So, in this paper we study the effects of FGSM attack on neural networks and eventually we try to prevent such adversarial attacks on neural networks by making it robust. Since we don't know what is happening under the hood in neural networks, it is essential to understand how these attacks could affect the network in order to prevent these attacks in future.

2 Related Work

The CNN model learns to extract features from images. The trained model can be used to solve similar problems known as transfer learning. The main reason to use transfer learning in medical related field is lack of proper annotated data. Healthcare economy is very big and fraud already exist[5]. There is possibility that in the future decision about medical reimbursement will be made by algorithms. The ground truth in medical imaging can be controversial even specialist can disagree on results. In these scenarios attack on classifier can be very devastating. [6] In dermatology it is more difficult to perform as many procedure as possible, this can result in performing as many procedure as possible even if many of them are unnecessary to increase their revenue. A dermatologist from Florida was sentenced 22 years in prison for performing unnecessary procedures[6]. In a hypothetical scenario an insurance company can use a deep learning method for diagnosis to avoid fraud scenario mentioned above. A bad agent could modify input to the system to obtain the result they want. There is some research done on vulnerabilities of deep learning networks for classification[4]. various defence methods have been proposed for adversarial input attacks[1][2][3].

3 Methodology

3.1 Dataset

The dataset used in this project is from the National Institute of Health's Malarial Parasite infection research work. Malaria is caused by parasites that are transmitted through the bites of infected mosquitoes. The current standard method for malaria diagnosis in the field is light microscopy of blood films. The dataset has large scale blood smear images. It is a collection of segmented red blood

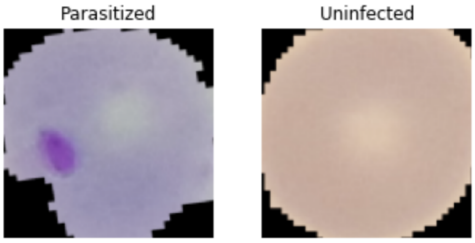

Figure 1: A sample from the dataset.

cells using stained glass collected from patients. We take a subset of thin blood smear images which is labelled into two categories - parasitic(infected) or uninfected. It is hard to classify them with naked eye at some cases. But In some clear cases, we can see a red parasitic element in the infected images dataset. In total there are 27,558 images in the dataset. See a sample of dataset from Figure 1

3.2 Custom Architecture:

To solve this classification problem, we create a custom model with 11 layers. The input images are preproceesed with various techniques like image resize, flattening them into an array, normalizing the input and then passing it into 2D convolutional layers with increasing filter size 32, 64, 128, following by 2D Max pooling, then flatten them and pass them into a dense layers. There are two dense layers. Dropout technique is also used to prevent overfitting. The model is shown in Figure 2

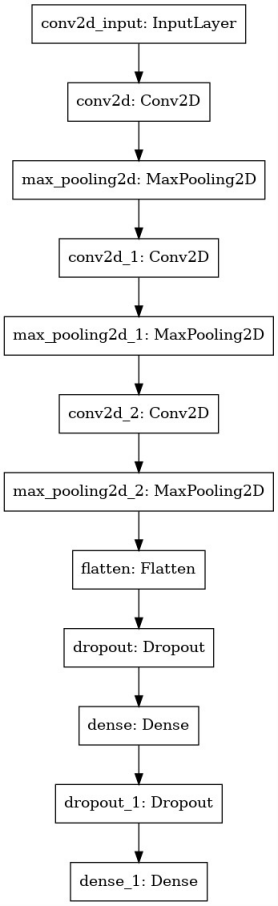

Figure 2: CNN Architecture

3.3 Adversarial Attacks

Adversarial attacks in machine learning (ML) refer to the deliberate manipulation of input data in order to deceive a trained model. These attacks can take various forms, such as adding noise to an image or modifying its pixel values in a subtle way that is imperceptible to humans but still causes the model to make incorrect predictions. Adversarial attacks can be extremely challenging to detect and prevent, as they often involve sophisticated techniques and require a deep understanding of the ML algorithms and their vulnerabilities. Moreover, they can be highly effective in disrupting the performance of ML models, potentially leading to serious consequences in real-world applications. One of the key challenges in dealing with adversarial attacks is the fact that they often involve finding a delicate balance between making the perturbations small enough to avoid detection, while still being large enough to cause the model to make errors. In addition, adversarial attacks can also be tailored specifically to target a particular model or dataset, making them difficult to generalize across different scenarios.

3.3.1 Fast Gradient Sign Method

The fast gradient sign method (FGSM) is a technique that manipulates the input data to a neural network in order to cause the model to make incorrect predictions. It does this by calculating the gradients of the loss function with respect to the input image, and then adding a small perturbation to the original image that maximizes the loss. This perturbed image, called the adversarial image, is specifically designed to deceive the model and cause it to make mistakes.

One unique aspect of FGSM is that the gradients are calculated with respect to the input image, rather than the model parameters. This is done in order to create an adversarial image that maximizes the loss, and therefore deceives the model. The process of finding the gradients is relatively simple, as it involves using the chain rule and calculating the contribution of each pixel to the loss value. This allows for fast and efficient generation of adversarial images. Furthermore, since the model is already trained and the gradients are not taken with respect to the model parameters, the model remains fixed and the only goal is to fool it using the adversarial image.

$$adv_X = x + \epsilon * sign(\nabla_x J(\theta, x, y))$$

Here,

- adv_X: Adversarial image
- x: Original input image.
- y: Original input label.
- ϵ : Multiplier to ensure the perturbations are small.
- θ : Model parameters.
- J: Loss.

FGSM attacks are particularly effective because they are computationally efficient and can be easily implemented using standard gradient-based optimization techniques. In addition, they are relatively simple to understand and can be applied to a wide range of ML algorithms and tasks. Despite their effectiveness, FGSM attacks can be detected and mitigated using various techniques, such as adversarial training and defense mechanisms. These techniques aim to improve the robustness of ML models against FGSM attacks by exposing them to a range of adversarial perturbations during training, and by incorporating additional constraints and regularization terms into the optimization objective.

We take our CNN classifier custom designed for Malaria parasite cell classification and assume it is subjected to FGSM attacks and hence defend this attack by adversarially retraining the model

4 Experiment

We start with training a custom model for malaria detection from cell images. We start by attacking a model with FGSM attack. Test accuracy is determined for different values of epsilon. This shows a significant drop in accuracy when FGSM attack is used. Figure 3 shows the perturbation pattern obtained and shows the image modified with the perturbation pattern. Adversarially modified images can influence the prediction output of the network and adversely affect test accuracy.

Now, we train a ResNet50v2 model which was initially trained on the Imagenet dataset and produced 98% accuracy. We train the network by adding an Average Pooling layer followed by a dense layer with sigmoid activation to accommodate our dataset and labels. Also, we added a dropout of 0.5. The final training accuracy of the model was 96% after 10 epochs.

We tried attacking the ResNet50v2 model using an FGSM attack and the accuracy dropped to as low as 38% for different epsilon values. Now, we trained both models using different set of adversarial images generated using epsilon values in the range (0.04, 0.4). Finally, we tried attacking the networks after implementing the defense and tabulated the results below. The main problem here is, the perturbations in some cases are very subtle in that it is not visible to the naked eye but still it is powerful enough to alter the model's prediction. The perturbation pattern and the perturbed Image are shown Figure 4

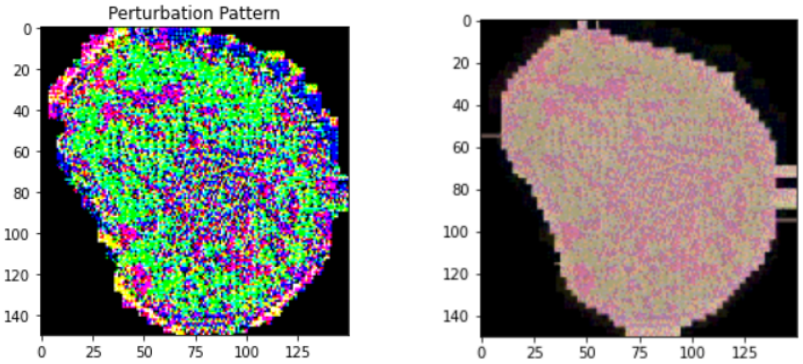

Figure 3: Original Perturbation Pattern(Left), Modified Perturbation Pattern(Right) for Custom Model

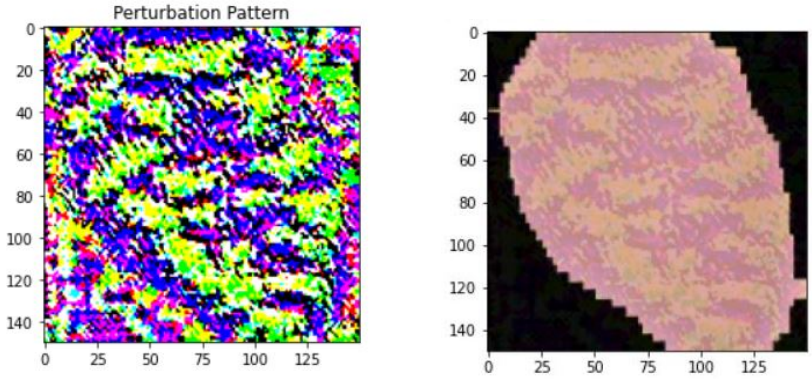

Figure 4: Original Perturbation Pattern(Left), Modified Perturbation Pattern(Right) for ResNet50v2 model

5 Results

Results for Custom Network after attack	
Epsilon	Accuracy
0.004	87.85
0.01	75.21
0.04	68.95
0.078	51.16

Results for Custom Model after adversarial training	
Epsilon	Accuracy
0.004	89.90
0.011	81.10
0.019	77.55
0.027	75.40

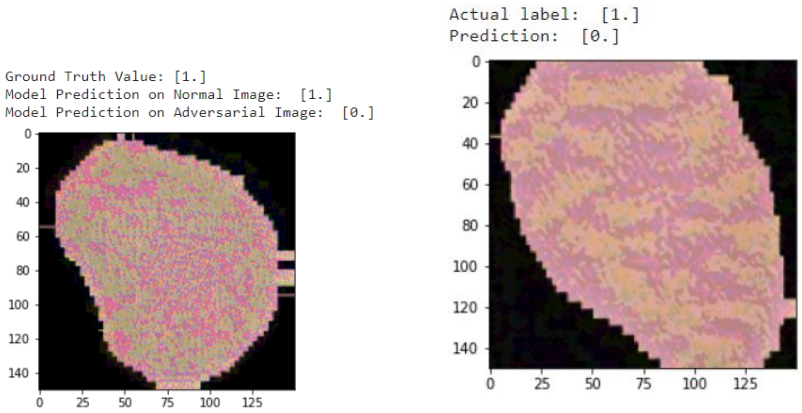

Figure 5: Output Prediction for Custom Model (Left), for ResNet50v2 model (Right)

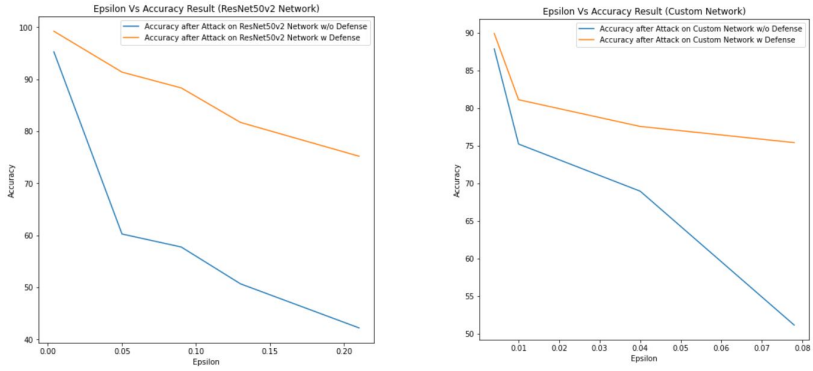

Figure 6: Accuracy vs Epsilon plot for Custom Model (Left), for ResNet50v2 model (Right)

Results for ResNet50v2 after attack	
Epsilon	Accuracy
0.004	94.28
0.05	60.26
0.09	57.75
0.13	50.67
0.21	42.20

Results for Resnet50v2 after adversarial training	
Epsilon	Accuracy
0.04	99.23
0.25	91.37
0.30	88.33
0.35	81.70
0.40	76.20

6 Conclusion

We trained 2 different networks and adversarially attacked each of them. We also trained the networks with adversarial inputs to be robust against attacks. From the results, it is evident that a perturbation which is not visible to the naked eye added to the image can change the output of network. As we can see from the tables, when we increase the epsilon value, the accuracy deteriorates. After training the network with perturbed images, The network becomes robust to FGSM attack. However, this defense technique is specific to FGSM attack. Also it is evident from the results that this defense technique is immune to FGSM attacks but still the improved accuracy rate is not sufficient in the case of the custom model. With the defense implemented, the network accuracy for a higher epsilon value was around 89% which would still not be sufficient in the healthcare domain.

Future work on this could include implementing different type of attack on the network trained with adversarial images and checking the performance. Also, a different defense technique could be implemented. Having understood the importance of accuracy in healthcare domain, it is essential for the networks to have maximum accuracy and immune to any to adversarial attacks.

References

[1] LDL: A Defense for Label-Based Membership Inference Attacks https://doi.org/10.48550/arXiv.2212.01688

[2]Adversarial Detection by Approximation of Ensemble Boundary https://arxiv.org/ftp/arxiv/papers/2211/2211.10227.pdf

[3]A Review of Adversarial Attack and Defense for Classification Methods https://doi.org/10.48550/arXiv.2111.09961

[4]https://doi.org/10.48550/arXiv.1903.07054

[5]Adversarial Attacks Against Medical Deep Learning Systems https://doi.org/10.48550/arXiv.1804.05296

[6]William J Rudman, John S Eberhardt, William Pierce, and Susan Hart-Hester. 2009. Healthcare fraud and abuse. Perspectives in Health Information Management/AHIMA, American Health Information Management Association 6, Fall (2009).

[7]Adversarial Examples: Attacks and Defenses for Deep Learning. Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li, National Science Foundation Center for Big Learning, University of Florida

[8]Adversarial Attacks and Defenses in Images, Graphs and Text: A Review, Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, Anil K. Jain

[9]Adversarial attacks and defenses on AI in medical imaging informatics: A survey by SaraKavianiKi, JinHanInsooSohn

[10] Uwimana, Anisie Senanayake, Ransalu. (2021). Out of Distribution Detection and Adversarial Attacks on Deep Neural Networks for Robust Medical Image Analysis.