Analysis of the Impact of White Box Adversarial Attacks in ResNet While Classifying Retinal Fundus Images

D. P. Bharath Kumar, Nanda Kumar, Snofy D. Dunston, V. Rajam

Analysis of the Impact of White Box Adversarial Attacks in ResNet While Classifying Retinal Fundus Images

Bharath Kumar D P, Nanda Kumar, Snofy D Dunston and Mary Anita Rajam V

Abstract Medical image analysis with deep learning techniques has been widely recognized to provide support in medical diagnosis. Among the several attacks on the deep learning (DL) models that aim to decrease the reliability of the models, this paper deals with the adversarial attacks. Adversarial attacks and the ways to defend the attacks or make the DL models robust towards these attacks have been an increasingly important research topic with a surge of work carried out on both sides. The adversarial attacks of the white box category, namely Fast Gradient Sign Method (FGSM), the Box-constrained Limited Memory Broyden-Fletcher-Goldfarb–Shanno (L-BFGS-B) attack and a variant of the L-BFGS-B attack are studied in this paper. In this work, we have used two defense mechanisms, namely, Adversarial Training and Defensive distillation-Gradient masking. The reliability of these defense mechanisms against the attacks are studied. The effect of noise in FGSM is studied in detail. Retinal fundus images for the diabetic retinopathy disease are used in the experimentation. The effect of the attack reveals the vulnerability of the Resnet model for these attacks.

Bharath Kumar D P
Bharath Kumar D P, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University e-mail: bharath292001@gmail.com

Nanda Kumar
Nanda Kumar, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University e-mail: nandakumar2001au@gmail.com

Snofy D Dunston
Snofy D Dunston, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University e-mail: snofydunston@gmail.com

Mary Anita Rajam V
Mary Anita Rajam V, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University e-mail: anitav@annauniv.edu

1 Introduction

Deep learning models have been used in medical diagnostic systems for a long time now. The deep learning models are well accepted for their performance and their aid in decision making for the medical professionals. Some of the convolutional neural network models commonly used for this purpose are AlexNet, GoogleNet, ResNet and Inception Net. The other deep learning models commonly used are auto encoders, Long Short Term Memory networks, Recurrent Neural Network, Generative Adversarial neural network and so on. However, recent studies reveal that deep learning models could be tampered by attacks known as adversarial attacks. These attacks aim to reduce the reliability and performance of the deep learning system by the introduction of data samples which mislead the system to wrong predictions. These wrong predictions may also have higher confidence value in certain cases.

The adversarial attacks create images with slight variation from the original images. The difference between the newly created images and the original image is imperceptible to human vision. However, to the learning model, the variation leads to misprediction. Some of the methods for generating these adversarial images listed in the literature are Limited Memory Broyden-Fletcher-Goldfarb–Shanno attack (L-BFGS) attack [**?**], Fast gradient sign method (FGSM) [**?**], iterativative least-likely class, Jacobian based Saliency Map attack, Deep Fool, Carlini and Wagner attack (C&W), compositional pattern-producing network-encoded evolutionary algorithm, Zeroth order optimization based attack,Universal Perturbation, One-Pixel Attack, Feature Adversary, Hot/Cold, general generative adversarial networks, Model-Based Ensembling Attack and Ground-Truth Attack [**?**].

The adversarial attacks perturbate the original input, then present them to the model and lead to misclassification. Based on the intention of misclassification, the attacks are classified as targeted attacks and untargeted attacks. In targeted attacks, the intention of the adversaries is to misclassify the records to a particular class. In untargeted attacks, the intention of the adversaries is to misclassify the records to any other class other than its true class.

Based on the knowledge of the model being targeted, the attacks are classified into White box attack, Black box attack and Grey box attack. In case of white box attacks, the structure and parameters of the model are known to the adversaries. The structure of the model alone is known to the adversaries in the case of grey box attacks. In case of black box attacks, these details are not known to the adversaries.

Defense mechanisms are used to improve the robustness of the DL models against adversarial attacks. The defense mechanisms to countermeasure adversarial attacks can be proactive or reactive. Various defense methods have been proposed in literature. Adversarial detecting, input reconstruction and network verification are some of the reactive methods, and adversarial training and network distillation are some of the proactive methods.

In this paper, the vulnerability of Resnet model towards the optimization based L-BFGS-B, and the gradient based Fast Gradient Sign method (FGSM) white-box adversarial attacks is studied. These attacks are studied on the retinal fundus images of the Diabetic Retinopathy (DR) patients and their effects are discussed. The suc-

cess rate of the attacks and the trade-off between producing less distorted and more confident adversaries is studied. We have used two defense mechanisms, namely, Adversarial Training and Defensive distillation-Gradient masking. The reliability of these defense mechanisms against the attacks are tested.

The rest of the paper is organized as section II on the related work, section III on the different attacks used in this work, section IV on results and discussion and section V concludes the paper.

2 Literature Survey

The study of universal perturbation as an adversarial attack was done by Guohua Cheng and Hongli Ji on U-Net. It was found that, among the four modalities in the brain MRI images, if all the modalities were altered the perturbation has an effect on the classifier [**?**]. Hassan et al. demonstrated the FGSM attack and Basic Iterative Model of FGSM on a ResNet model with 85 public datasets which contained time series data [**?**].

Gerda Bortsova et al. experimented the FGSM and projected gradient descent (PGD) attacks on Inception-v3 and Densenet121. The datasets used were diabetic retinopathy dataset with 88,702 color fundus images, ChestX-Ray14 dataset with 112,120 frontal-view X-rays and patchCamelyon (PCam) with 327,680 patches extracted from histopathology whole-slide images of lymph node sections [**?**].

Yupeng Cheng et al. examined the camera exposure to the Retinal Fundus images and proposed an exposure based adversarial attack on the retinal fundus images. Their experiments were carried out in three stages as multiplicative-perturbation based attack, adversarial bracketed exposure fusion and convolutional bracketed exposure fusion. The convolutional bracketed exposure fusion was found to be more effective in Resnet50, MobileNet and EfficientNet [**?**].

Saeid Asgari et al. have demonstrated 10 attacks belonging to gradient based, score based and decision based adversarial attacks on chest X-ray images with Inception-ResNet-V2 and Nasnet-Large neural network models. The gradient based attacks were superior in fooling the network in comparison to the other attacks. Secondly, by modifying the pooling layer of the network, the effect of adversarial attacks were reduced [**?**]. A susceptibility score has been proposed by Mengying Sun et al. using the global maximum perturbation (GMP), global average perturbation (GAP) and the probability of being perturbed across all records (GPP). This score provides the efficiency of an attack by uncovering the locations which are susceptible for an attack. A perturbation distance has been defined by the authors and is used to find the optimal adversarial records [**?**].

Rida El-Allami et al. have performed a study on the Spiking neural network's robustness against adversarial attacks. The study reports that the parameters of the network namely Spiking Voltage threshold, spiking time window and attack's noise budget play a major role in the robustness of the network [**?**].

3 Materials and Methods

3.1 Attacks studied

The attacks studied in this work are explained in this section.

FGSM Attack

The goal of FGSM attack is to find the optimal direction in the input space where the loss function ascends/descends steeply and so pushing the inputs in that direction will find an input that is close to the given image and also is misclassified.

The perturbed input x' is given by

$$x' = x + \varepsilon \cdot sign[\nabla x J(\Theta, x, y)] \tag{1}$$

where

x is the input value, ε is the epsilon value that determines the magnitude of the perturbation allowed on a pixel value, $J(\Theta, x, y)$ is the loss function and ∇x is the gradient of the loss function.

L-BFGS-B Attack

In this type of attack, finding adversarial examples is modeled as an optimization problem. The optimization aims at finding an adversarial example with minimum perturbation and classifies into non-true class.

The perturbed input x' is found by solving the below minimization (optimization) equation

$$f(x) = c \cdot L2norm(x - x') + loss(x') \tag{2}$$

where x' is chosen between 0 to 1

3.2 Deep neural network classifier used

The basic deep learning model used in this work is Resnet50, a variant of ResNet model with 48 Convolution layers, one MaxPool layer and one Average Pool layer. It has been widely used for transfer learning in the classification of medical images. The final network architecture used in this work consists of additional layers to the Resnet50 namely, a global average pooling layer, dropout layer and the final classifying dense layer. The model is initialized with pretrained weights from ImageNet and then trained on the retinal fundus dataset specified in Section 4.1, that is, transfer learning from ImageNet is done. The ImageNet dataset contains 14,197,122 anno-

tated images and is used in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a benchmark in image classification and object detection.

3.3 Defense mechanisms studied

This section describes the defense mechanisms studied in this work.

3.3.1 Adversarial training

In adversarial training, adversarial samples that are misclassified by the classifier are created. The neural network classifier is then trained with the created samples and their correct labels.

3.3.2 Defensive distillation

Hinton et al. [**?**], introduced knowledge distillation, a process to extract knowledge from one model and apply it to a compressed model and make it perform with the same accuracy. Papernot et al. [**?**] exploited this process and modified it to act as a defense. Distillation, is training a teacher model by introducing a temperature parameter T in the final softmax activation. The softmax output of the teacher model (soft labels) is used for training another model with the same temperature T.

The intuition behind this is that the knowledge acquired by the model during training is encoded not only in its weights, but also in the output probability vector, which holds the relative information about classes. Informally, in case of hard labels, the model only gets to know that it belongs to one particular class and no information about its similarity with the remaining classes. All samples belonging to a class are treated with the same weights, even though some samples might be less similar to that class. Soft labels use a vector of probabilities and the similarities of the given sample, with all classes, are learned by the model. This helps the model understand the relative difference between the classes.

3.4 Details of the study on the impact and working of Adversarial attacks

A binary classifier is first built using the deep learning model specified in Section **??**, which classifies retinal images as 'DR affected' or 'Unaffected'. The FGSM and the L-BFGS-B attacks are performed on 100 image samples which were correctly classified by the Resnet classifier model used as 'DR affected' or 'Unaffected'. The number of samples in each class is chosen to be approximately the same. Different

metrics are studied for the attacks performed. The robustness of the Resnet model against the FGSM and the L-BFGS-B attacks using adversarial training and distillation defense mechanisms is studied. Various studies are also done on the perturbations generated using the attacks.

4 Results and Discussion

4.1 Dataset

The dataset used for training the Resnet classifier model has 35,638 images, out of which 35,126 retina scan images are taken from the diabetic retinopathy(DR) Kaggle dataset and 512 retina scan images are taken from IDRiD(Indian Diabetic Retinopathy Image Dataset) dataset.

The classes in the dataset are grouped into two classes as affected by DR and unaffected by DR. The dataset is first balanced and the images are resized to 224 x 224 pixels. The dataset is divided into train set, validation set and test set. The details of the dataset used for training are given in table **??**

Table 1: Details of training dataset

Dataset Name	Number of images	Unaffected by DR	Affected by DR	Number of images in Training set	Validation set	Testing set
Kaggle	18966	9556	9400	14956	3720	290
IDRiD	554	184	370	444	-	110

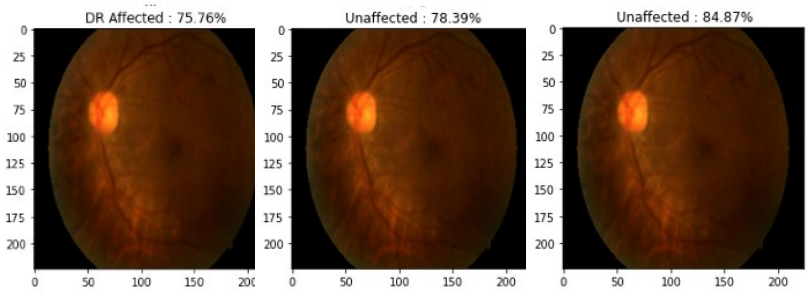
4.2 Hyper parameters used for tuning the Resnet model

Batch Size	A batch size of 16 is used for training since it provides better generalization and is small enough to fit in memory.
Learning Rate	A learning rate of 0.0001 is found to give the best results based on experimentation.
Loss function	Categorical cross entropy loss function is used in the training process since the model is a multi-class classifier.
Optimizer	Several optimizers were tested for stochastic gradient descent parameter updates namely Adam, RMSprop and SGD. No optimizer showed significant increase in performance over the other two. It is found that the Adam optimizer performed the best overall.

4.3 Attack framework

100 original retina image samples that are correctly classified by the Resnet model explained in Section **??** are used for implementing the attacks. As it is a binary classification problem, only targeted attacks have been tested, and we believe untargeted attacks should also provide similar results. The norm or distance metric is taken as L2 norm because it is widely used in the literature for the implementation of FGSM and L-BFGS-B attacks.

4.4 Sample images after the adversarial attacks



(a) Sample image (b) FGSM attack (c) L-BFGS-B attack

Fig. 1: Effect of Adversarial attacks

Figure **??** depicts the effects of the different adversarial attacks considered on a sample image from the dataset. While the sample image is classified as DR affected with a confidence of 75.76%, after the FGSM and the L-BFGS-B attacks, the image is classified as unaffected with confidence values of 78.39% and 84.87% respectively. Thus, it is seen that the attacks were successful in misleading the classifier.

4.5 Metrics used for evaluation

Success Rate The number of adversarial examples created successfully.
Confidence Confidence value for the classification of the adversarial image.
Strong adversaries The number of adversarial examples which were classified with more confidence than their original counterparts (here, out of the 100 samples considered).
Norm L2 distance between the original image and perturbed image.

4.6 Analysis of the effect of the FGSM attack

Experimentation was done with the FGSM attack with different epsilon values namely 0.1, 0.25, 0.5, 1, 2 and the results are tabulated in Table **??**. It is seen that, when the epsilon value is 0.25, FGSM produces adversarial images with minimum average norm value and achieved a success rate of 95%.

The step size in the direction of gradient is controlled by epsilon, with smaller values of epsilon giving less distortion and adversaries with low confidence. It is observed that, very high epsilon values also did not help, as it prevented the gradient descent to converge to an optimal perturbation and so the success rate decreased for very high values of epsilon. (Started decreasing from epsilon = 3)

Table 2: FGSM results

epsilon	Success rate (%)	Avg.conf (%)	Max.conf (%)	Avg.norm	Max.norm	Strong adversaries
0.1	85	69.78	94.37	38.77	38.78	64
0.25	95	83.54	98.78	96.9	97	71
0.5	99	91.62	99.9	193.98	193.98	86
1	99	94.77	99.98	387.97	388	92
2	99	95.12	99.97	775.95	776	92

4.6.1 Analysis of the effect of the L-BFGS-B attack

Experimentation is done with the Box constrained L-BFGS attack and the results are tabulated in Table **??**. L2-norm is used as the distance metric, and the constant c, which measures the relative importance of the norm and loss in the objective function is computed using binary search. To control the perturbation, a parameter, epsilon is used. This value controls the lower and the upper thresholds of the box. High values of epsilon give large perturbation and the confidence of adversaries is increased.

We have also tried a variant of the L-BFGS-B attack by removing the norm constraint in the objective function and optimizing only the loss function. The motive is to observe the extent of optimization without the norm constraint and whether this could produce images with acceptable perturbations with a better confidence level. The results are tabulated in Table **??**.

The perturbation increased as expected, but the confidence remained almost similar. The same maximum confidence was observed and the average confidence increased slightly. The effect of epsilon was minimal and the results stayed almost constant. So, it is concluded that the cost given for high perturbation is not balanced

Table 3: L-BFGS-B attack

epsilon	Success rate (%)	Avg.conf (%)	Max.conf (%)	Avg.norm	Max.norm	Strong adversaries
0.07	61	66.891	92.01	29.1	31.4	36
0.2	90	76.71	96.23	70.7	75.1	71
0.5	100	99.43	99.86	175	189.76	99
1	100	99.43	99.86	346.35	360.99	100

Table 4: Modified L-BFGS-B attack

epsilon	Success rate (%)	Avg.conf (%)	Max.conf (%)	Avg.norm	Max.norm	Strong adversaries
0.07	100	96.2	99.92	171.56	421.79	100
0.2	100	99.21	99.93	146.56	301.57	100
0.5	100	99.65	99.94	158.65	279.88	100
1	100	99.66	99.87	250.52	1067.82	100

in the confidence level and so the previous method (L-BFGS-B) produces better adversaries. Though the confidence value is higher compared to the unmodified attack, the norm value is also higher. The higher norm values indicate that the difference in the images may be much visible.

4.6.2 Experiments using perturbation

For the creation of adversaries, the attack models generate perturbation that are added to the original image. We wanted to analyse the impact of the noise / perturbations generated by the attacks.

Figure **??** shows the perturbations / noises created by FGSM to transform a 'DR affected' sample to 'Unaffected' for different values of epsilon. It is seen that the perturbation is high when the value of epsilon is larger.

We first try to find how the Resnet classifier classifies, when just the perturbation is given as input to the classifier. Figure **??** shows a perturbation image created by FGSM to push the classification towards 'Unaffected' class. Interestingly, this noise (just the perturbation) shown in figure **??** is classified by the classifier as 'Unaffected' with 81.68% confidence.

Now, if this perturbation in figure **??** is given as input further to FGSM, it successfully completes the attack on the perturbation image and the resultant image (perturbation using FGSM + FGSM) is classified with the Resnet classifier as DR affected with a confidence of 55.42% (Figure **??**). Though the attack is successful, the low confidence value signifies again that the original perturbation given as input had the properties of unaffected retina image strongly, as expected.

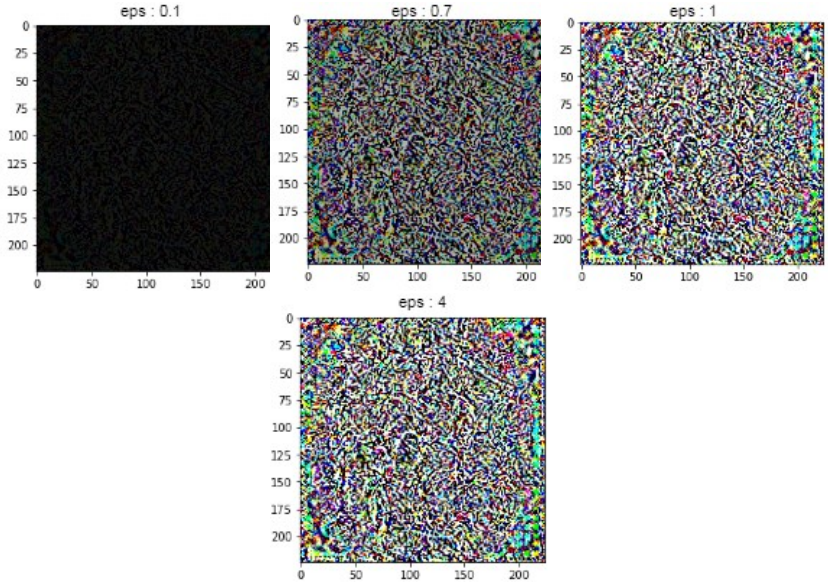


Fig. 2: Perturbation images created by FGSM for different epsilon values

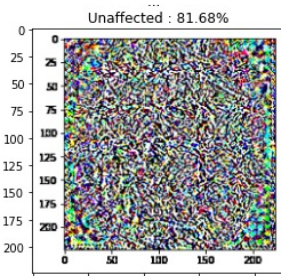


Fig. 3: Perturbation created by FGSM to push towards Unaffected class

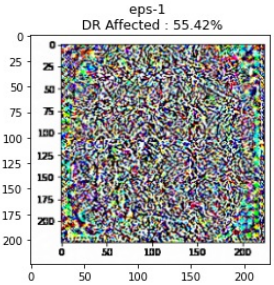


Fig. 4: FGSM on perturbation created by FGSM

When the perturbation in figure **??** is given as input further to L-BFGS-B (perturbation using FGSM + L-BFGS-B), the resultant image is misclassified as DR affected with 61.3% confidence and a high norm of 350.5 (Figure **??**). So, now we have created an artificial image sample of an unaffected retinal fundus image.

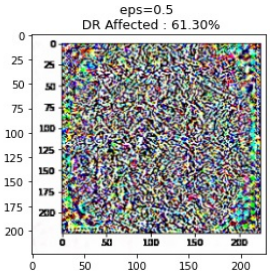
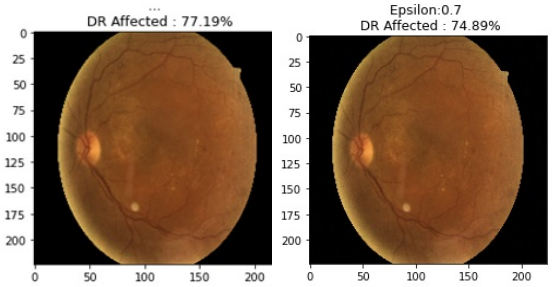


Fig. 5: L-BFGS-B on perturbation created by FGSM

We have also tested if the perturbation created from one input image is enough to convert another image to an adversary. For this, we added the perturbation obtained from one image of the dataset on a different image of the dataset and tested if the second image behaved as an adversary. When we add the perturbation shown in Figure **??** to a different DR affected image (Figure **??**), the result shown in Figure **??** is got.



(a) DR affected input (b) After adding perturbation

Fig. 6: Image before and after adding the perturbation shown in figure **??**

The confidence is observed to have reduced. Even though the added noise didn't misclassify the retinal image as unaffected, it has pushed the sample towards 'Unaffected' direction and we believe a stronger noise sample will perform the misclassification too, proving the claim by Goodfellow et al. [**?**] that the adversarial directions stay the same for all samples within a training set.

To examine whether any random noise can cause a misclassification, a random noise is generated (Figure **??**) and given as input to the Resnet classifier. This noise is unrelated to the DR image but the classifier classifies the image as 'Unaffected' with 73.24% confidence.

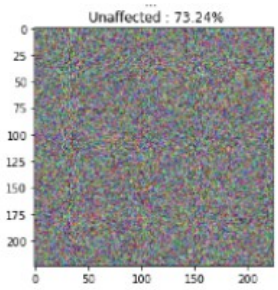


Fig. 7: Random noise

When the L-BFGS-B attack is applied on this random noise shown in Figure **??**, the image was easier to attack and it produced an adversarial image with less epsilon value of 0.39 and with a less norm of 130.5 (Figure **??**).

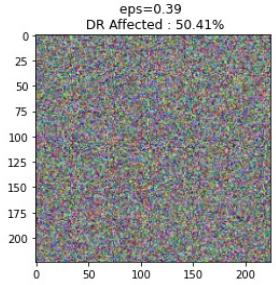


Fig. 8: Adversary created from Random noise

This observation signifies that the perturbation noise added in FGSM is not just a random noise but carefully structured one to push a benign sample towards target direction and so it captured the features of Unaffected image, which made it hard to misclassify it as seen above.

4.7 Defenses

4.7.1 Adversarial Training

We created Adversarial examples with FGSM and retrained the Resnet model along with the adversarial examples with the correct labels.

We tested the adversarial training defense mechanism against the FGSM attack and the results are summarized in Table **??**. It is observed that training with the adversarial examples decreased the success rate of the attacks and decreased the confidence of the adversaries while the norm remained the same. However,it couldn't still prevent attacks and misclassifications. The number of strong adversaries reduced to 0, showing that the adversaries produced had low confidence values. The more the adversarial samples we used for training, the less the confidence of the adversaries.

Table 5: FGSM results after adversarial training

epsilon	Success rate (%)	Avg.conf (%)	Max.conf (%)	Avg.norm	Max.norm	Strong adversaries
0.1	38	57	94	38.77	38.78	0
0.25	41	56	98.78	96.9	97	0
0.5	44	56	99.9	193.98	193.98	0
1	44	55	99.98	387.97	388	0

The results of the adversarial training defense mechanism against the L-BFGS-B attack are summarized in Table **??**. It is seen that norm values decreased more than confidence values. The number of strong adversaries reduced to 0 again, similar to FGSM attack. The success rate didn't reduce but as the number of strong adversaries is 0, it is inferred that the produced adversaries are of low confidence as expected.

Table 6: Modified L-BFGS-B results after adversarial training

epsilon	Success rate (%)	Avg.conf (%)	Max.conf (%)	Avg.norm	Max.norm	Strong adversaries
0.2	100	76.71	96.23	78.7	163.7	0
0.5	100	96	99.86	115	148.44	0
1	100	97.6	99.86	205.46	246.36	0

4.7.2 Defensive distillation

We trained the teacher network with a softmax output at temperature 100. The teacher network has the same architecture as the previously discussed ResNet model. We utilized the same hyperparameters that were previously shown to be the most ideal. We used a custom loss function that computes the softmax cross entropy between logits and labels.

A new training set was formed with soft labels obtained from the teacher model. We trained the student model with the same network architecture as the teacher model using the new training set and the temperature of the softmax layer was kept at 100. This new model is referred to as the distilled model.

We tested the distilled model against the two adversarial attacks and the results are summarized in Table **??**.

Table 7: Success rate of attack after defensive distillation

Adversarial attack	Success rate against original model	Success Rate against Distilled Model
FGSM	99%	60%
L-BFGS-B	100%	82%

5 Conclusion

The paper aims to demonstrate the applicability of adversarial attacks and their implications on classifying retinal images. This enforces the importance of making neural networks robust and the need to combat these attacks. The experiments on the FGSM and the L-BFGS-B attacks have revealed that the performance of the ResNet model could be reduced by the adversarial attacks. Defense mechanisms can reduce the effect of the attacks. It is also observed that a random noise may also pose the risk of reducing the confidence of the classifier in its true prediction.

Acknowledgements This research is funded by Science & Engineering Research Board (SERB).

References

1. Asgari Taghanaki, S., Das, A., Hamarneh, G.: Vulnerability Analysis of Chest X-Ray Image Classification Against Adversarial Attacks. Understanding and Interpreting Machine Learn-

ing in Medical Image Computing Applications. 87-94 (2018), doi:10.1007/978-3-030-02628-8_10
2. Bortsova, G., González-Gonzalo, C., Wetstein, S., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., van Ginneken, B., Pluim, J., Veta, M., Sánchez, C., de Bruijne, M.: Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Medical Image Analysis. 73, 102141 (2021), doi:10.1016/j.media.2021.102141
3. Carlini, N., Wagner, D.: Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy (SP). (2017),
4. Cheng, G., Ji, H.: Adversarial Perturbation on MRI Modalities in Brain Tumor Segmentation. IEEE Access. 8, 206009-206015 (2020), doi:10.1109/ACCESS.2020.3030235
5. Cheng, Y., Juefei-Xu, F., Guo, Q., Fu, H., Xie, X., Lin, S., Lin, W., Liu, Y.: Adversarial Exposure Attack on Diabetic Retinopathy Imagery. arXiv. (2020), arXiv:2009.09231
6. Christian, S., Wojciech, Z., Ilya, S., Joan, B., Dumitru, E., Ian, G., Rob, F.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. (2013),
7. El-Allami, R., Marchisio, A., Shafique, M., Alouani, I.: Securing Deep Spiking Neural Networks against Adversarial Attacks through Inherent Structural Parameters. 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE). (2021), 10.23919/DATE51398.2021.9473981
8. G. Hinton, O. Vinyals, and J. Dean: Distilling the knowledge in a neural network. Deep Learning and Representation Learning Workshop at NIPS 2014. arXiv preprint arXiv:1503.02531. (2014),
9. Ian J, G., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572. (2014).
10. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.: Adversarial Attacks on Deep Neural Networks for Time Series Classification. 2019 International Joint Conference on Neural Networks (IJCNN). (2019), 10.1109/IJCNN.2019.8851936
11. Newaz, A., Haque, N., Sikder, A., Rahman, M., Uluagac, A.: Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems. GLOBECOM 2020 - 2020 IEEE Global Communications Conference. (2020), 10.1109/GLOBECOM42002.2020.9322472
12. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. 2016 IEEE Symposium on Security and Privacy (SP), pp. 582-597, doi: 10.1109/SP.2016.41, (2016).
13. Sun, M., Tang, F., Yi, J., Wang, F., Zhou, J.: Identify Susceptible Locations in Medical Records via Adversarial Attacks on Deep Predictive Models. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (2018), 10.1145/3219819.3219909
14. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial Examples: Attacks and Defenses for Deep Learning. IEEE Transactions on Neural Networks and Learning Systems. 30, 2805-2824 (2019), doi: 10.1109/TNNLS.2018.2886017