Analysis of Adversarial Attacks on Skin Cancer Recognition

Aminul Huq
Department of Computer Science & Technology
Tsinghua University
Beijing, China
aminul.huq11@gmail.com

Mst. Tasnim Pervin
Department of Computer Science & Technology
Tsinghua University
Beijing, China
tasnimpervin@gmail.com

***Abstract*—Cancer is one of the most detrimental diseases having the highest death rates in recent years. There are various types of cancers; among them, skin cancer is the most familiar one. Early detection and treatment of it can lead to full recovery for the patients. Deep learning based image classification models have been proven to perform exclusively well to classify images. However, in recent years researchers have shown that adding small calculated noises can induce these models to generate wrong answers. In this regard, here we performed adversarial training based on Projected Gradient Descent (PGD) to increase the robustness of two popular deep learning models, namely MobileNet and VGG16, against white-box attacks of PGD and FGSM attacks. We performed our experiments on a dataset of 10015 images and have shown that our models are much robust than standard training ones and achieved almost similar results as them.**

***Index Terms*—Skin Cancer, Adversarial Machine Learning, Deep Learning, PGD Attack**

I. Introduction

In a human body, the most extensive part that occupies a person is skin. It is also the first section that is evaluated by a practitioner in medical science when they see patients as it provides a lot of insights and background information about the patients' situation [1]. Due to human habits like alcohol intake, smoking, lifestyle, current environment and other factors, human beings are more prone to getting diagnosed with cancer. Skin cancer is the most common form of cancer, according to the American Nurses Association. If diagnosed early enough it can be treated properly. However, it also causes problems as it is similar to some other benign lesions and thus makes this task very difficult. There are various types of skin cancers. Here, we mention some of the common ones, which are melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis/Bowen's disease (intraepithelial carcinoma), etc.

From the beginning of the twenty-first century, deep learning models have been effectively classifying images in various sectors such as medical imaging, self-driving cars, face recognition, and many others [2]–[4]. Deep learning models use lots of layers for identification of complex hidden features of the images to discriminate and perform classification. In this experiment, we have implemented two deep learning models called MobileNet and VGG16 [5], [6]. They are very popular models and have been quite effective against the ILSVRC dataset.

In recent years adversarial machine learning has been an emerging research field. It is mainly comprised of different techniques to fool the classification models into providing wrong results by adding tiny noises that are imperceivable to humans. It also includes how to defend from these attack techniques as well. These issues were first recorded by C. Szegedy [8]. After this, many innovative attacking models were proposed like FGSM, MI-FGSM, DeepFool, CW, PGD, Universal Adversarial Perturbation, and many more [7]–[13]. In order to defend from these attacks many defense schemes have also been proposed including image quilting, cropping, JPEG compression [14]–[16]. However, adversarial training has been proved as the most effective defense technique among all.

To make a robust model for skin cancer classification, we used MobileNet and VGG16 classification model on 10015 images. We performed Projected Gradient Descent (PGD) and Fast Gradient Sign Method (FGSM) based white-box attack on these models to test their robustness. We also analyzed the transferability effect of PGD approach as well. After that, we performed adversarial training on these models and later checked their robustness by performing white-box attacks.

Our contribution can be summarized as follows:

- We implemented high performing MobileNet and VGG16 models for correctly identifying skin cancers from images into seven different categories.
- PGD and FGSM based white-box attack have been implemented here to test the robustness of these models.
- Adversarial training based on PGD has been performed to increase model robustness, and it has been checked against white-box attacks.
- Analysis of white-box attacks and adversarial image transferability has been given here for skin cancer-based images.

This paper is arranged in the following manner. Section 2 provides a small discussion about previous works, classification models that have been used here, introduction to adversarial machine learning and the attack models. Description of the dataset, model training, results and analysis of the experiment

is shown in section 3. Finally we have provided our conclusion in section 4.

II. Background Studies

A. Related Works

Many researchers have worked on this field in the past and research is still going on as this is an important topic. R. Oliveira and his fellow researchers provided an excellent discussion about different feature extraction techniques, models for feature selections, and also future trends in their review paper [17]. They compared their results with other research works. Working with 350 images and with multiclass SVMQ. Abbas et al. showed their approach achieved 93% accuracy [19]. B. Janney and his fellow researchers performed adaptive k-means clustering to segment the melanoma from the background. They then performed feature extractions, which were mainly focused on texture and intensity-based. For classification purposes, they used KNN, SVM, MLP, Decision Tree, and Random Forest approach [20]. Esteva and his fellow researchers worked with CNN on 129,450 clinical images, which consisted of 2,032 different diseases. The results they achieved were in a level of competence comparable to dermatologists [21]. A hybrid feature extraction technique has been used by E.A. Al-Mansour and his fellow researchers [22]. They used LBP, and GLCM approaches for local feature and color based global features for this hybrid representation. For classification purposes, they used SVM. Stoecker et al. proposed the use of statistical approaches such as a gray level co-occurrence matrix to analyze texture in skin images, which can accurately be used for both segmentation and classification from dermoscopy images [23]. In order to classify 12 skin diseases, Han and his fellow researchers used CNN to achieve around 96% accuracy [24]. U.O. Dorj used a pre-trained AlexNet model for feature extraction and implemented ECOC SVM for classification [18]. They chose their dataset by searching on the internet. M.A. Kadampur et al. created a system in the cloud, which was based on model-driven architecture and used deep learning algorithms to classify images with 99.77% accuracy [25].

B. Classification Models

We have used here MobileNet and VGG16 for this experiment. These models are very well defined and established as they have been tested against the ImageNet dataset, and their top 5 accuracy is 89.5% and 90.1%, respectively [28].

1) MobileNet: MobileNet introduces a low latency and efficient network configuration specifically helpful for embedded vision applications and mobile devices [5]. This lightweight model became much popular because of the fewer number of parameters and fewer multiplications-additions. A set of two parameters, Width Multiplier (α) and Resolution Multiplier (ρ), have been used for easy tuning. For adding filtering effect on the input channels, depth-wise separable convolutions have been used for this model by which a single convolution is performed on every color channel and flattened despite utilizing a combination of all three channels at once. After that, point-wise convolution (1x1 convolution) have been addressed to combine the depth-wise convolution outputs. The combination of depth-wise convolution and point-wise convolution is named as depth-wise separable convolution. This approach sums two layers: a separable layer for filtering and a layer for combining. The use of 3x3 depth-wise separable convolutions helped the model to reduce computation drastically compared to the standard convolution. This model needs comparatively low maintenance and operates with high speed.

2) VGG16: VGG16 is one of the modifications of the convolutional neural network from ILSVRC-2014 proposed by Simonyan and A. Zisserman from the University of Oxford [6].

Large kernel-sized filters of AlexNet has been replaced by a row of smaller filters (kernel size: 3x3). The first convolutional layer of filter size 3x3 with a single stride works as an input layer of the model accepting RGB images of size 224x224. The spatial padding of the convolutional layers is maintained accordingly to protect the spatial resolution after being convolved. Spatial pooling is done by five max-pooling layers of filter size 2x2 with stride two after the first convolutional layers. After that, three fully connected layers follow previous max-pooling layers. The depth of fully connected layers differs from architecture to architecture. For each class, the first two layers have 4096, and the third layer has 1000 channels approximately. All these hidden layers use ReLU non-linearity. The softmax layer is the final output layer, which predicts the probability of classes. Among these layers, only a single network has Local Response Normalization (LRN), resulting in significant memory allocation and amplified computation time.

C. Adversarial Machine Learning

Modern machine learning and deep learning has achieved a whole new height because of its high computational power and fail proof architecture. However, recent advances in adversarial training have broken this illusion. A compelling model can misbehave by a simple attack by adversarial examples.

An adversarial example is a specimen of input data that has been slightly transformed in such a way that can fool a machine learning classifier resulting in mis-classification. The main idea behind this attack is to inject some noise to the input to be classified so that the resulting prediction is changed from actual class to another class. After adding these noises the perturbed image appears similar to original image to human eyes. Thus we can understand the threat of this kind of attack to classification models.

Based on the goal or intention of attacker it can be classified into two types.

- Targeted attack: The goal of a targeted attack is to make the model classify inputs as a specified target class that is defined by the attacker.
- Non-targeted attack: This type of attack does not have any preference or target, it just fools the generally unknown model by slightly changing input making it unable to

predict correctly. The main purpose here is to reduce model accuracy.

Depending on the knowledge of the attacker about the model attacks can be classified into two types.

- Black-box Attack: This is the most real-life scenario where the adversary does not have any access to the related parameters of the targeted model, so the adversary tries to break the model with either random assumption or calculated guess with trial and error.
- White-box attack: In this kind of attack, the adversary has access to the underlying model parameters. As the training policy of the model is known to the adversary, it is prominent to impose a drastic effect on the model performance.

1) Fast Gradient Sign Method (FGSM): Proposed by I. Goodfellow and his fellow researchers FGSM is one of the earliest, simplest, and fast adversarial attack techniques [8]. It can be used for white-box, targeted, and non-targeted attacks. It generates adversarial perturbation by looking into the gradients. This approach calculates the loss across the predicted labels and original labels and calculates gradients across all images. Then takes the sign of the gradients and uses it to generate noises.

$$x_{adv} = x + \epsilon * sign(\nabla_x L(\theta, x, y) \tag{1}$$

Here, x_{adv} represents the adversarial image of the original image x, y represents the original label of that image, ϵ is the perturbation to add, L calculates the loss function and ∇_x is the gradient with respect to x.

2) Projected Gradient Descent (PGD): A. Madry et al. studied robustness of neural networks through the lens of robust optimization. The formulation which they constructed is based on natural saddle point (min-max) to capture the notion of security against adversarial attacks [12]. This is an iterative attack which starts from random positions to determine the best spot for perturbing images. This also has much better potential in adversarial training for improving model robustness. It can be represented as

$$\begin{gathered} min_\theta P(\theta) \\ where, P(\theta) = E_{(x,y)\sim D}[max_{\delta \epsilon S} L(\theta, x+\delta, y)] \end{gathered} \tag{2}$$

In this equation, we can see that the inner maximization aims at achieving higher loss. At the same time, the outer side minimization tries to attain model parameters in such a way that the inner attack is minimized. In order to perform a PGD attack, we can start with random perturbation in the L2 or L-infinity ball around a sample. After that, we can take a gradient step in the direction of the most considerable loss and project the perturbation into the same direction and repeat until we get convergence.

III. Experimental Results & Analysis

A. Dataset Description

The dataset that has been used here is called "Humans Against Machines with 10000 Training Images" HAM10000 [27]. These dermatoscopic images were collected from different populations, acquired and stored by different modalities. The overall dataset comprises of 10015 dermatoscopic images. It is mainly part of the ISIC 2018 competition [26]. Only the training set is made public and for this reason; we used this data and divided it into three parts of training, validation, and testing sets. We trained on 8111 samples performed validation on 902 samples and tested against 1002 samples. In total, there are seven classes to classify here called: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis/Bowen's disease (intraepithelial carcinoma), benign keratosis (solar lentigo/seborrheic keratosis/lichen planuslike keratosis), dermatofibroma and vascular lesion. All the images have been reshaped to 224 x 224 x 3 dimensions. Since all the images are RGB images, the pixel values are in the range of 0 to 255. We scaled the values from 0 to 1 for simplicity.

B. Model Training

Both models were trained twice for this experiment. One of them is called standard or regular training, and the second is called adversarial training. The parameters that were used here are mentioned below:

- Loss Function: Categorical Cross-Entropy
- Optimizer: SGD
- Learning Rate: 0.1
- Momentum : 0.9
- Batch-size : 16

We set a reduction factor of 0.5 to the learning rates after monitoring three epochs if the validation accuracy did not improve.

For standard training, we used the original data. We then tested its robustness against two attack techniques PGD and FGSM. In order to perform adversarial training, we need adversarial images to train the model along with original images. We created adversarial image data based on original data for the purpose of performing adversarial training.

C. Results & Analysis

The results of this experiment are shown in Table 1, and 2. We generated the same number of adversarial images as the test set for both MobileNet and VGG16 models. As we can see with regular training of both models, we achieve a moderate testing accuracy of 77.24% & 71.72% for MobileNet and VGG16, respectively. With PGD based adversarial images, both models performed very poorly, and accuracy came down to 2.89% and 8.68% respectively. The same is the situation for FGSM based adversarial images. It pulls down the accuracy of these models to 2.49% and 9.88% respectively. Thus indicating that these models are indeed vulnerable to adversarial noises. We now look at the transferability of these attacks. We tested the attack images created using MobileNet on both VGG16 and MobileNet model. Again, we did the same but this time we generated the images on VGG16 and checked against VGG16 and MobileNet. Table 2 shows the results. When testing MobileNet's robustness against VGG16's generated images, the adversary performs 6.08% accuracy, and for VGG16, it

achieves 12.77% accuracy against MobileNet's adversarial images. Hence, we can conclude that these adversarial attacks can be transferable as well.

TABLE I
WHITE-BOX TESTING ON STANDARD TRAINING

	Testing Accuracy(%)	PGD (%)	FGSM (%)
Standard Training (MobileNet)	77.24	2.89	2.49
Standard Training (VGG16)	71.72	8.68	9.88
Adversarial Training (MobileNet)	76.14	**75.94**	**63.17**
Adversarial Training (VGG16)	70.47	**71.05**	**50.79**

After introducing adversarial training based on PGD attacks, we again perform similar tests as the previous white-box. From table 1, we can see that after adversarial training, these models' performance improves significantly. These models are now more robust than before and provides much better accuracy against adversarial examples. We can see that after adversarial training, results against adversarial test sets are almost similar to standard training test results.

TABLE II
TRANSFERABLITY OF ADVERSARIAL EXAMPLES

	MobileNet (%)	VGG16 (%)
MobileNet	2.89	6.08
VGG16	12.77	8.68

IV. CONCLUSION

We employed popular deep learning models for classifying skin cancer from images and experimented with their robustness against popular white-box attacks. In order to overcome these issues, PGD based adversarial training has been performed, and it has been shown that after conducting adversarial training, these models' robustness increase significantly. Also, from the White-box attack as well, we can see that PGD based adversarial training has good generalization value in defending against other attacks. Other adversarial training on FGSM, CW do not have a good generalization score [12]. This experiment was performed on the publicly available dataset which lacks the sufficient number of training samples . In the future, we would like to expand the dataset and also include other attack and defense models for adversarial attacks.

REFERENCES

[1] B.A. Lowell, C.W. Froelich, D.G. Federman and R.S. Kirsner. "Dermatology in primary care: Prevalence and patient disposition", *Journal of the American Academy of Dermatology (JAAD)*, vol. 45, no. 2, pp. 250-255, August, 2001.

[2] Y. Xue et al. "Application of deep learning in automated analysis of molecular images in cancer: a survey", *Contrast Media & Molecular Imaging*, 2017.

[3] J. Kim and C. Park. "End-to-end ego lane estimation based on sequential transfer learning for self-driving cars", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1194-1202. 2017.

[4] Y. Sun, D. Liang, X. Wang and X. Tang "Deepid3: Face recognition with very deep neural networks", *arXiv preprint arXiv:1502.00873* , 2015.

[5] A.G. Howard et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications", *arXiv preprint arXiv:1704.04861*, 2017.

[6] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition", *International Conference on Learning Representations*, 2015.

[7] C. Szegedy et al. "Intriguing properties of neural networks", *International Conference on Learning Representations*, 2014.

[8] I. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples" *International Conference on Learning Representations*, 2014.

[9] Y. Dong et al. "Boosting adversarial attacks with momentum", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185-9193, 2018.

[10] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. "Deepfool: a simple and accurate method to fool deep neural networks", *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2574-2582, 2016.

[11] N. Carlini, and D. Wagner. "Towards evaluating the robustness of neural networks", *IEEE Symposium on Security and Privacy (sp)*, 2017.

[12] A. Madry et al. "Towards deep learning models resistant to adversarial attacks", *International Conference on Learning Representations*, 2018.

[13] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. "Universal adversarial perturbations", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765- 1773. 2017.

[14] C. Guo, M. Rana, M. Cissé and L.V.D. Maaten, "Countering adversarial images using input transformations", *International Conference on Learning Representations*, 2018.

[15] A. Prakash, N. Moran, S. Garber, A. Dilillo and J.A. Storer, "Deflecting adversarial attacks with pixel deflection", *IEEE Conference on Computer Vision and Pattern Recognition* pp. 8571-8580, 2018.

[16] C. Xie, Y. Wu, L.V.D. Maaten, A.L. Yuille and K. He, "Feature denoising for improving adversarial robustness", *IEEE Conference on Computer Vision and Pattern Recognition* pp. 501-509, 2019.

[17] R. B. Oliveira, J. P. Papa, A. S. Pereira and J. M. R. S. Tavares, "Computational methods for pigmented skin lesion classification in images: review and future trends". *Neural Computing & Application* vol. 29, no. 2, pp. 613–636, 2018.

[18] U. Dorjet, K. K. Lee, J. Y. Choi and M. Lee, "The skin cancer classification using deep convolutional neural network", *Multimedia Tools and Applications*, vol. 77, no. 8, pp. 9909–9924, 2018.

[19] Q. Abbas, M. Sadaf, A. Akram, "Prediction of Dermoscopy Patterns for Recognition of both Melanocytic and Non-Melanocytic Skin Lesions", *Computers*, vol. 5, no. 3, pp. 13, 2016

[20] J. B. Janney, S. Roslin, "Classification of melanoma from Dermoscopic data using machine learning techniques", *Multimedia Tools and Applications*, vol. 79, no. 5-6, pp. 3713–3728, 2020.

[21] A. Esteva, et al. "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[22] E. Alamsour and M. A. Jaffar, "Classification of Dermoscopic skin cancer images using color and hybrid texture features", *International Journal of Computer Science and Network Security*, vol. 16, no. 4, pp. 135–139, 2016

[23] W. V. Stoecker, C. S. Chiang, and R. H. Moss, "Texture in skin images: comparison of three methods to determine smoothness", *Computerized Medical Imaging and Graphics*, vol. 16, no. 3, pp. 179-190, 1992.

[24] S.S. Han et al. "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm", *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529-1538, 2018

[25] M.A. Kadampur and S.A. Riyaee. "Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images", *Informatics in Medicine Unlocked*, vol. 18, 2020.

[26] N. Codella et al. "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)", *arXiv preprint arXiv:1902.03368*, 2019.

[27] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", *Scientific Data*, vol. 5, pp. 180161, 2018.

[28] "Keras Applications", [Online]. Available: https://keras.io/applications/. [Accessed on 1st April 2020].