Machine Learning Summer School - Indonesia

ANALYSIS OF ADVERSARIAL ATTACKS ON SKIN CANCER RECOGNITION

Aminul Huq, Mst. Tasnim Pervin
Tsinghua University, Beijing, China

huqa10@mails.tsinghua.edu.cn, pervinmt10@mails.tsinghua.edu.cn

Abstract

Cancer is one of the most common diseases having the highest death rates. Among these cancers, 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 for image classification. 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 to increase the robustness of two popular deep learning models, namely MobileNet and VGG16, against both white-box attacks of PGD and FGSM attacks. The experiments shows that our models are much robust than standard training ones and achieved almost similar results.

Dataset Description

"Humans Against Machines with 10000 Training Images" - HAM10000 comprising of 10015 dermatoscopic images of seven classes [1].

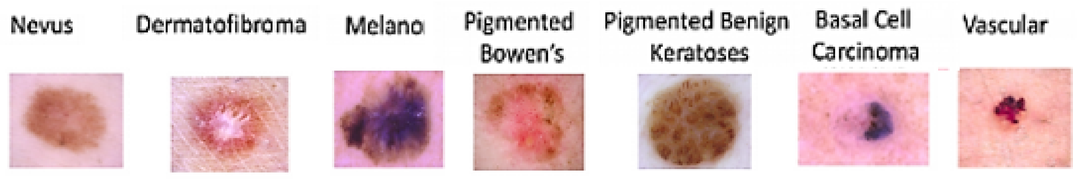

Figure 1: Sample of dataset

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. These models have been used for both standard training, adversarial training with adversarial examples.

Adversarial Machine Learning

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 attacks is to inject some noise to the input to be classified so that the resulting prediction is changed from actual class to another class.

Fast Gradient Sign Method

FGSM, proposed by I. Goodfellow and his fellow researchers, is one of the earliest, simplest, and fast adversarial attack techniques [2].

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

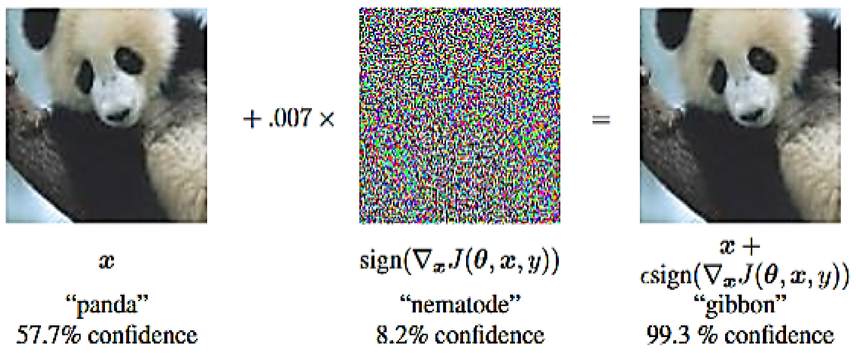

Figure 2: Effect of FGSM Attack

Projected Gradient Descent (PGD)

A. Madry et al. constructed a formulation based on natural saddle point (min-max) to capture the notion of security against adversarial attacks [3]. Following equation states the algorithm:

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

Experimental Implementation

Effect of Attacks(Standard Training):

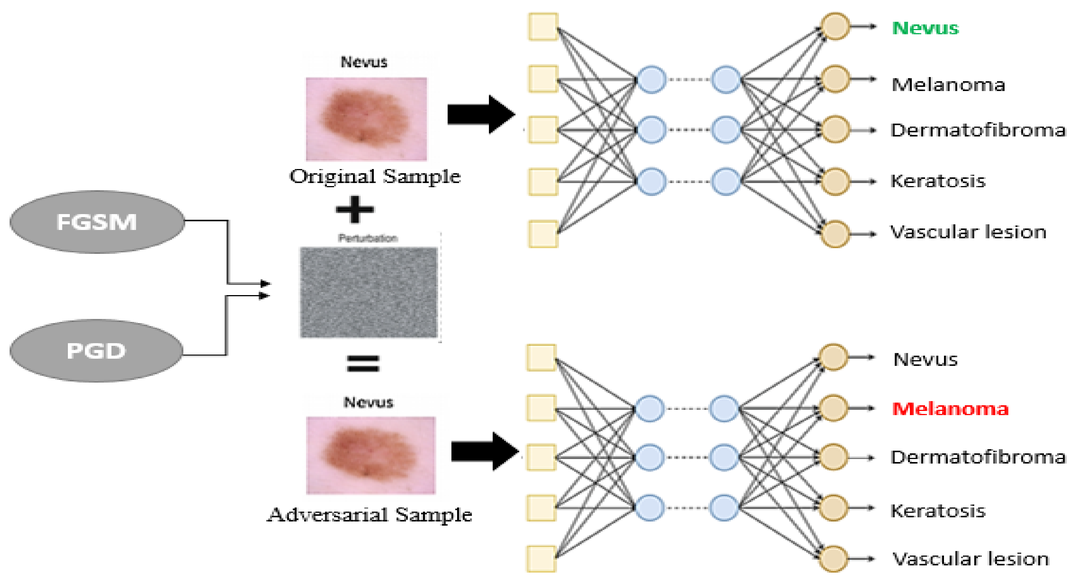

Figure 3: Flowchart of standard training after adversarial attacks

Table 1: FGSM and PGD attack results (% accuracy) on Standard Training

Training Approach	Models	Testing Approach		
		Standard	PGD	FGSM
Standard	MobileNet	77.24	2.89	3.48
	VGG16	71.72	8.68	9.88

Effect of Attacks(Adversarial Training):

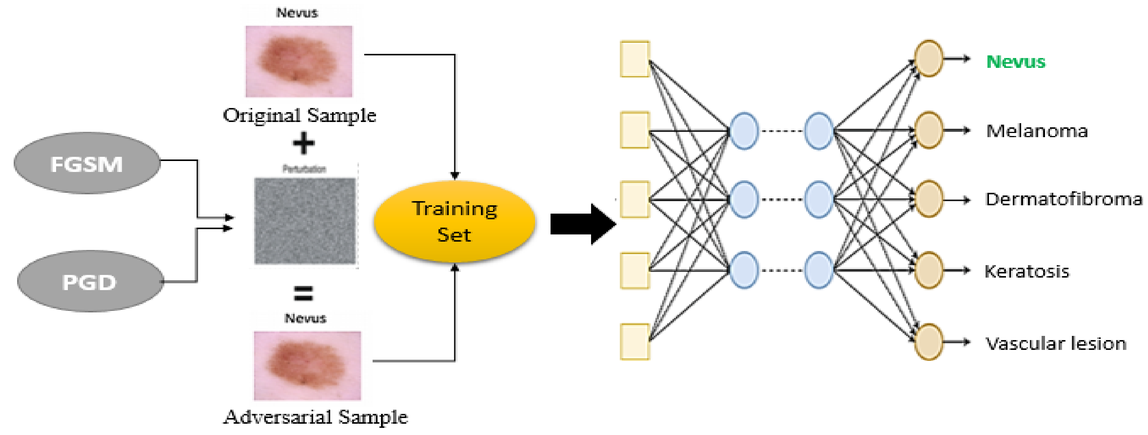

Figure 4: Flowchart of standard training after adversarial attacks

Table 2: FGSM and PGD attack results (% accuracy) on Standard Training and Adversarial Training

Training Approach	Models	Testing Approach		
		Standard	PGD	FGSM
Standard	MobileNet	77.24	2.89	3.48
	VGG16	71.72	8.68	9.88
Adversarial	MobileNet	76.14	**75.94**	**63.17**
	VGG16	70.47	**71.05**	**50.79**

Table 3: Transferability of standardly trained models(% accuracy) by testing with PGD Adversarial Images

Models	MobileNet	VGG16
MobileNet	2.89	6.08
VGG16	12.77	8.68

References

[1] 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.

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

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