Analysis of Adversarial Attacks on Skin Cancer Recognition

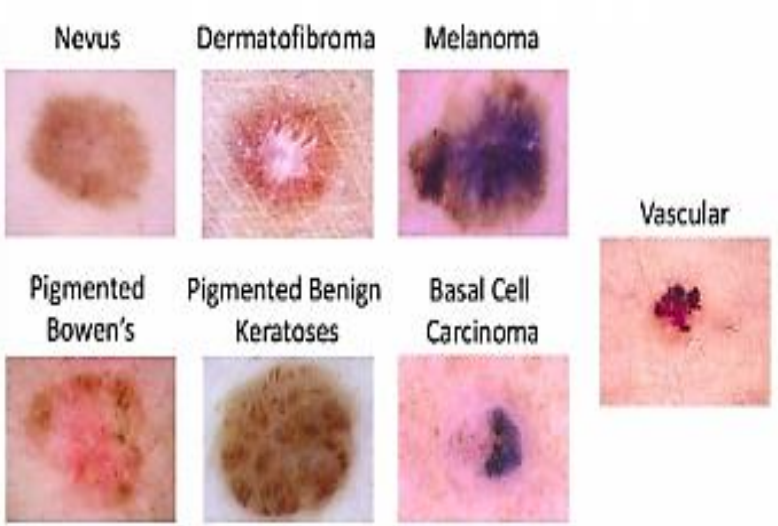
Aminul Huq, Mst. Tasnim Pervin

Dept. of Computer Science & Technology

Tsinghua University, Beijing, China

Introduction

- “Humans Against Machines with 10000 Training Images" HAM10000[1] comprising of 10015 dermatoscopic images
- Seven classes:
    1. Melanoma
    2. Melanocytic nevus
    3. Basal cell carcinoma
    4. Intraepithelial carcinoma
    5. Benign keratosis
    6. Dermatofibroma
    7. Vascular lesion



Classification Models

Models	Top-1 ImageNet Accuracy	Million Parameters
VGG16 [2]	0.715%	138M
MobileNet[3]	0.706%	4.2M

Adversarial Machine Learning

- FAST GRADIENT SIGN METHOD[4]

$$x^{adv} = x + \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{true}))$$

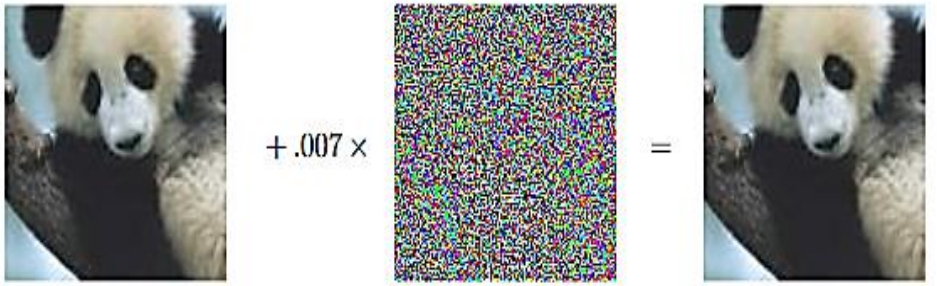
where

x is the input (clean) image,

x^{adv} is the perturbed adversarial image,

J is the classification loss function,

y_{true} is true label for the input x.



Adversarial Machine Learning

- ITERATIVE ADVERSARIAL ATTACKS (PGD) [5]

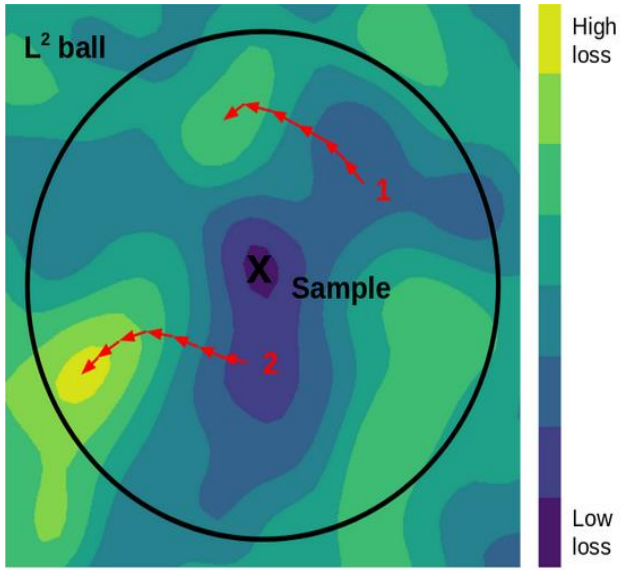
$$min_\theta\, p(\theta), \qquad where\, p(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, [max_{\delta\epsilon S}\, L(\theta, x+\delta, y)]$$

Step 1 Start from a random perturbation in the L^p ball around a sample

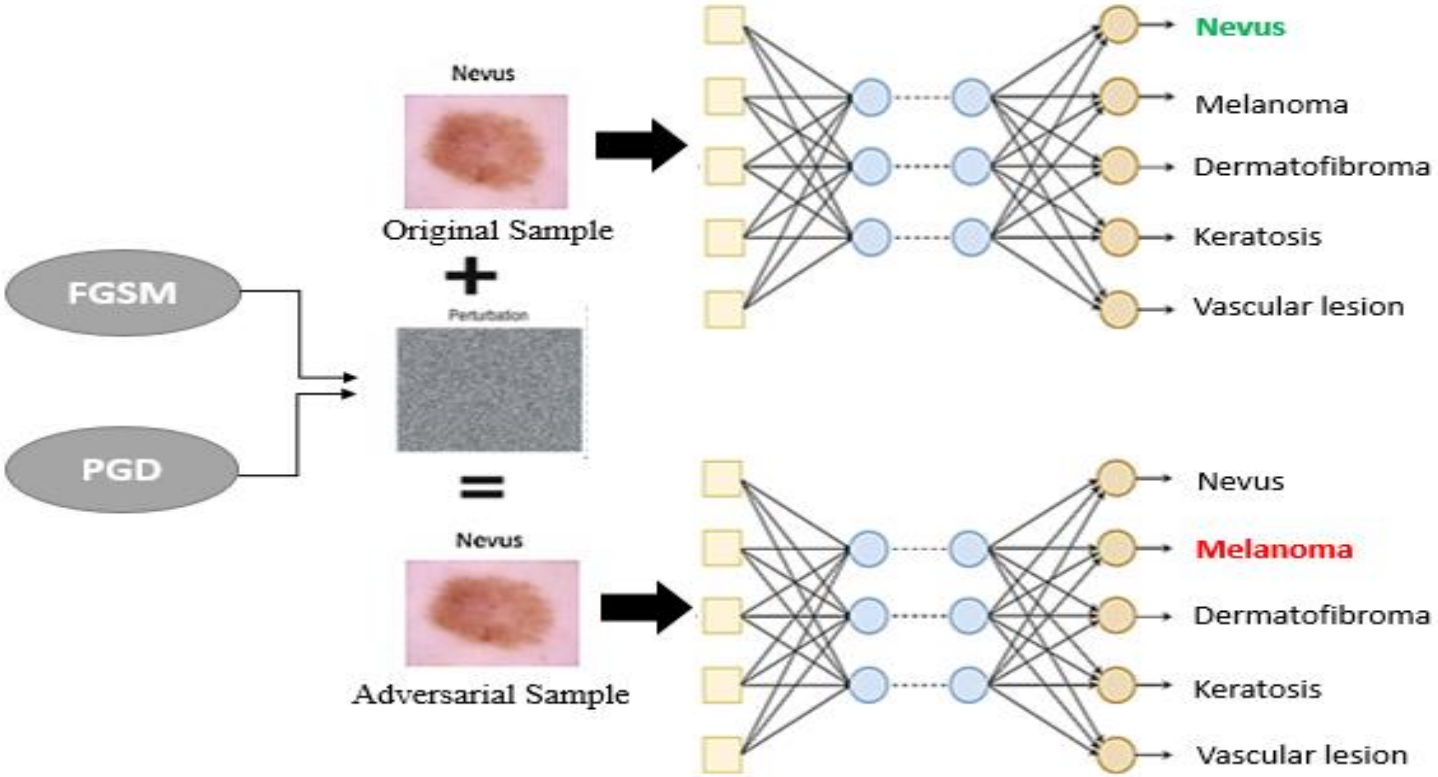
Step 2 Take a gradient step in the direction of greatest loss

Step 3 Project perturbation back into L^p ball if necessary

Step 4 Repeat 2–3 until convergence



Effect of Attacks(Standard Training)

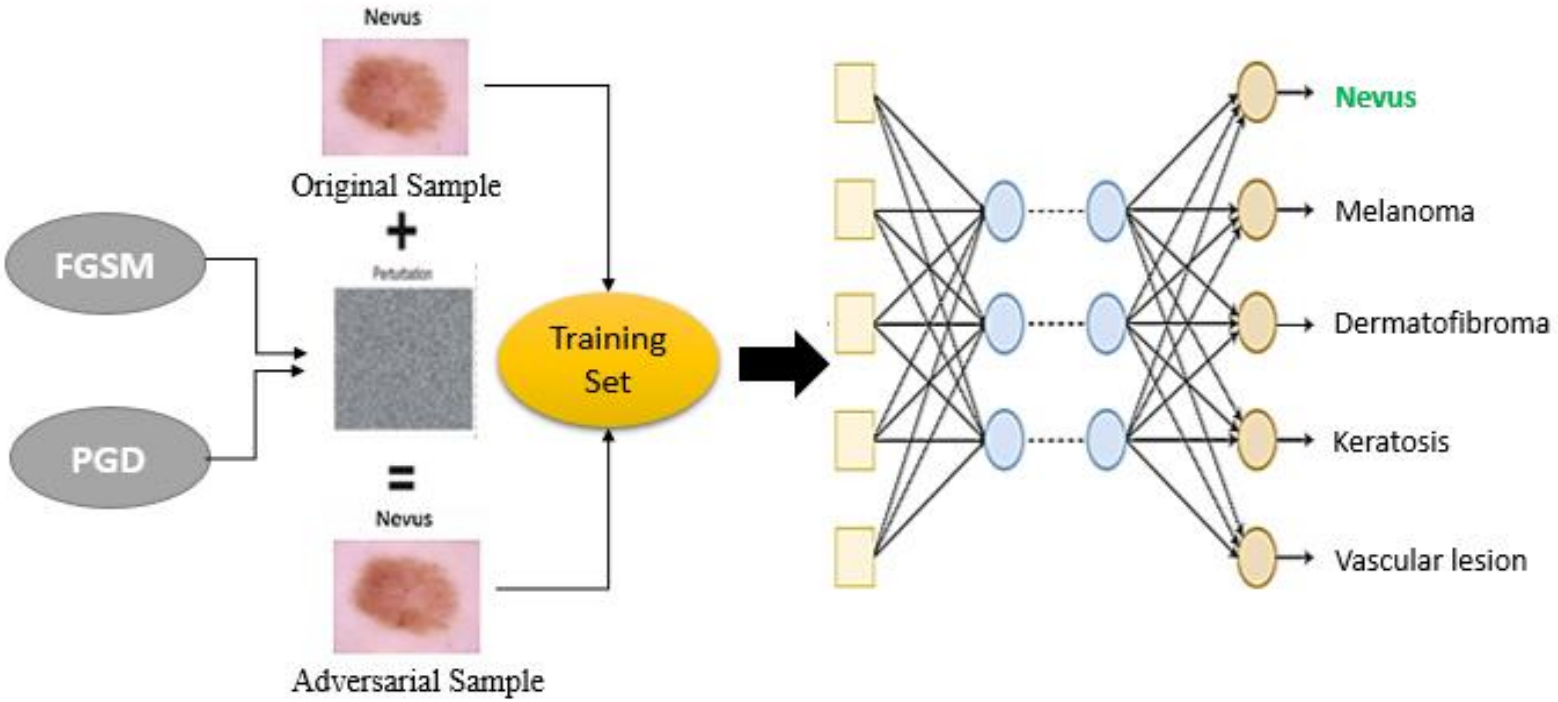


Performance Analysis

Table 1. FGSM & 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)



Performance Analysis

Table 2. FGSM & 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**

Performance Analysis

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] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi:10.1038/sdata.2018.161 (2018).

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

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[4] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[5] Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).

Thank You