IEEE TENSYMP 2020
Technology for Impactful Sustainable Development
5-7 June 2020, Dhaka, Bangladesh
IEEE region 10
IEEE Bangladesh Section
5-7 June, Dhaka, Bangladesh

Robust Deep Neural Network Model for Identification of Malaria Parasites in Cell Images

Aminul Huq & Mst. Tasnim Pervin

Overview

Introduction to Adversarial Machine Learning

Malaria Cell Image Analysis

Classification and Robustness

- Dataset & Models
- Fast Gradient Sign Method
- Defense Techniques
- Results and Analysis

Conclusion

Adversarial Machine Learning

Given a classifier $f(x) : x \in X \rightarrow y \in Y$, which maps the input sample x to the label y.

Normal Example:

- $y = f(x)$

Adversarial Example:

- $f(x) \neq f(x + \delta)$ where $||\delta||_p$ is small

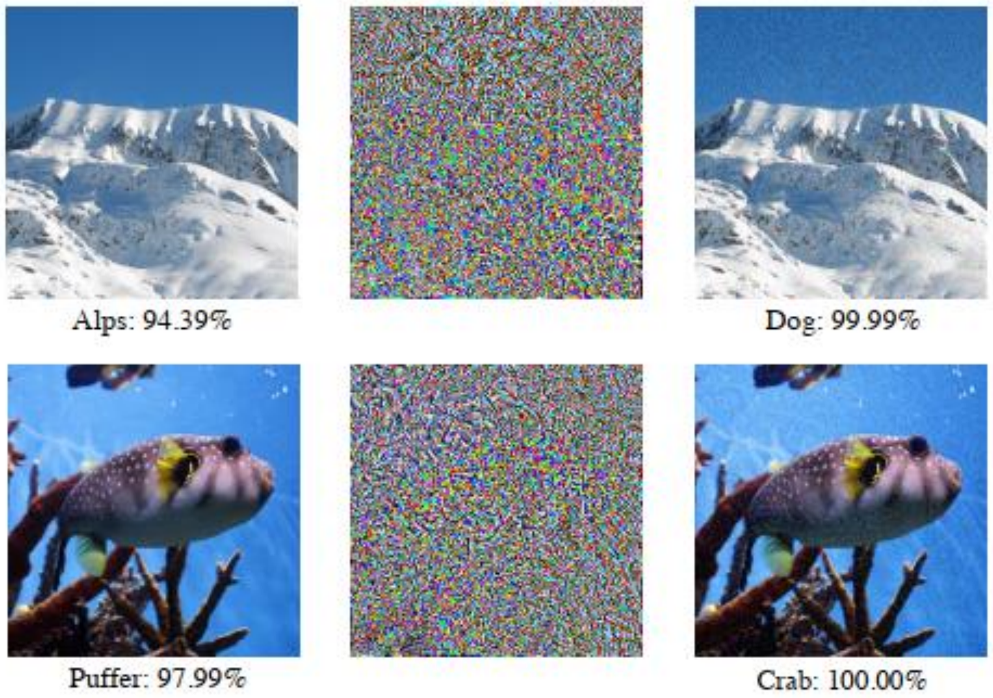


Fig. 1 Adversarial Image Example [1]

[1] Y. Dong et al. "Boosting adversarial attacks with momentum." Proceedings of the *IEEE conference on computer vision and pattern recognition*. 2018.

Classification of Attacks

Knowledge

White-Box
- I/P & O/P
- Architecture
- Parameters

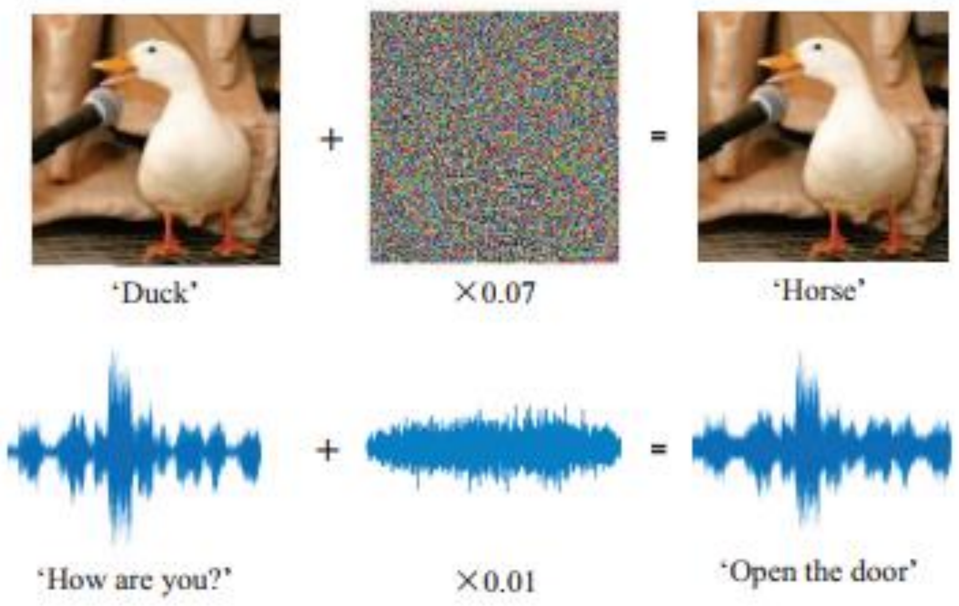
Black-Box
- I/P & O/P
- Using queries tries to get additional info

Goal of Attacker

Non-Targeted
- Lower model accuracy adversary doesn't care about labels
- $f(x + \delta) \neq y$

Targeted
- Alter the label to a adversary given fixed label.
- $f(x + \delta) = y'$

Motivation



spider-man is better than any summer blockbuster we had to endure last *boring* summer , and hopefully , sets the tone for a summer of good stuff . if you're a comic fan , you can't miss it . if you're not , you'll still have a good time .

Fig 2. Applications of Adversarial Machine Learning [1][2].

[1] Y. Gong et al. "Protecting Voice Controlled Systems Using Sound Source Identification Based on Acoustic Cues." Proceedings of the 27th International Conference on Computer Communications and Networks (ICCCN), Hangzhou, China.

[2] B. Liang et al. "Deep Text Can be Fooled" Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).

Malaria Cell Images

- Malaria is a common mosquito-borne disease that is transmitted through humans by mosquito bites.
- In 2019 the World Health Organization (WHO) had published a report saying that there were 228 million instances of this disease around the whole world [1].
- Thick and thin blood smears are needed to check the presence of the plasmodium parasite and count the number of infected and uninfected cells.
- Expertise is needed from the part of the observer to correctly distinguish a normal cell from the infected cell.
- CAD systems are capable of doing this efficiently with the help of DNN.

[1] WHO, 2019, World Malaria Report. Available at https://www.who.int/news-room/feature-stories/detail/world-malariareport-2019 (Accessed on 27th February 2020)

Dataset & Model Description

- Dataset : National Institute of Health & Kaggle [1].
- Total of 27,558 images ,13,779 normal cell images & 13,779 infected cell images.
- Train, Test & Validation set.
- VGG16. Top 5 acc 90.1% on ImageNet dataset.

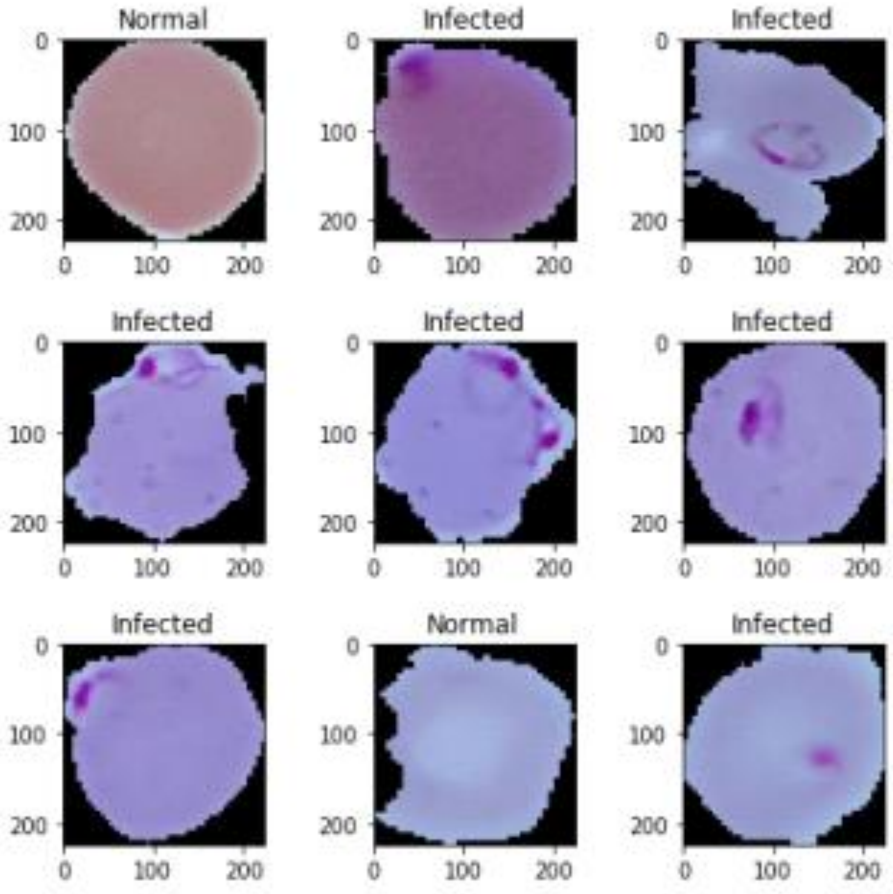


Fig 3. Sample Dataset [1]

[1] S. Rajaraman et al. (2018) “ Pre-trained convolutional neural networks as feature extractors toward improved Malariaparasite detection in thin blood smear images. ” PeerJ6:e4568 https://doi.org/10.7717/peerj.4568

Generating Adversarial Images

- Fast Gradient Sign Method (FGSM)[1]

- Simple and Fast

$$x_{adv} = x + \in * \, sign\left(\nabla_x J(\theta, x, y)\right)$$

x_{adv} = adversarial image, x = input image, $\in$ = epsilon, J= loss function, θ = model parameters & ∇_x = gradient with respect to input.

- Using this method we generated adversarial testing images using original test set.

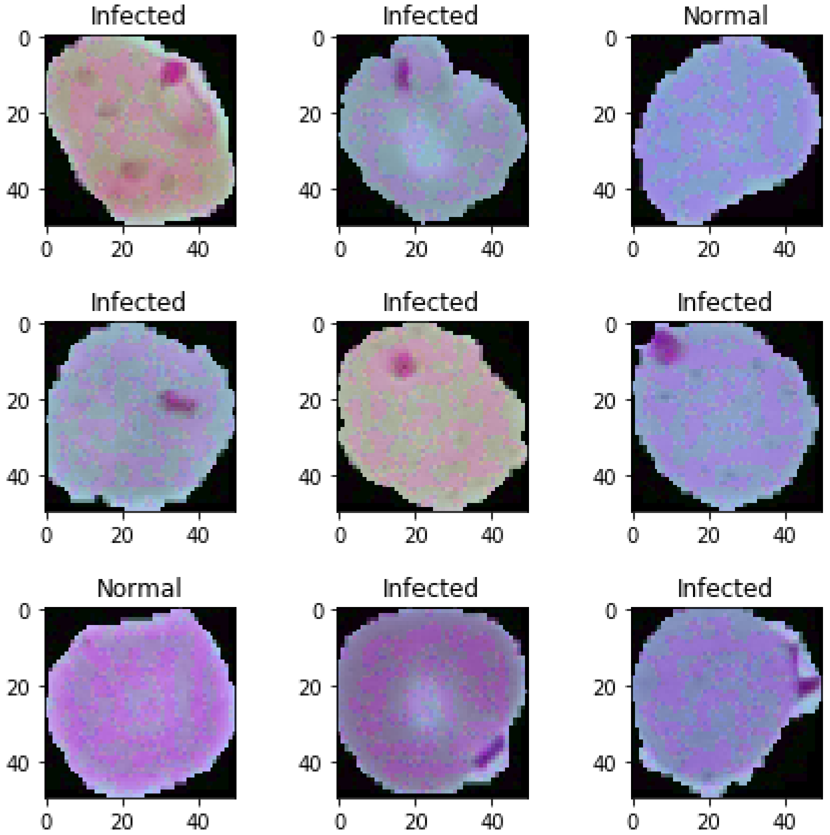


Fig 4. Adversarial Images

[1] I. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014)

Initial Results

	Accuracy	Precision	Recall	F1 Score
Original Test Data	95.96	96.78	95.10	95.59

	Accuracy	Precision	Recall	F1 Score
Adversarial Test Data	29.40	36.70	56.29	44.43

Defense

- Three types of approaches for defense:
  - Image Transformation : Image cropping, JPEG compression.
  - Distillation : Denoising AE, HGD.
  - Training : Adversarial Training.
- Most effective and currently most used is Adversarial Training.
  - Generate adversarial images for training purpose.
  - Train the model with both original image and adversarial image.

Results

Before Adversarial Training

	Accuracy	Precision	Recall	F1 Score
Adversarial Test Data	29.40	36.70	56.29	44.43

After Adversarial Training

	Accuracy	Precision	Recall	F1 Score
Adversarial Test Data	93.38	90.59	96.86	93.62
Original Test Data	95.79	95.73	95.87	95.80

Conclusion

- We have showed here that DNN are susceptible to noises and it can be exploited.
- We implemented VGG16 for classifying Malaria Cell Images and for evaluating the robustness we generated adversarial images using FGSM.
- Using Adversarial Training we were able to improve our robustness of the model.
- Couldn’t compare with other researchers work as most of the works focused on classification accuracy rather than robustness.
- We are continuing our work and analyzing our results with other models & attack techniques.

Thank you for your Time !