ADVERSARIAL MACHINE LEARNING: AN IN-DEPTH EXAMINATION OF ATTACKS AND DEFENSE MECHANISMS

Prateek Goel
201570211
MSc Data Science and Analytics
Supervised by Dr. Luisa Cutillo
In colaboration with The MathWorks

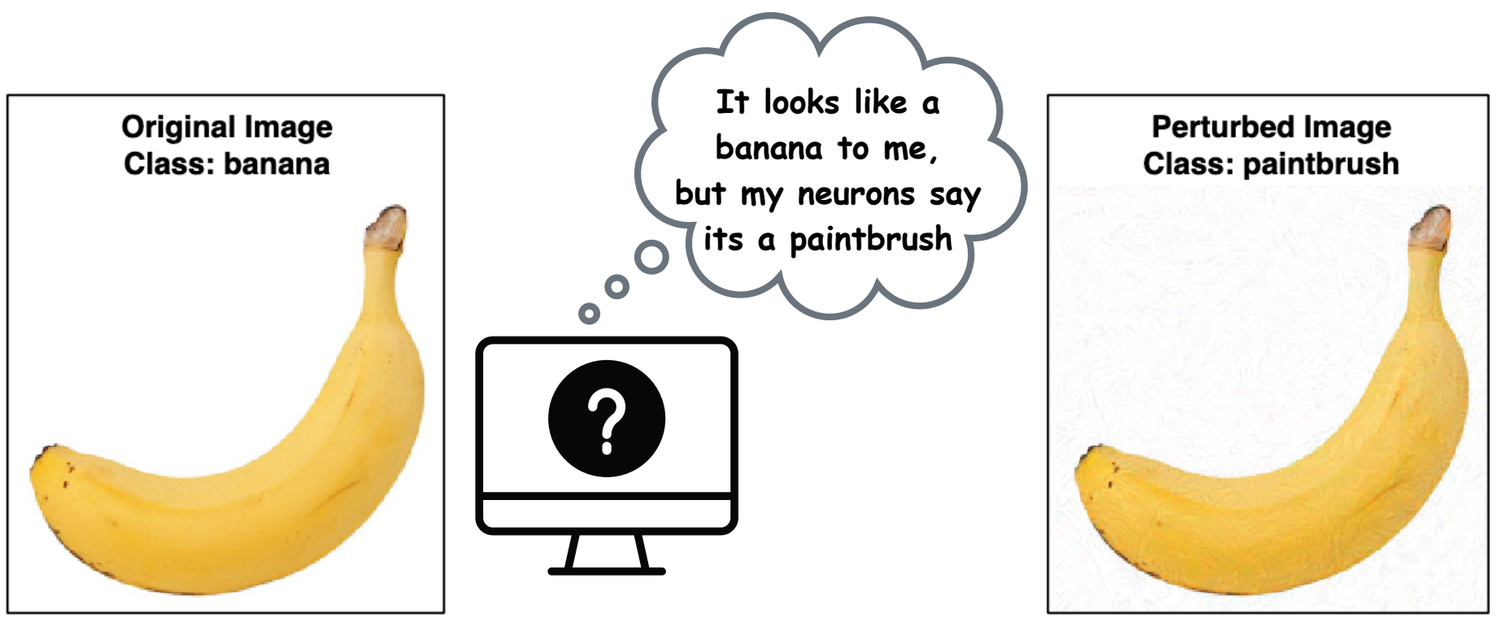
UNIVERSITY OF LEEDS, 2022/23

BACKGROUND

Neural networks, the backbone of modern AI solutions have been playing a pivotal role, especially in critical situations where security is paramount. However, with the reliability comes challenges. They are being mislead/fooled by adversaries in order to produce wrong outputs. Adversarial Machine Learning explores the intriguing intersection of cybersecurity and artificial intelligence. This field aims to study how minor changes in the inputs can easily fool a robust neural network. Using the Grad-CAM technique, we investigate these model responses and implement defense strategies. The objective is to strengthen model robustness, ensuring they can withstand these adversarial attacks and provide reliable, accurate outputs, even when facing deceptive input data

METHODOLOGY

Our methodology in this project leverages the power of Gradient Descent, a popular optimization algorithm, to generate adversarial examples. The inputs are perturbed in the opposite direction using the gradient to get the input misclassified, thus effectively land with a plethora of adversarial examples using minor perturbations. The difference in the actual and perturbed inputs look imperceptible to the human eyes.



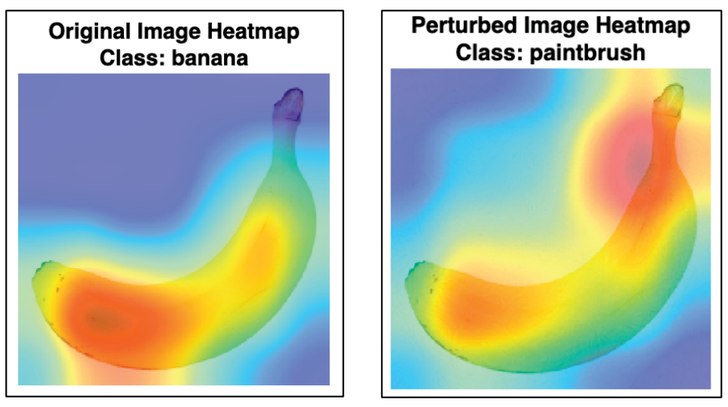
HOW DOES IT WORK?

The gradient is calculated for the current model and is used to create the new/perturbed input from the original input.

$$\tilde{x} = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$$

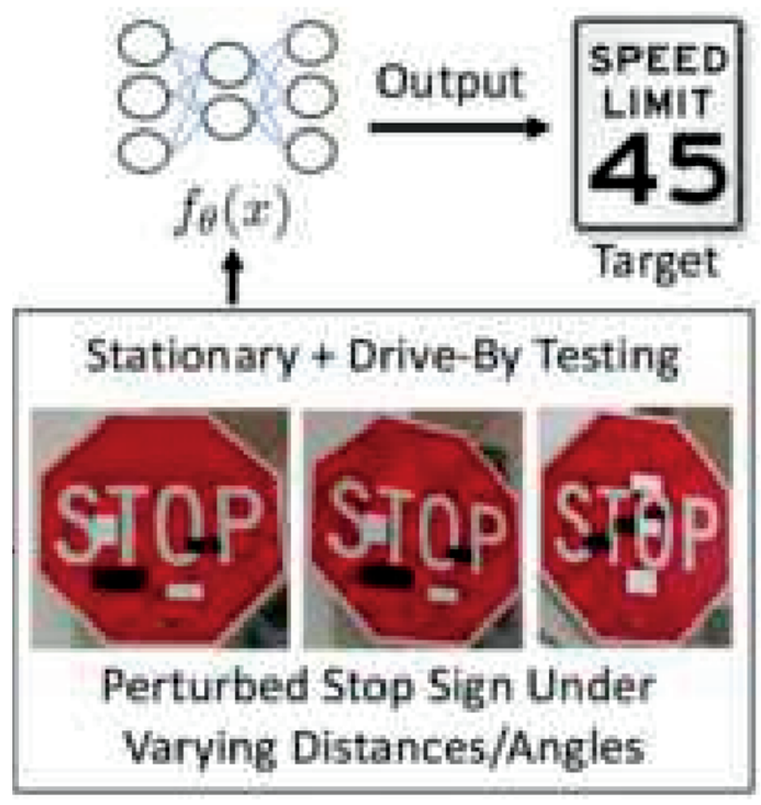
$$\tilde{x} = x - \epsilon \text{sign}(\nabla_x J(\theta, x, t))$$

The adversaries change the focus of the classifier from the part which was used earlier for classification to some other part of the image. This can be seen in the heatmaps generated using Grad-CAM in MATLAB.



The red portion shows the region, which is being focused on by the model to classify the image. The image on the right, which is perturbed shows the change in the area of focus being shifted.

REAL WORLD ATTACK



FUTURE SCOPE

- **Adversarial Training:** The model is trained directly with adversarial examples, enhancing its ability to counteract such attacks.
- G**radient Masking:** A defense strategy that obscures gradients, making it challenging for adversaries to generate successful adversarial examples

REFERENCES

- Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification.
- Goodfellow, I. J., Shlens, J., \& Szegedy, C. (2014). Explaining and harnessing adversarial examples.

ACKNOWLEDGMENTS

I am deeply grateful for the invaluable support provided by Dr. Luisa and MathWorks' Dr. Mike Croucher throughout the entire process for guiding me in the right direction.

MathWorks®

UNIVERSITY OF LEEDS