Mitigating Adversarial Examples in Neural Networks

Moustafa Alzantot (UCLA), **Supriyo Chakraborty** (IBM Research), **Mani Srivastava** (UCLA)

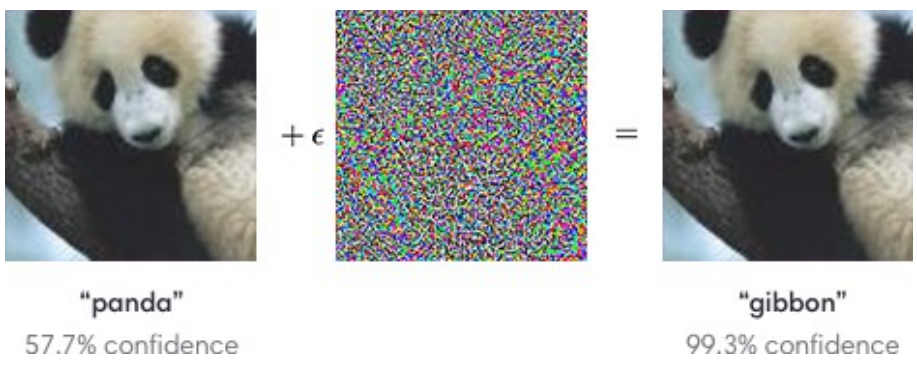
Adversarial Attacks

Machine learning models based on deep neural networks are now achieving super-human accuracy in many domains such as image recognition, speech recognition, machine Translation, and playing games. However, despite their outstanding performance, researchers have also demonstrated that these models can be easily fooled and forced to produce wrong results.

This kind of attack, known as adversarial examples, is performed by adding small amounts of noise that are imperceptible for a human but are sufficient to make a classifier perform mistake on examples that were originally correctly classified.

Researchers have found different efficient methods to generate adversarial noise such as fast gradient sign method (*FGSM*)

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

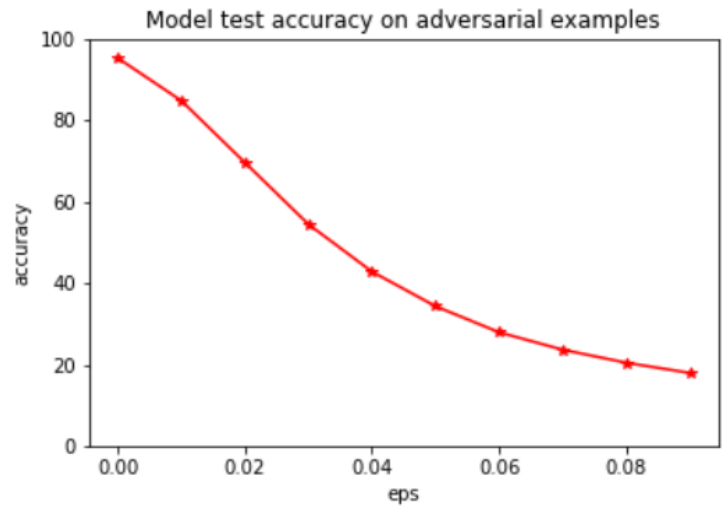


The Jacobian-based saliency map Attack (JSMA) can be used to perform targetted attacks that allow an attacker to force the classifier to mis-classify the input and produces a certain label. JSMA can generate adversarial noise with 97% success while modifying on average 4.02% of the input features.

EXPERIMENTS

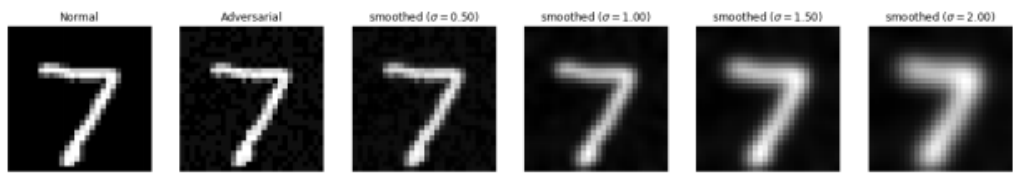
We trained a shallow neural network to classify the MNIST dataset of handwritten digits. MNIST dataset consists of 60, 000 training images and 10, 000 test images.

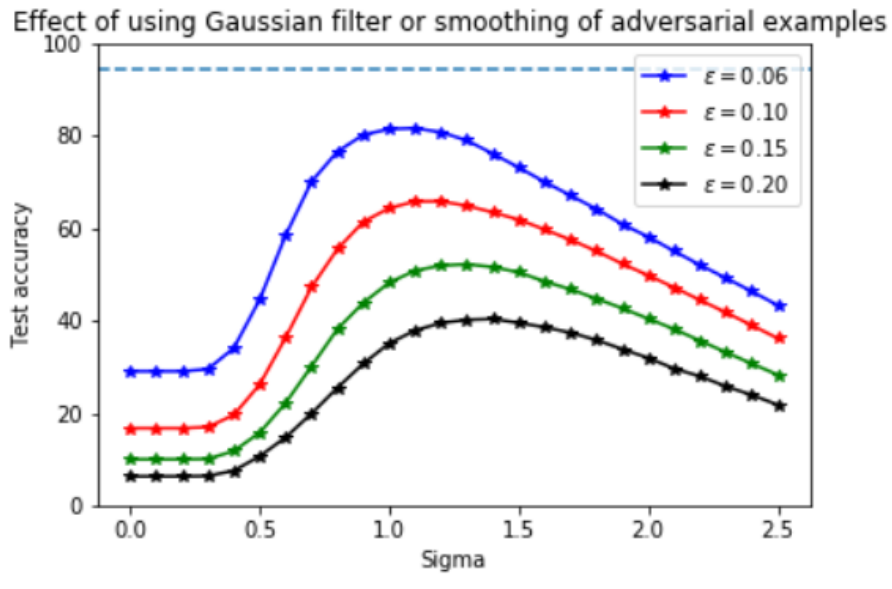
We generate adversarial examples to fool our classification model. The accuracy of the model drops from 95.16% to 17.9% after adding the noise.



Mitigating using Smoothing Kernel

We study the effect of applying a Gaussian kernel to the input image before feeding it into the neural network classification model.





Distributed Analytics and Information Science International Technology Alliance
Annual Fall Meeting, September 2017