# Mitigating Adversarial Examples in Neural Networks

*Project 6, Task 2 (Deep Learning for Multi-Layer Situational Understanding)*

Moustafa Alzantot
University of California, Los Angeles
malzantot@ucla.edu

Supriyo Chakraborty
IBM T. J. Watson Research Center
supriyo@us.ibm.com

Mani Srivastava
University of California, Los Angeles
mbs@ucla.edu

***Abstract*—While deep neural networks models have exhibited astonishing results in a variety of tasks over the past few years, researchers have recently shown how these models can be easily fooled and attacked using *adversarial examples*. Adversarial examples originate from benign examples which are correctly classified by the machine learning model, but an attacker adds a small amount of noise, usually unnoticeable by a human, which are deliberately crafted to make the classifier mistake on classifying the example after adding this adversarial noise. In this paper, we study the characteristics of adversarial noise and present approaches for increasing the robustness of models against adversarial examples attacks.**

## I. Introduction

Machine learning models based on deep neural networks are now achieving super-human accuracy in many domains such as image recognition [8], machine translation [9], and playing games [7]. However, despite the out-sanding performance of deep learning models, 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* [4] are performed by adding small amounts of noise that are unnoticeable for a human but are sufficient to make a classifier perform mistake on examples that were originally correctly classified. Researchers have studied the existence of adversarial examples for different machine learning models [4]. For example, a fast and robust method for generating adversarial noise called the 'Fast Gradient Sign Method' (FGSM) was proposed in [2]. This method, although being very simple and quick, was found powerful enough to make a classifier misclassify 99.9% of examples it was correctly classifying [2]. In FGSM, the adversarial example $\tilde{x}$ is obtained from the original example $x$ by finding:

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

where $\epsilon$ is a small number indicating the maximum amount of noise to be added, $\theta$ is the model parameters, and $J$ is the model cost function.The Jacobian-based saliency map Attack (JSMA) was introduced in [5]. Compared to the FGSM, JSMA can perform an adversarial attack while modifying *only a subset* of input features. Also, JSMA can be used to perform *targetted* attacks that allow an attacker to force the classifier to misclassify 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.

Researchers have also studied the methods of defense against adversarial examples. For example, [2] shows how re-training a model using a subset of adversarial examples makes it more robust against these examples. Other approaches for defense are the 'defensive distillation' [6], and MagNet [3] which uses an auto-encoder to reform the input examples before being feed into the classification model. However, retraining the model with adversarial examples and defensive distillation requires retraining the model which sometimes cannot be feasible to do. Also, defensive distillation was found to be incapable of providing defense against recent attacks [1], and MagNet requires training a large number of auto-encoders to be able to defend against adversarial examples.

In this paper, we study the feasibility of defending adversarial examples by low-cost input pre-processing that does not require retraining the target model or other auxiliary networks (such as the autoencoders used in [3]). We evaluate the effects of applying smoothing low-pass filter and frequency-domain high-pass filter to adversarial examples.

## II. 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. Our model can classify the test dataset with $95.16\%$ accuracy. We generate adversarial examples to fool our classification model. Figure II shows how the accuracy of our model on the test dataset drops from $95.16\%$ to $17.9\%$ as a result of adding the adversarial noise.

### A. Mitigation using Smoothing Kernel

Our first mitigation approach is based on applying a Gaussian kernel to the input image before feeding it into the neural network classification model. The effect of Gaussian kernel is equivalent to a smoothing low-pass filter where the level of smoothing depends on the kernel standard deviation parameter $\sigma$. The leftmost image in Figure 4 shows a sample image from the test dataset. This image of digit 7 was correctly classified by the classification model. However, after adding the adversarial noise it becomes misclassified with output label 3, although it still looks unchanged for a human's eye. Figure 2 shows the effect of different smoothing levels applied to the

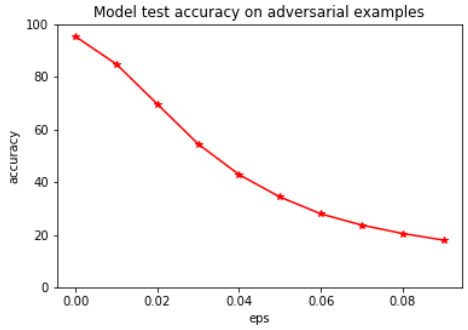


Fig. 1. Accuracy of our classification model evaluated on adversarial examples datasets with different noise $\epsilon$ levels.

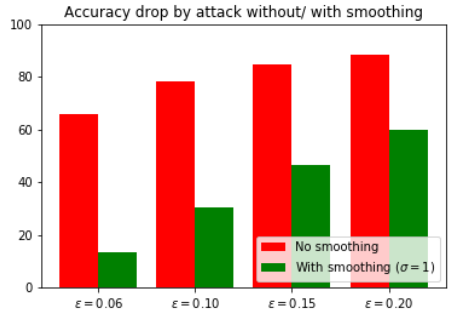


Fig. 4. Percentage of successful attacks by adversarial examples on a model with/without inputs smoothing.

adversarial image. We notice as the $\sigma$ increases, the image becomes more blurry.

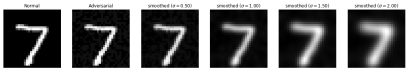

Fig. 2. Effect of smoothing on the adversarial image.

*1) Effect on Adversarial Examples:* When we apply the Gaussian kernels to different datasets of adversarial examples with different noise levels $\epsilon$. We observe that for small levels of adversarial perturbations, the smoothing can compensate *part of* the adversarial noise effect. However, accuracy remains below the test accuracy on normal examples. Also, if we use a large $\sigma$ value for smoothing kernel, the accuracy starts going down again because of the images becoming too blurry which implies information loss

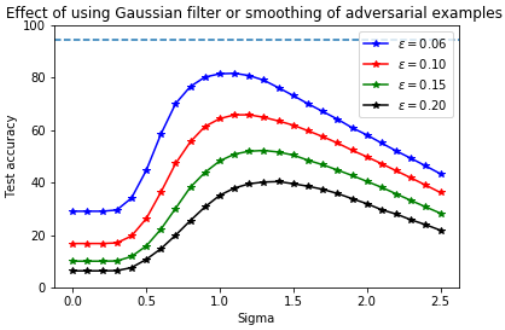


Fig. 3. Effect of the smoothing kernel $\sigma$ value on model accuracy as evaluated against adversarial examples datasets.

Figure 4 shows the percentage of successful attacks by adversarial examples when smoothing is used and when it is not used.

## III. Conclusion

In this study, we analyzed the characteristics of adversarial examples. We studied how Gaussian kernel smoothing increases the model robustness to adversarial noise attacks. Our result shows that a smoothing kernel can slightly increase the model robustness against adversarial noise generated by the FGSM. Our ongoing work is to evaluate the robustness of our results against other attack methods (e.g. JSMA approach) and different classification models and datasets.

## Acknowledgement

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

## References

[1] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

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

[3] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. *arXiv preprint arXiv:1705.09064*, 2017.

[4] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[6] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.

[7] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[9] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.