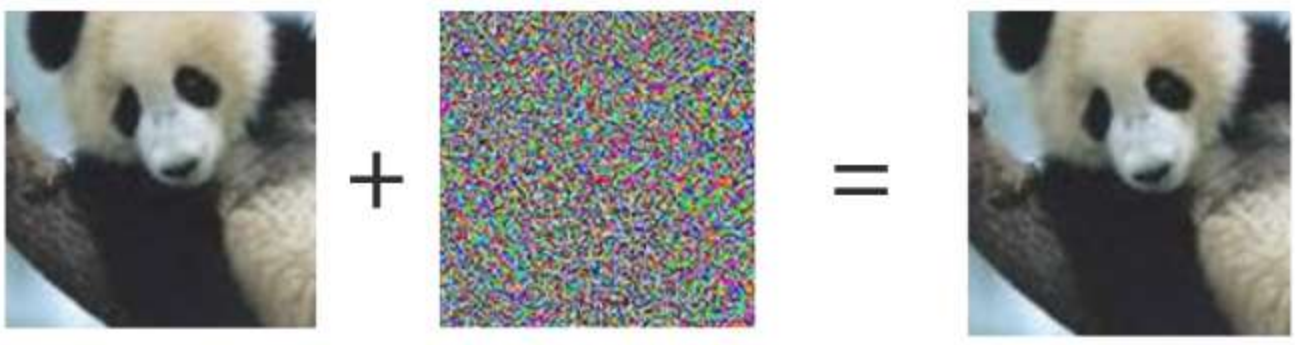
Defense against Adversarial Machine Learning

Jesús Alejandro Noguera Ballén <janoguerab@unale.du.co>
Jorge E. Camargo, PhD <jecamargom@unal.edu.co>

1. The Problem - Attacks to gain miss classification.

In recent years it has been shown that many systems of image recognition are susceptible to ***adversarial attacks***, these are intentional modifications of the images, are imperceptible to the human eye, and lead to miss classification by the system, in other words they are like ***optical illusions for systems***.



"Panda"
57.7% confidence

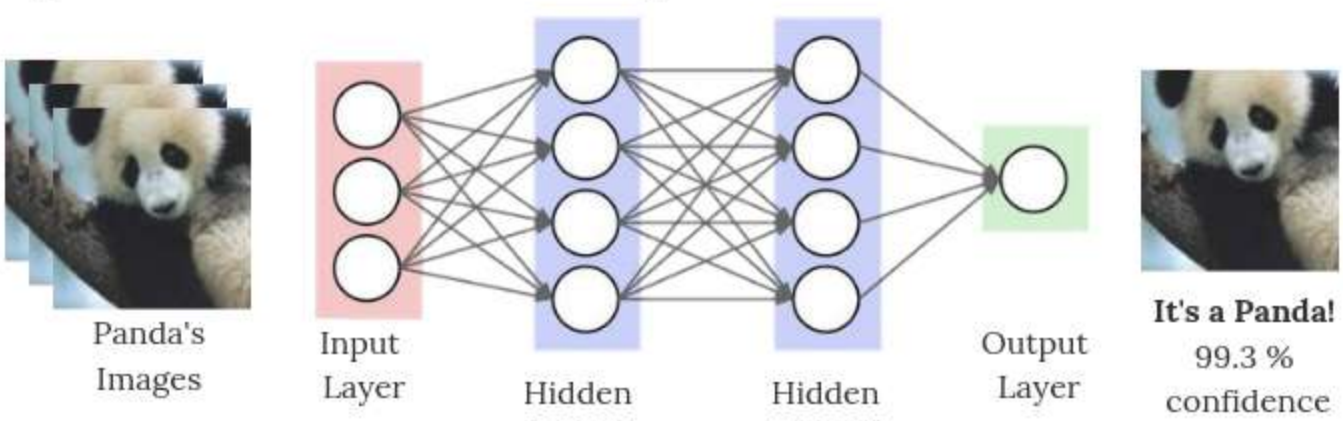
"nematode"
8.2% confidence

"Gibbon"
99.3 % confidence

The resulting image show on the right is classified as **"Gibbon"** by a pre-trained neural network with a 99.3% confidence.

2. How It works?

Train a neural network with images and its associations (well known databases) so the neural network gets a high confidence in the classification of the image.



It is extremely important to give robustness to the learning algorithms so that it can responds effectively against the missclasification attacks.

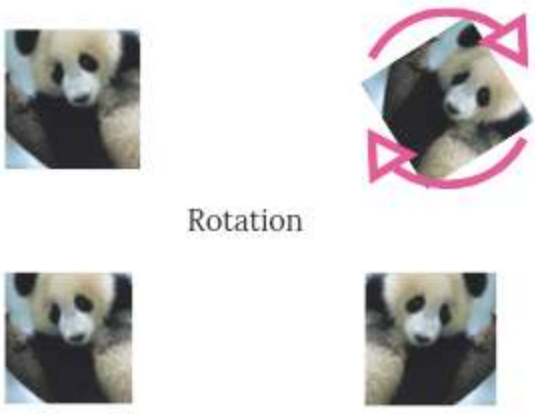
3. Objective:

The main goal in this paper is show a strategy to give robustness to the learning algorithms and can therefore be more effective in the process of classification.

Justification: Adversarial examples have the potential to be dangerous. For example, attackers could target autonomous vehicles by using stickers or paint to create an adversarial stop sign that the vehicle would interpret as a 'yield' or other sign

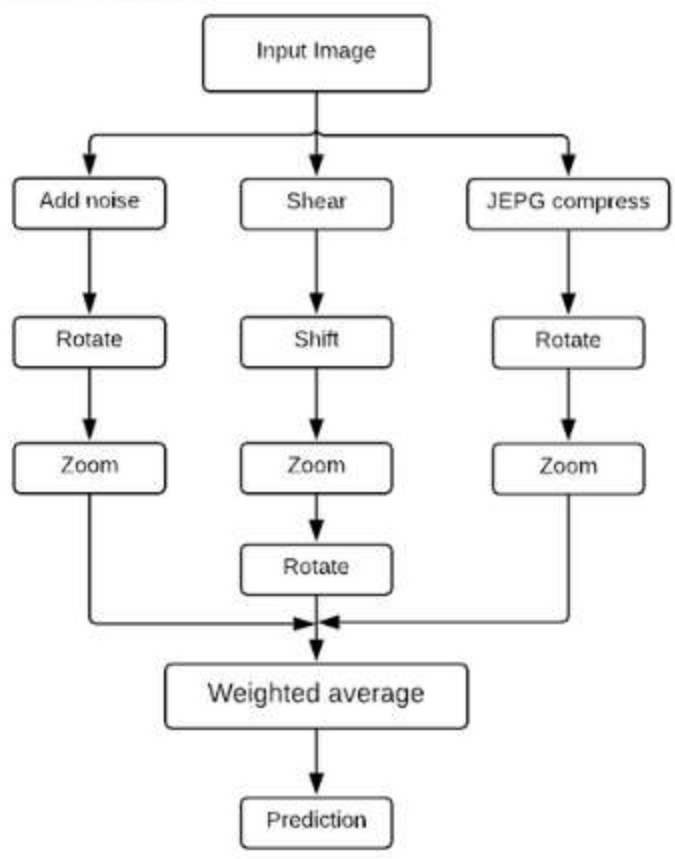
4. Methodology:

4.1) Apply several transformations to the input image to make the attack perturbations less effective



4.2) Classify using networks trained with adversarial examples. This kind of training helps the networks generalize better to images that are not just outside the training set, but do not even occur in "nature"

5. Strategy:



References

- *Alpaydin, E. (2014). Introduction to machine learning*
- *Dasgupta, B., Liu, D., & Siegelmann, H. T. (2007). Neural networks. In Handbook of approximation algorithms and metaheuristics*
- *Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing.*
- *Helmbold, D. P. & Long, P. M. (2016). Dropout versus weight decay for deep networks.*
- *Hu, W. & Tan, Y. (2017). Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN.*
- *Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. (100), 1–14.*
- *Papernot, N., McDaniel, P. D., Wu, X., Jha, S., & Swami, A. (2015). Distillation as a defense to adversarial perturbations against deep neural networks.*
- *Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical Black-Box Attacks against Machine Learning. (November).*
- *Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., . . . Rabinovich, A. (2014). Going deeper with convolutions.*
- *Wu, F., Hu, P., & Kong, D. (2015). Flip-rotate-pooling convolution and split dropout on convolution neural networks for image classification.*