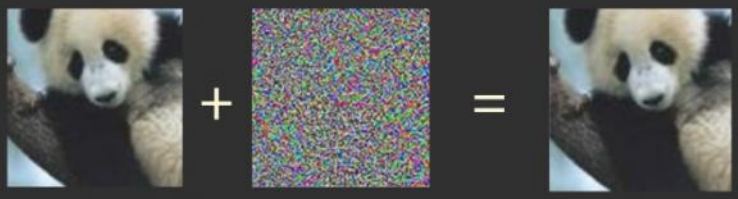
Adversarial Machine Learning in Cybersecurity

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

UNSECURELAB

UNIVERSIDAD NACIONAL DE COLOMBIA

1. The Problem - Attacks to gain miss classification.



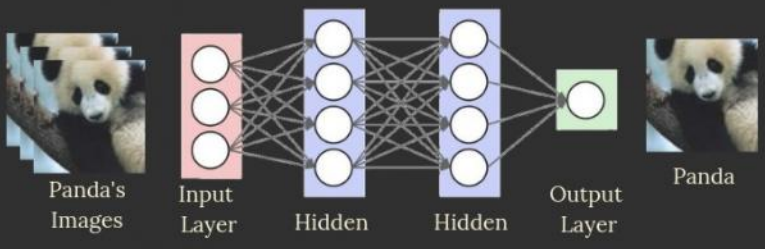
"Panda" 57.7% confidence

"nematode" 8.2% confidence

"Gibbon" 99.3 % confidence

2. How to It Works

Train a neural network with images and its associations (well known databases)



3. The Goal

Reduce missclassification

4. Possible Solutions:

Image Transformations

Rotation

Reflection

References

- Xiao, H. (2017). *Adversarial and Secure Machine Learning*, 153. *Retrieved from https://mediatum.ub.tum.de/1335448.*
- *Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world, http://arxiv.org/abs/1607.02533. Journal Article, 1-14.*
- *Goodfellow, I. (2016). Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples Practical Black-Box Attacks against Deep Learning Systems, (November). http://arxiv.org/abs/1602.02697.*
- *Sinha, A., & Krishnamurthy, B. Understanding Adversarial Space through the lens of Attribution. http://www.research.ibm.com/labs/ireland/nemesis2018/pdf/paper4.pdf*
- *Hayes, J., & Danezis, G. (2017). Learning Universal Adversarial Perturbations with Generative Models. https://doi.org/10.1109/SPW.2018.00015.*
- *NIPS 2017 (2017) - Adversarial Learning – Adversarial attacks and defenses. https://github.com/anlthms/nips-2017.*