Examining Attacks on Neural Networks

University of New Hampshire
College of Engineering and Physical Sciences
[1]Landon Buell
Adviser: [2]Prof. Qiaoyan Yu
[1]Dept. of Physics and Astronomy
[2]Dept. of Electric and Computer Engineering
University of New Hampshire, Durham New Hampshire, USA

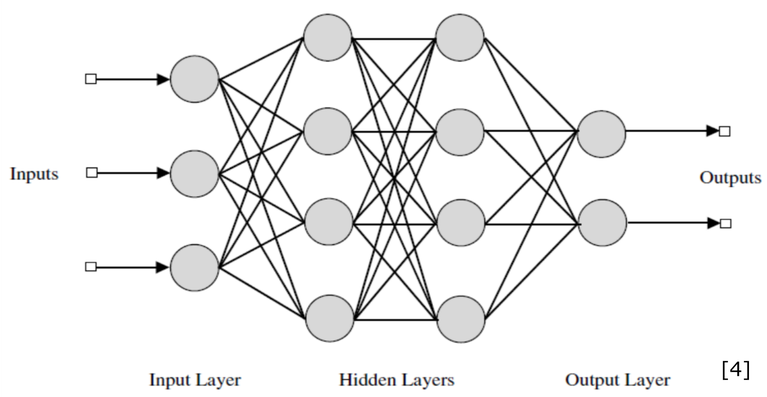
Introduction

- Neural Networks are implemented all over the modern world [1,2]
- What Happens when someone *Attacks* a Neural Network?
- Can we detect an attack based on a Networks behavior?
- We demonstrate initial concepts that Network designers can Explore
- As a practical demonstration, we attack a digit-image classification Program [2,3]

Network Model

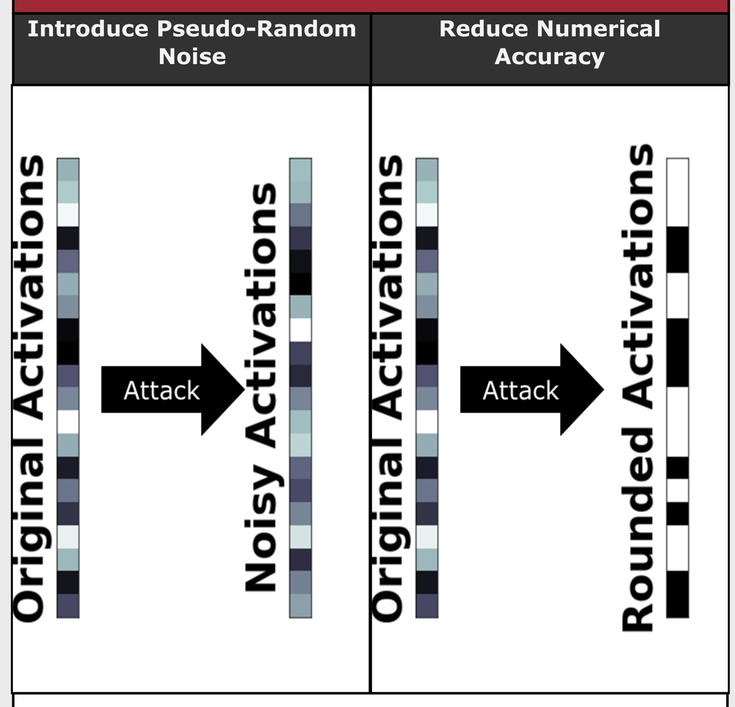
$$\vec{x}^{(l+1)} = f\left[\hat{W}^{(l)}\vec{x}^{(l)} + \vec{b}^{(l)}\right] \quad (1)$$

$$\vec{x}^{(l+1)} = f\left[A\left(\hat{W}^{(l)}\vec{x}^{(l)}\right) + \vec{b}^{(l)}\right] \quad (2)$$



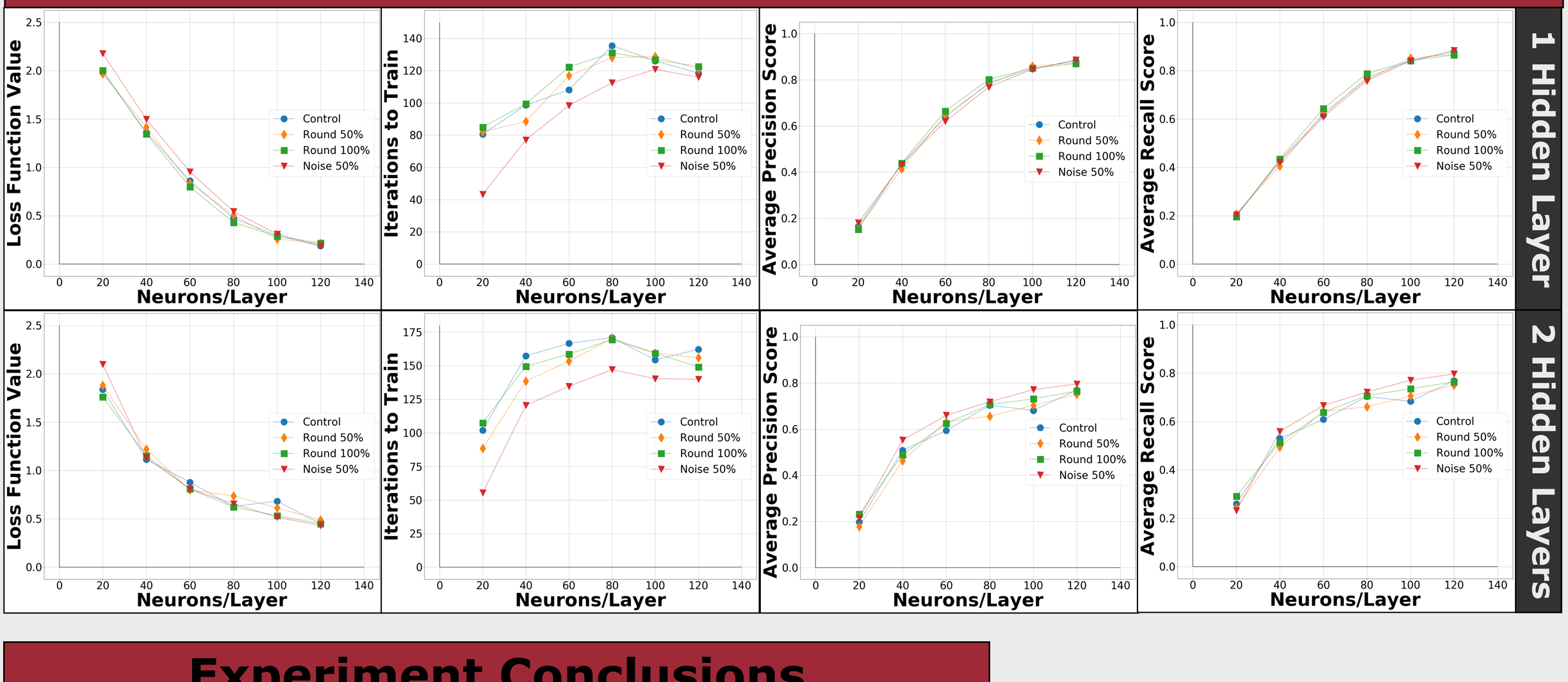
[4]

Attack Models



Two different "*A*" functions are modeled using a color map to indicated how activations are changed by an attack

Experiment Results



Experiment Conclusions

- Attack Functions show changes in Neural Network's *Loss function*, an the number of *iterations* required in training
- No substantial changes are indicated by precision or recall metric scores
- We can expand a future exploration to include conclusions are neuron density and layer numbers providing different results

References

[1] Géron Aurélien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly, 2017.
[2] Goodfellow, Ian, et al. *Deep Learning*. MIT Press, 2017.
[3] Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[4] Choudery, Haroon. "What Are Neural Networks?" Aiforanyone.org, 13 Aug. 2018.

We would like to acknowledge support from the National Science Foundation, award number CNS-1652474.