Examining Attacks on Neural Networks

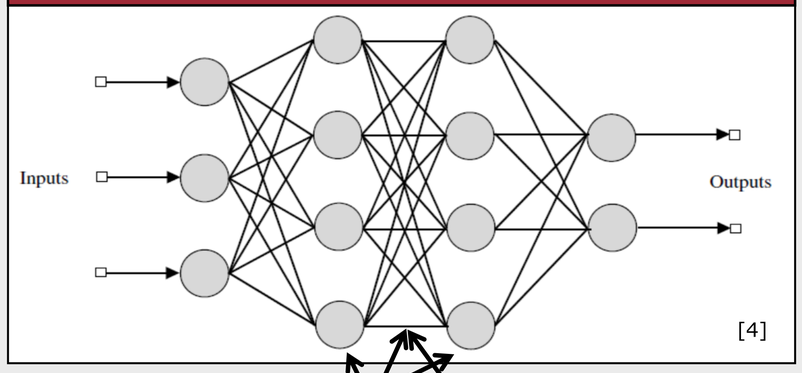
University of New Hampshire
College of Engineering and Physical Sciences

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

Introduction

- Neural Networks are applied in numerous systems worldwide — search algorithms, pattern detection, image recognition, and security. [1,2]
- This widespread use makes them possible targets for Cyber Attacks, which may lead to large consequences including data leakages, and further security vulnerabilities.
- It is imperative that Networks have proactive measures in place that may counter-act an attack if it is detected.
- Using a Image-Classification Neural Network [1,3], we explore how an attack can show changes within the model, over varying layer depth and neuron density.

Network Model



Standard Feed Forward Model

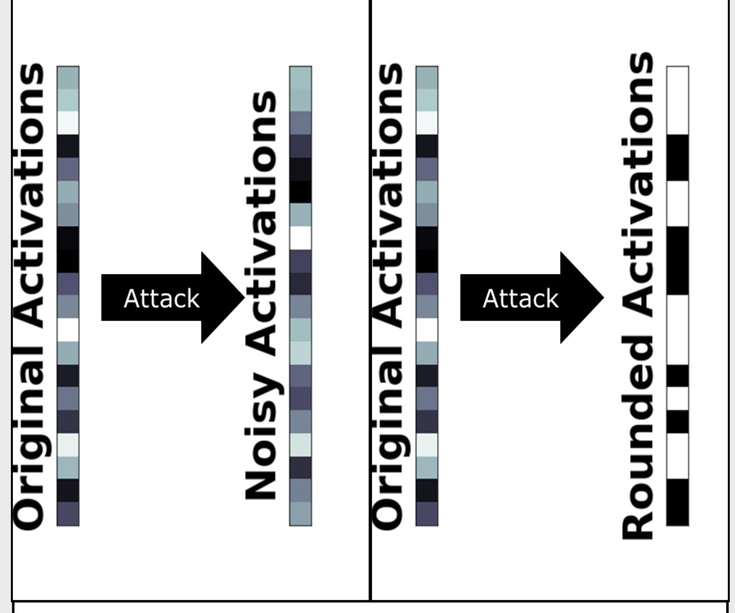
$$\vec{x}^{(l+1)} = f\left[\hat{W}^{(l)}\vec{x}^{(l)} + \vec{b}^{(l)}\right] \quad (1)$$

Attacked Feed Forward Model

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

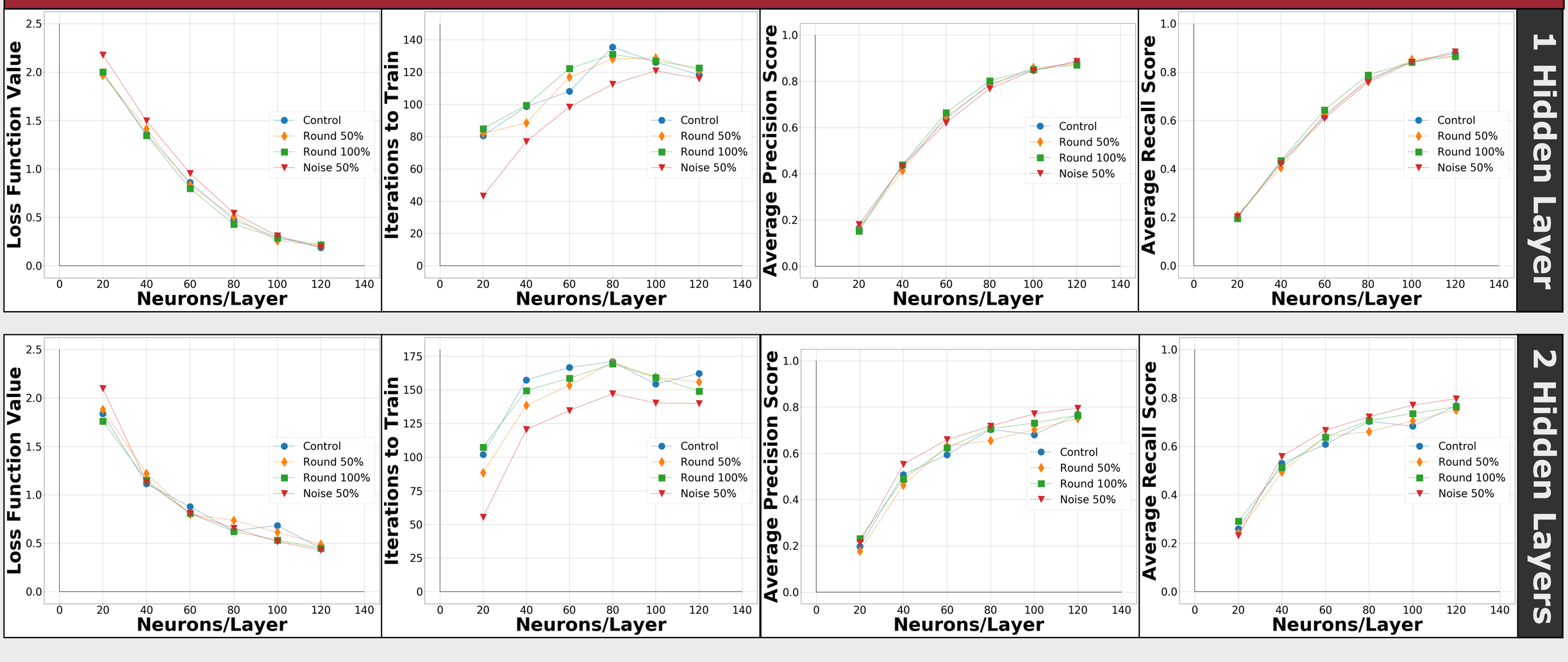
Attack Functions

Introduce Pseudo-Random Noise	Reduce Numerical Accuracy



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

Experimental Results



Conclusions

- For the studies on layer depths and neuron densities, attacks that target numerical accuracy show minor deviations from baseline models. These attacks would be consider *stealthy* as they are hard to detect with the given metrics.
- Both attacks that introduce noise, and those that reduce numerical precision show fewer iterations before either converging or arriving at a *stopping criteria* [2,3].
- The two network depths shown indicates that for both attacks types, precision and recall scores are greater affected by networks with more layers and higher neuron densities.
- We can expand future explorations into studying how attack functions change precision and recall score metrics given network depth and neuron densities.

Acknowledgements

- This work was partially supported by the National Science Foundation, award number CNS-1652474.
- I would like to thank Dr. Kevin Short in the UNH mathematics department for recommending me to this research position and providing additional consultation on this topic

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.