Generating and defending against adversarial examples in vision-optimized neural architectures

Daniel Donoghuel
Email: ddonogh1@jhu.edu

Nicholas Lines
Email: nicholasalines@gmail.com

Arnaldo Pereira
Email: aepereira@gmail.com

***Abstract*—As automated decision-making becomes more popular and more dependent upon artificial intelligence, securing sensitive models from adversarial behavior has become essential. Neural networks are particularly vulnerable to so-called adversarial examples [1], and various attacks and defences have been explored in the literature.**

Our intention in this paper is to demonstrate and confirm the results of such attacks at an informative but modest scale. We apply two common attacks to both the wide ResNet and GoogLeNet neural models, and test two defences, in a reproducible computational environment. We show that significant improvements in network robustness are available with minimal defence measures.

The authors are listed alphabetically, and all made equal contributions. This work is performed in association with the Johns Hopkins Engineering for Professionals Program, as a project for EN.625.638.8VL2.FA20 Neural Networks.

Contents

I. Executive Summary

II. Computational Work

A. Creating Adversarial Examples

B. Defences

III. Analysis

IV. Conclusions

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, *Intriguing properties of neural networks*, 2014.