Landon Buell

Qaioyan Yu

URC 2020 Abstract

18 April 2020

Examining Attacks on Neural Networks

Recent advancements in high performance computing, large information densities, and the need for complex pattern recognition, have found learning models such as neural networks near the forefront of advanced problem-solving. This widespread use of networks in several fields makes them likely targets for potential cyber-attacks or other external programs that seek to manipulate the performance of such a model. Most standard machine learning algorithms do not yet possess a reliable set of measures that proactively monitor system behaviors and properties to counter these possible vulnerabilities. To explore how Neural Networks respond to cyber-attacks, we train and test a series of Multilayer Perceptron Image-Classifiers subjected to different forms of attack functions over varying network depths and neuron densities. Each attacked model is used to examine qualities of the network such as average training iterations, training loss function and average classification performance metrics as compared to an unperturbed baseline. This research will allow for the identification of behavioral differences in classification neural networks based on the complexity of the model, which can then be implemented as a series of counter measures to secure a network against external attacks.