Global Robustness of Several Classifiers

Neta Katz
Advisor: Dana Drachsler Cohen

A thesis proposal presented for the degree of
MSc in Electrical and Computer Engineering

Technion - Israel Institute of Technology

1 Introduction

Deep neural networks have been proven to be effective as classifiers, while often being vulnerable to adversarial attacks. An adversarial attack is a malicious input designed to deceive the neural network into a wrong prediction or output. Therefore, there is a growing need for evaluation methods and algorithms to determine whether a network is robust against these attacks. Existing verification techniques such as Reluplex, ReluVal, and DeepPoly provide formal guarantees for a single network. While these methods contribute significantly to network robustness evaluation, they overlook an essential aspect: the interplay between classifiers. This gap in evaluation methods motivates our research.

Robustness Part of the challenge is defining what is robustness in the notion of multiple neural networks. For a single classifier (neural network) a popular property for understanding its robustness is local robustness. At a high level, we can say that by given a classifier, an input, and a perturbation limit, the classifier is locally robust if perturbing the input up to the perturbation limit does not change the network's classification. The set of perturbed inputs is called the input's neighborhood. However, local robustness is limited to reasoning about a single neighborhood at a time. Therefore, this approach can be problematic when evaluating a large number of inputs and neighborhoods. Specifically, when we want to examine all possible inputs and their neighborhoods, it becomes impractical. As a result, there is a growing need to reason about a network's robustness over all possible inputs, known as global robustness. This means that no matter where you are in the input space, the network should exhibit stable behavior and avoid drastic changes due to minor input perturbations. Global robustness is more challenging to evaluate, compared to local robustness, due to the complexity of the property.

Robustness of Several Classifiers Today many classifiers aim to classify correctly the same inputs. Although there are many studies regarding how to evaluate the robustness property of each of those classifiers individually, there is a lack of research on the relationship between the robustness of different classifiers. We can identify two main challenges in this research: (1) choosing the robustness property, and (2) constructing the evaluation algorithm that would capture the relationships between classifiers while maintaining low running time and scalability at a certain level. In this thesis, we would like to construct an algorithm that would cover all possible input regions. Therefore, the main focus is on the property of global robustness in the notion of several classifiers.

Key Idea To scale the analysis, our key idea is to rely on existing methods that encode global robustness properties for classification networks, such as [INSERT MIP LINK], and extend their usage for several networks instead of a single one. We propose an algorithm that finds dependencies and constraints between layers and neurons in the classification networks, then encodes the networks and finds the lower and upper bounds of the objective property using optimization algorithms, such as Gurobi.

Preliminary Results In our preliminary research, we implemented a basic version of our approach. We evaluated it on classifiers that were trained on the MNIST database, whose architectures are fully connected or convolutional. More specifically, we evaluated our basic version on two classifiers that are almost identical in terms of architectures and weights, where the only difference between them is a small addition to one of the entries of the vector of the last layer of the network. Results show that by using our proposed constraints, we managed to reduce the running time of the evaluation significantly.

Future Goals As part of the thesis, we plan to extend this algorithm to networks that have more differences than a single neuron's weight. Moreover, our current implementation is for two classifiers alone. We would like to generalize it to N¿=2 classification networks.

References