Are Safety Guarantees in Neural Networks Safe?

Merkouris Papamichail[1,2*], **Konstantinos Varsos**[1#], **Giorgos Flouris**[1] and **Joao Marques Silva**[3,4]

[1] Foundation for Research and Technology Hellas

[2]Computer Science Department, University of Crete

[3]Catalan Institution for Research and Advanced Studies

[4]University of Lleida

Presenting author:Konstantinos Varsos, email:varsosk@ics.forth.gr

* Corresponding author: Merkouris Papamichail, email: mercoyris@ics.forth.gr

ABSTRACT

A primary challenge in AI safety is the existence of adversarial examples---slightly distorted inputs that cause a neural network (NN) to misclassify [2]. To mitigate this problem, recent research focuses on the computation of *robustness certifications*, which, for a given input, determine the largest distortion the input may receive without breaking the network's prediction. Robustness certifications can be interpreted as an axis-aligned hyper-rectangle (multi-dimensional intervals). Most existing approaches focus on maximizing the certification's *volume*, but recent intractability results prohibit the computation of volume-optimal certifications in reasonable time. In [1], we introduce the *apothem* measure---the minimum slack between the input and one of the interval's faces, and show how to compute *apothem-optimal* certifications in a *linear* number of calls to a NN verifier (*oracle*) w.r.t. the input domain's diameter. Moreover, we prove that we *cannot* have a volume-optimal, oracle-based algorithm, even if we *discard* the oracle costs. Also, we introduce *dual certifications*---an interval including all instances of a class---thus providing *apothem-minimum* upper bounds to a robustness certification. Our method offers a novel *formal robustness certification* framework for NN classifiers based on interval approximations, called *ParallelepipedoNN* [3].

REFERENCES

[1] M. Papamichail, K. Varsos, G. Flouris, and J. Marques-Silva, Are Safety Guarantees in Neural Networks Safe? How to Compute Trustworthy Robustness Certifications,, ICFEM'26 [TBP].

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and Harnessing Adversarial Examples, ICLR'15.

[3] ParallelepipedoNN, https://github.com/merkouris148/parallelepipedonn