Are Safety Guarantees in Neural Networks Safe?

[1,2]M. Papamichail, [1,2]K. Varsos, [1]G. Flouris, & [3,4]J. Marques-Silva

[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

{mercoyris,varsosk,fgeo}@ics.forth.gr, {jpms}@icrea.cat

Introduction

Despite their remarkable success, Neural Networks (NNs) can exhibit unpredictable behavior: even small, imperceptible perturbations of an input may cause a network to change its prediction. Such vulnerabilities raise serious safety and reliability concerns, particularly in safety-critical applications.

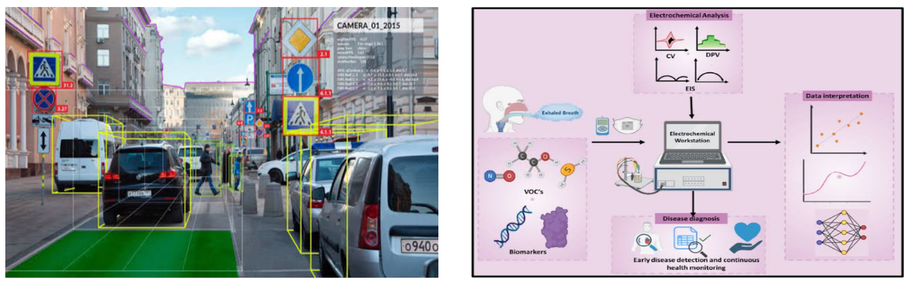


Figure 1: NNs in autonomous driving (left), and medical diagnosis (right).

E.g., the image on the left in Figure 2 is correctly classified as a panda. Adding a minimum amount of noise results in the image on the right. Despite the latter image being (almost) indistinguishable from the first to a human, the network misclassifies it as a *"gibbon"*.

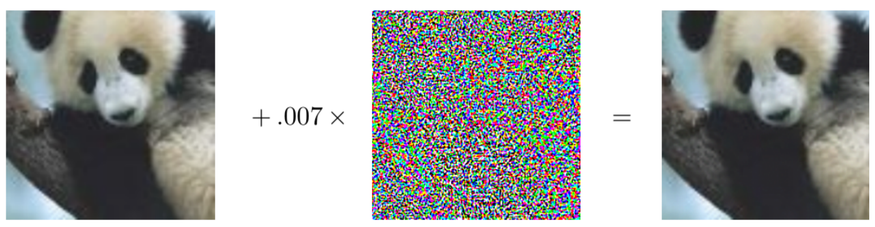


Figure 2: A panda and a gibbon [3].

This discrepancy seems to be inherent to the traditional training procedure and is independent of a model's accuracy, motivating the following,

> The challenge is therefore not only to build accurate AI systems, but to provide guarantees about when their decisions can be trusted.

From Predictions to Guarantees

To address the trustworthyness and safety issues we have to move from predictions to guarantees,

How much can an input change before the Neural Network changes its decision?

Robustness certification addresses this challenge by computing regions around an input for which the network's prediction is guaranteed to remain unchanged, see Figure 3.

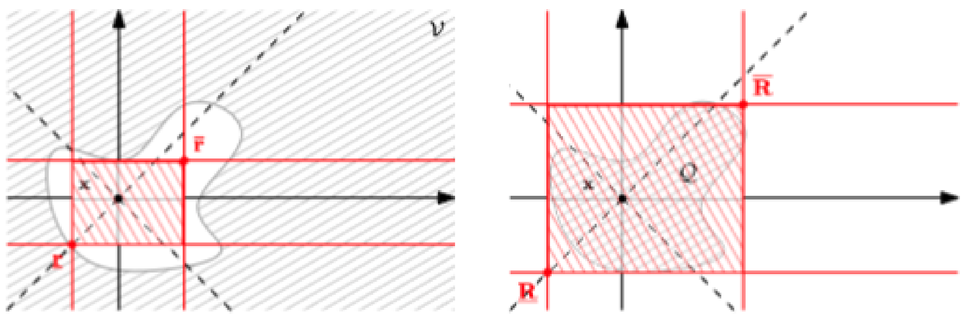


Figure 3: Interval based certifications [1].

Existing methods often underestimate robustness due to mathematical relaxations, making it difficult to distinguish genuinely unsafe networks from conservative certifications.

Compute robustness certifications with **provable optimality guarantees**, interacting with a NN verifier to refine the certified region.

Exploit the geometry and algebraic structure of multi-dimensional intervals to answer,

- the **size of the certification** region,
- if there exist a **better certification**.

Optimal local robustness – Apothem

Apothems: Finding the Nearest Safety Boundary. We introduce the **apothem**, measuring the smallest slack between an input and the boundary of its robustness certificate.

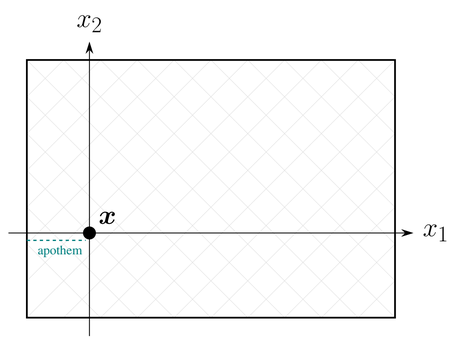


Figure 4: Apothem of x, [2].

Maximizing this quantity provides a certification that is optimal with respect to the nearest direction in which the prediction can fail.

If the optimal apothem does not satisfy a safety requirement, then the network genuinely fails that requirement.

Dual Certifications: Knowing the Limits

A robustness certificate provides a guaranteed region that is **certainly safe**.

A dual certification provides the complementary perspective: an interval containing all inputs associated with a given class.

This yields **how good any robustness certification can possibly be** under any increasing quality measure.

Together, robustness and dual certifications answer:

How robust do we know the network is and how robust could be.

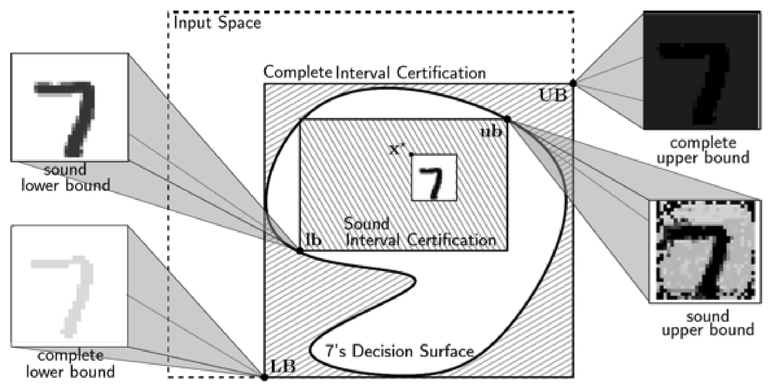


Figure 5: Interval based certifications for a MNIST image of "7".

Is volume optimization enough?

Intuitively, one might seek the robustness certificate with maximum volume.

- We show that finding a volume-optimal certification is computationally intractable [1, 2].

This highlights the importance of **identifying alternative robustness measures that admit both meaningful interpretations and efficient optimal algorithms.**

`ParallelepipedoNN` system

Our method offers a novel *formal robustness certification* framework for NN classifiers based on interval approximations, called **`ParallelepipedoNN`**.

- It uses verification oracles to compute local robustness certificates [1,2].
- It relies on efficient mixed integer linear programming solvers [1,2].
- Avoids convex relaxation methods [1,2].
- Provides exact results without excessive computational cost.

For interior points of the decision surface, our system is the first that *provably guarantees voluminous certifications*.

Beyond Neural Network Robustness

At its core, our approach studies how to construct optimal regions in a high-dimensional space while avoiding undesirable states.

The interval-based framework can potentially support applications wherever we need to answer:

How far can a system move before violating a safety property?

Examples include safe operating regions in control systems, parameter uncertainty analysis, verification of autonomous agents, and robust decision-making.

Acknowledgements

This work is partially funded by COFLED project.

ParallelepipedoNN

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 [TBA].
[2] M. Papamichail, K. Varsos, G. Flouris, and J. Marques-Silva, *The Cost of Relaxation: Evaluating the Error in Convex Neural Network Verification*, ArXiv.
[3] I. J. Goodfellow, J. Shlens, and C. Szegedy *Explaining and Harnessing Adversarial Examples*, ICLR'15.