

VERONA: A Framework for Neural Network Robustness Experiments

Aaron Berger, Annelot Bosman, Hendrik Baacke, Holger H. Hoos, Jan N. van Rijn



Universiteit
Leiden

RWTHAACHEN
UNIVERSITY

Small Input Perturbations



Label: Stop sign

Small Input Perturbations



Label: Stop sign



Label: Do not enter

Small Input Perturbations



Label: Stop sign



Label: Do not enter



Label: 120 km/h

[<https://kennysong.github.io/adversarial.js/>]

Local Robustness

A model is locally robust at x_0 if small changes to the input do not change its prediction.

$$\forall x : \|x - x_0\|_\infty \leq \varepsilon \Rightarrow f(x) = f(x_0)$$

Local Robustness

A model is locally robust at x_0 if small changes to the input do not change its prediction.

$$\forall x : \|x - x_0\|_\infty \leq \varepsilon \Rightarrow f(x) = f(x_0)$$

Neural Network Verification

- Neural networks tend to be sensitive to input perturbations
- Use formal reasoning techniques for robustness verification

Neural Network Verification

- Neural networks tend to be sensitive to input perturbations
- Use formal reasoning techniques for robustness verification

Key idea: ensure all x close to given input x_0 are classified with same (correct) label.

Neural Network Verification

- Neural networks tend to be sensitive to input perturbations
- Use formal reasoning techniques for robustness verification

Key idea: ensure all x close to given input x_0 are classified with same (correct) label.

$$x : \|x - x_0\|_\infty \leq \epsilon \Rightarrow f(x) = f(x_0)$$

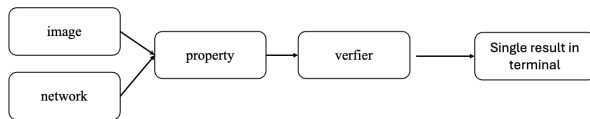
Neural Network Verification

- Neural networks tend to be sensitive to input perturbations
- Use formal reasoning techniques for robustness verification

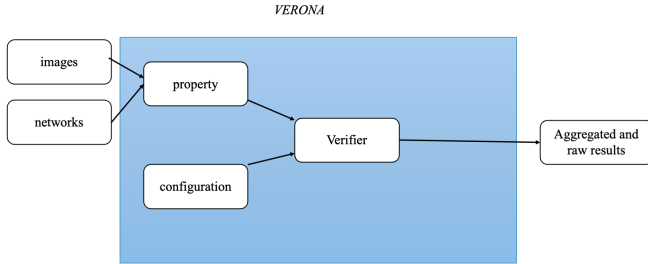
Key idea: ensure all x close to given input x_0 are classified with same (correct) label.

$$x : \|x - x_0\|_\infty \leq \epsilon \Rightarrow f(x) = f(x_0)$$

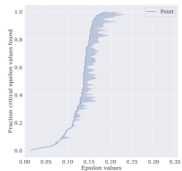
Expensive to solve!



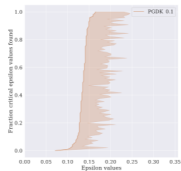
VERONA



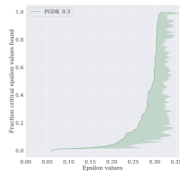
Analysing Robustness in Neural Networks



(a) Point – training
 $KS = 0.004$



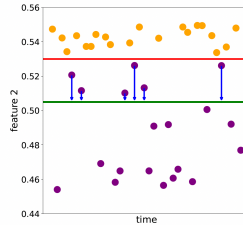
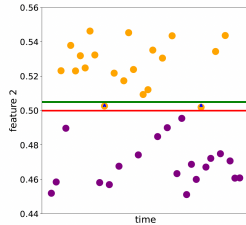
(b) PGDK w 0.1– training
 $KS = 0.081$



(c) PGDK w 0.3 – training
 $KS = 0.000$

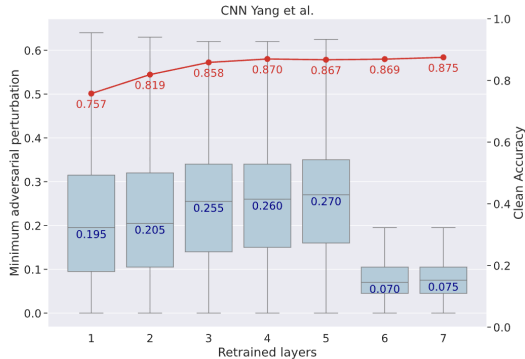
¹Bosman, A., Berger, A., Hoos, H. H., & Van Rijn, J. N. (2025). Robustness distributions in neural network verification. Journal of Artificial Intelligence Research, 83.

Changing data



²Rodi Laanen, *Annelot Bosman, Jan van Rijn, Bruno Veloso and Holger Hoos*. Concept Drift Detection using the Distance to the Decision Boundary. Master thesis

Robust transfer learning



³David P. Wünsch, Annelot Willemijn Bosman, Holger H. Hoos, Jan N. van Rijn (2026). Transfer Learning of Robustness for Image Classification: An Experimental Study using Robustness Distributions. AutoML 2026.

$$x' \in B_\varepsilon(x)$$

Is the prediction stable for all inputs within
a mathematically defined region?

$$x' \in B_\varepsilon(x)$$

Is the prediction stable for all inputs within a mathematically defined region?

What could we study?

- sensor degradation effect on accuracy
- spatial / spectral resolution
- geographic or temporal variation
- missing or noisy information

What could we study?

$$x' \in B_\varepsilon(x)$$

Is the prediction stable for all inputs within a mathematically defined region?

- sensor degradation effect on accuracy
- spatial / spectral resolution
- geographic or temporal variation
- missing or noisy information

How can we define properties relevant to earth observation?

To conclude

- What would you like to understand about your models?
- What would you like to see guaranteed?

`a.w.bosman@liacs.leidenuniv.nl`



Figure: VERONA code