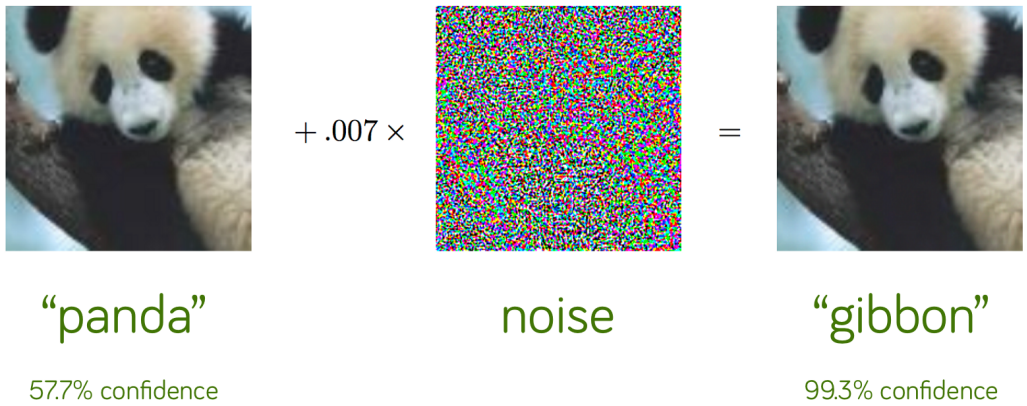
Verification of Neural Networks

Tobias Ladner

Prof. Dr.-Ing. Matthias Althoff
Cyber-Physical Systems Group
Technische Universität München

July 14th, 2022

Example: Image Classification



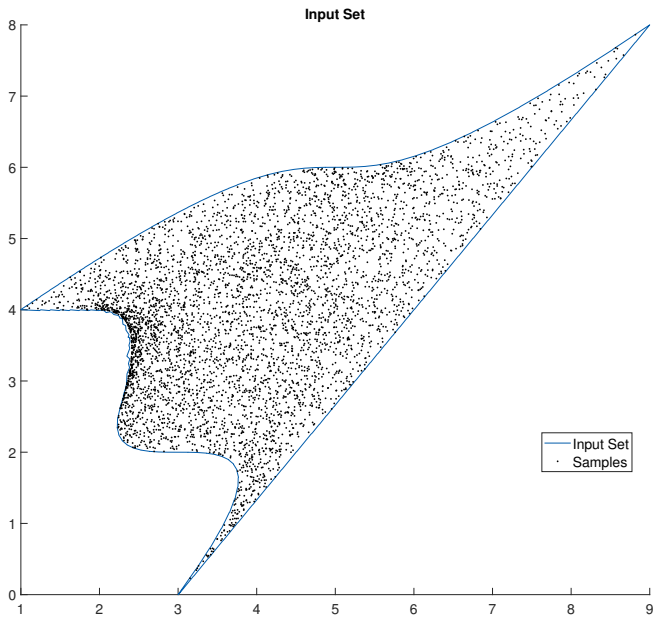
[1] Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

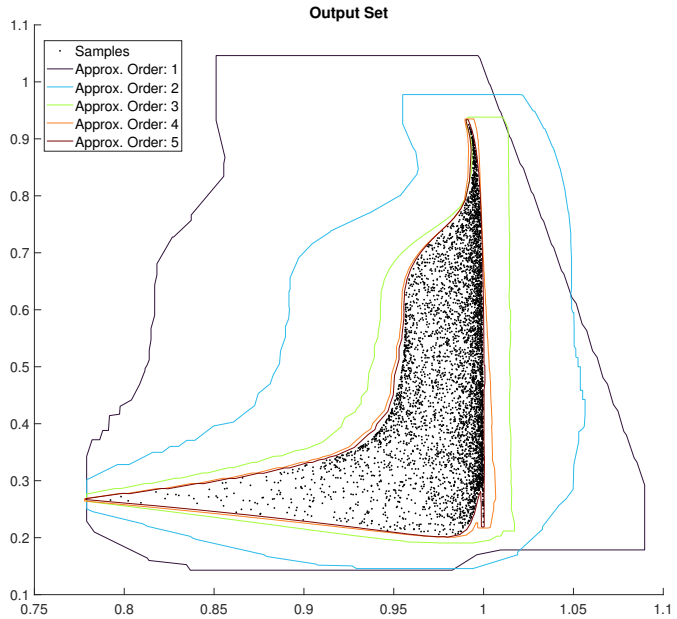
Goal:

Verify that NN is robust to adversarial attack on an image x_0:

- Specify max. perturbation ϵ,
- Verify correct prediction $\forall x$ with $\|x - x_o\|_\infty \leq \epsilon$.

How to Verify $\forall x$

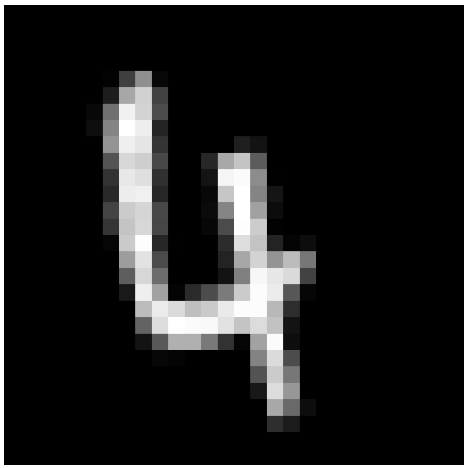


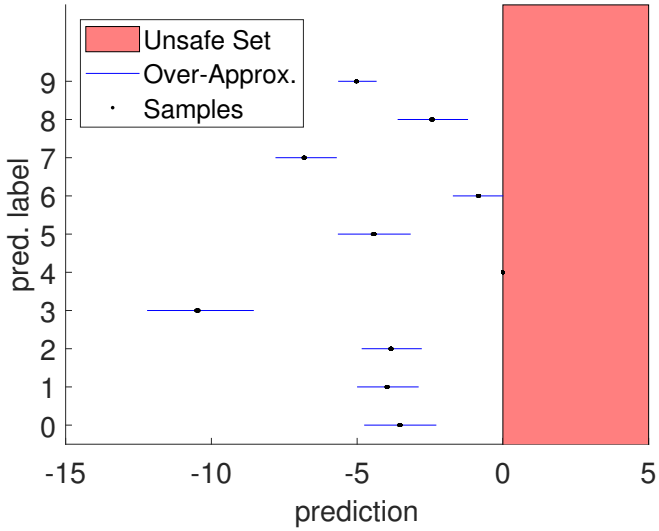


Verification of Neural Network:

- Propagate set containing all x through the network instead of individual samples
- Enclose output set as tight as possible

Topic 1: Evaluate Benchmarks





Your tasks:

- Understand the verification of neural networks
- Implement various benchmarks in CORA
- Evaluate results

Interested? Contact me!

Tobias Ladner
tobias.ladner@tum.de

Topic 2: Advanced NN Architectures

NN Architectures and Layers:

- Convolutional NN
- Recurrent NN
- ...

Your tasks:

- Understand the verification of neural networks
- Implement abstract transformers for new NN architectures in CORA
- Evaluate results

Interested? Contact me!

Tobias Ladner
tobias.ladner@tum.de

Topic 3: Restructuring Neural Networks

Smaller networks are usually easier to verify. Restructuring can help to decrease the size of neural networks.

Your tasks:

- Read relevant literature
- Implement promising approaches in CORA, e.g. [2]
- Evaluate results

Interested? Contact me!

Tobias Ladner
tobias.ladner@tum.de

[2] Elboher, Y. Y., Gottschlich, J., Katz, G. (2020, July). An abstraction-based framework for neural network verification. In International Conference on Computer Aided Verification (pp. 43-65). Springer, Cham.