Training Robust Neural Networks

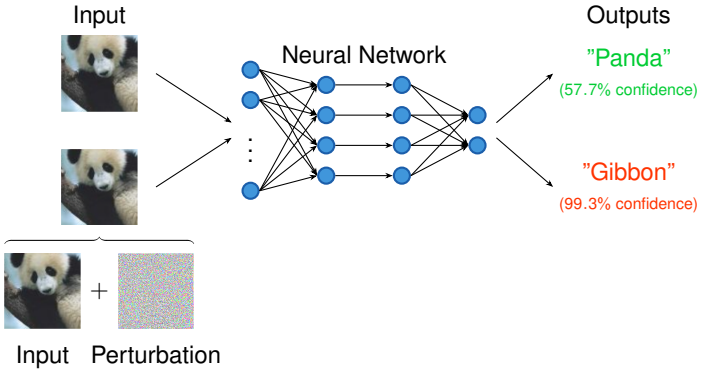
Lukas Koller

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

February 4, 2024

Neural Networks and Adversarial Attacks

Adversarial Attack: Small carefully chosen input perturbation that leads to a misclassification.

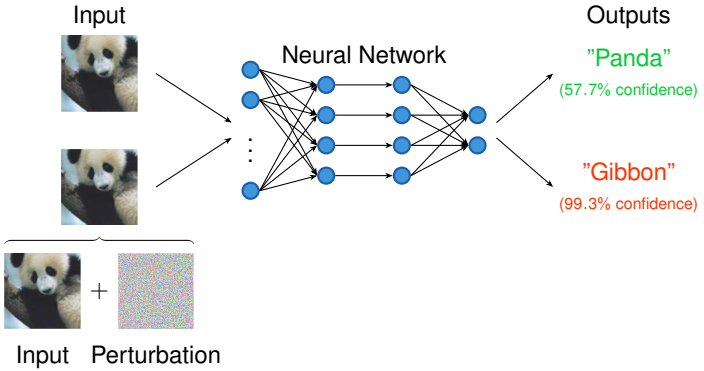


Neural networks are vulnerable to adversarial attacks![1]

[1] Goodfellow et al., "Explaining and Harnessing Adversarial Examples"

Neural Networks and Adversarial Attacks

Adversarial Attack: Small carefully chosen input perturbation that leads to a misclassification.



Neural networks are vulnerable to adversarial attacks![1]

[1] Goodfellow et al., "Explaining and Harnessing Adversarial Examples"

Training Robust Neural Networks

There are many different methods to train adversarially robust neural networks, e.g. [2], [3], [4], [5].

Your tasks: Review and compare different approaches to train and evaluate robust neural networks.

Interested? Contact me!

Lukas Koller
lukas.koller@tum.de

[2] Madry et al., “Towards Deep Learning Models Resistant to Adversarial Attacks”
[3] Gowal et al., “Scalable Verified Training for Provably Robust Image Classification”
[4] Mirman et al., “Differentiable Abstract Interpretation for Provably Robust Neural Networks”
[5] Zhang et al., “Theoretically Principled Trade-off between Robustness and Accuracy”