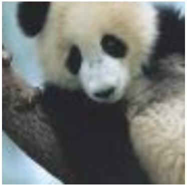
Formal Verification of Neural Networks

Tobias Ladner, Lukas Koller

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

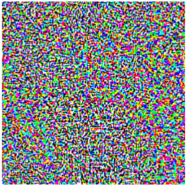
July 5th, 2024

Motivation



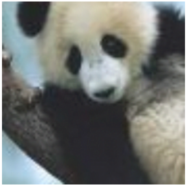
“panda”

$+\ 0.007 \times$



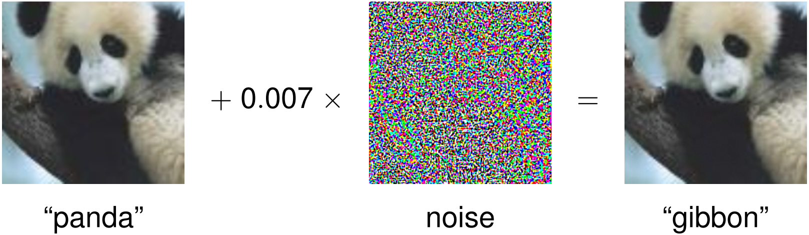
noise

$=$



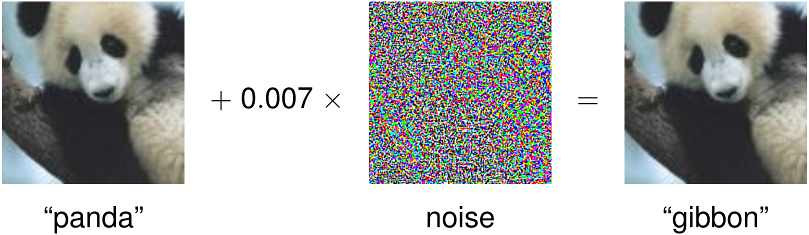
[1] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *International Conference on Learning Representations*. 2015

Motivation



[1] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *International Conference on Learning Representations*. 2015

Motivation

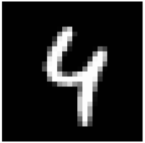


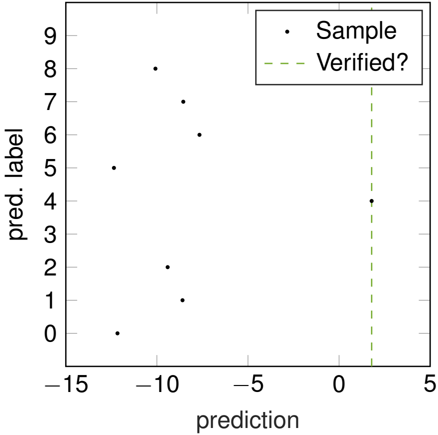
Adversarial attacks[1] limit the applicability of neural networks in cyber-physical systems!

[1] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *International Conference on Learning Representations*. 2015

Motivation

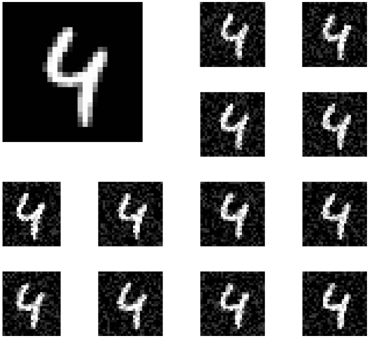
Let us demonstrate the formal verification of neural networks by an example:

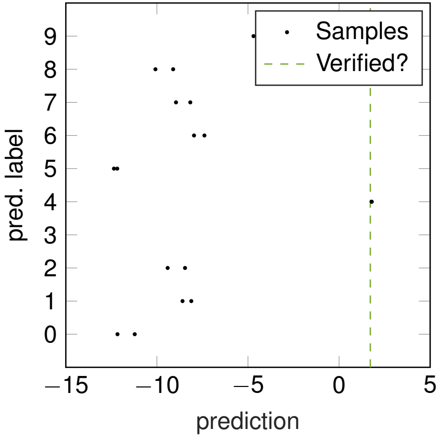




Motivation

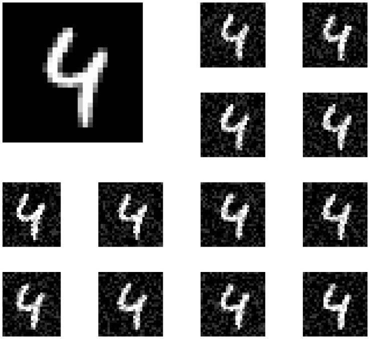
Let us demonstrate the formal verification of neural networks by an example:

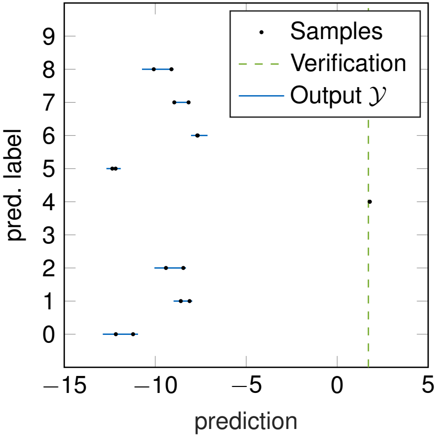




Motivation

Let us demonstrate the formal verification of neural networks by an example:





Topics

Possible topics:

- Design verification algorithms for general network architectures
- Enable easier verification by network design
- …

Furhter examples can be found on the CORA website[2].

Interested? Contact us!
Tobias Ladner, Lukas Koller
`tobias.ladner@tum.de`, `lukas.koller@tum.de`

[2] CORA website: `https://cora.in.tum.de/pages/neural-networks`