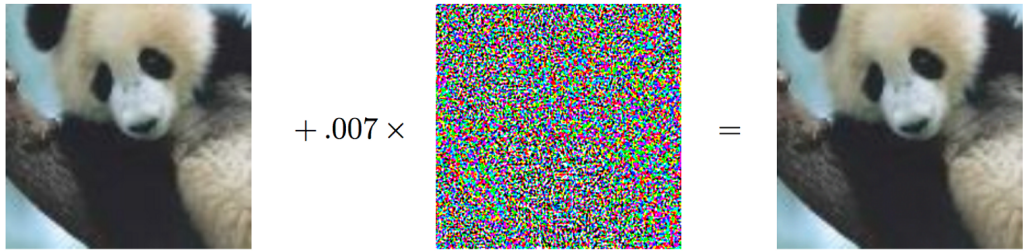
Neural Network Robustness Improving Techniques

Tobias Ladner

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

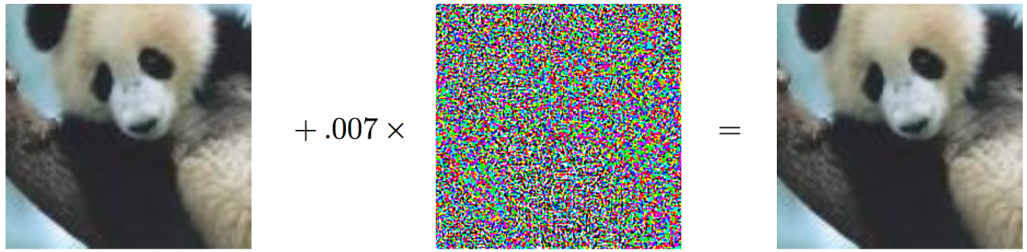
February 2nd, 2023

Motivation

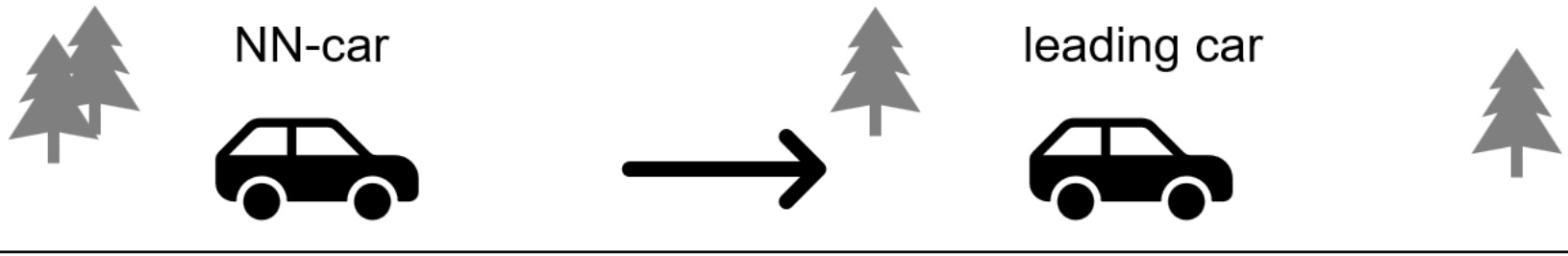


[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Motivation



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



Robustness of Neural Networks

Application: How can neural networks be made safe? Important to make them applicable in safety critical system. Examples: [2, 3]

Your tasks:

- Explore literature on robustness of neural networks
- Report + Presentation

Interested? Contact me!

Tobias Ladner
tobias.ladner@tum.de

[2] Parno, B., Pasareanu, C. Self-correcting Neural Networks for Safe Classification.
[3] Meng, M. H., et al. (2022). Adversarial robustness of deep neural networks: A survey from a formal verification perspective. IEEE Transactions on Dependable and Secure Computing.