TECHNISCHE UNIVERSITÄT MÜNCHEN
LEHRSTUHL FÜR INFORMATIONSTECHNISCHE REGELUNG
ORDINARIA: UNIV.-PROF. DR.-ING. SANDRA HIRCHE

MASTER'S THESIS

Safety verification for neural networks

Problem description:

Neural networks are a powerful tool for classification tasks, and have produced groundbreaking results in multiple application settings [4, 3]. However, making neural networks robust to adversarial attacks is still a highly challenging task, as the performance of most neural networks often falters in the presence of adversarial components [1]. In order to mitigate this shortcoming, different methods have been developed so far [2]. However, these algorithms tend to provide estimates that are overly conservative. The goal of this thesis is to identify safety critical adversarial attacks in the spirit of sensitivity analysis and eigenvalue perturbation. This in turn is to be employed to render the neural network more robust to such attacks.

Tasks:

- Literature research - Safety verification of neural networks
- Development of method for identification of safety critical attacks
- Development of robust training approaches
- Testing of developed approach using benchmark problems and subsequent comparison with established methods

Bibliography:

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Changliu Liu, Tomer Arnon, Christopher Lazarus, Clark Barrett, and Mykel J Kochenderfer. Algorithms for verifying deep neural networks. *arXiv preprint arXiv:1903.06758*, 2019.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[4] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Supervisor:	M. Sc. Alexandre Capone
Start:	XX.XX.2019
Intermediate Report:	XX.XX.2019
Delivery:	XX.XX.2020

(S. Hirche)
Univ.-Professor