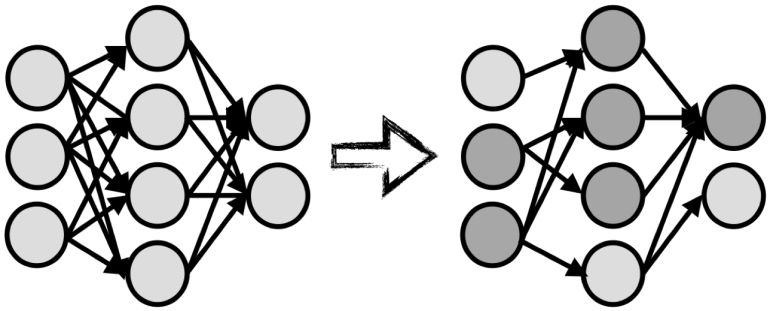
Efficient Adversarial Training for Robust and Sparse Deep Neural Networks



Deep neural networks (DNNs) are shown to be largely vulnerable to naturally or deliberately caused perturbations which are hardly perceptible by humans. This concern regarding DNN robustness becomes more pronounced in resource-constrained settings, when the model capacity and complexity can not be arbitrarily increased to compensate for robustness. In this context, adversarially robust training for DNNs under high model sparsity becomes an interesting research problem. In this project we will develop efficient and accelerated implementations for state-of-the-art sparse adversarial training algorithms that we have recently proposed [1]. We will also perform benchmark image classification experiments and adversarial robustness evaluations for these sparse DNNs.

[1] O. Özdenizci & R. Legenstein, "Training adversarially robust sparse networks via Bayesian connectivity sampling", International Conference on Machine Learning (ICML) 2021.

Goals & Tasks

- Review of the state-of-the-art on training adversarially robust and sparse DNNs.
- Efficient implementation of sparse adversarial training using sparse tensor operations on hardware accelerators.
- Experimenting with state-of-the-art adversarial attack benchmarks on sparse DNNs.

Qualifications

- Interest in deep learning.
- Experience with Python based deep learning frameworks such as TensorFlow and PyTorch are essential.
- Registered to one of the following:
  - ✓ **Bachelor Thesis**
  - ✓ **Seminar Project**
  - ✓ **Master Thesis**

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at