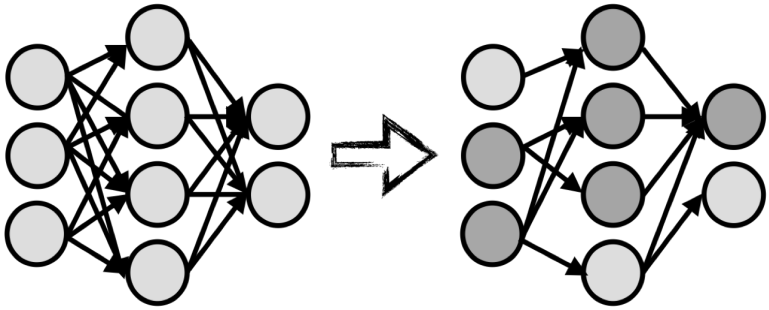
Robust Pruning of Deep Neural Networks by Harnessing Adversarial Decisions



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 pruning of deep neural networks to achieve sparsity becomes an interesting -and yet open- research problem. In this project we will develop robust pruning methods that are guided by state-of-the-art DNN interpretability pipelines. We will explore latent space vulnerability and relevances of DNNs in adversarial settings, and develop novel approaches to exploit these insights for robust pruning.

Goals & Tasks

- Review of the state-of-the-art on adversarially robust pruning methods.
- Implementing and experimenting with adversarial attacks on sparse networks.
- Exploring DNN interpretability methods for image classification to develop ideas for robust pruning.

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at

Qualifications

- Interest in deep learning.
- Experience with Python based deep learning frameworks such as TensorFlow or PyTorch are beneficial.
- Registered to one of the following:

  ✓ **Bachelor Thesis**

  ✓ **Seminar Project**

  ☐ **Master Thesis**

  Extension for a Master Thesis is possible on interesting results.