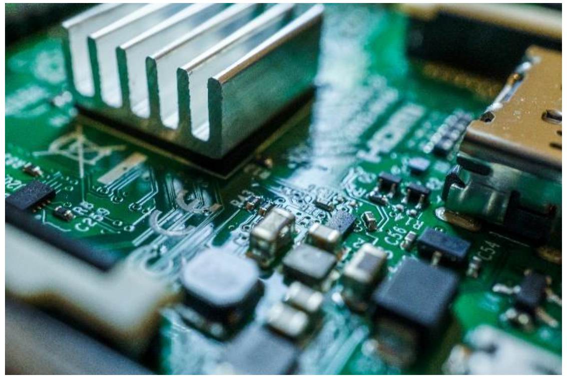
Complexity Reduction for Neural Networks with Focus on Embedded Applications

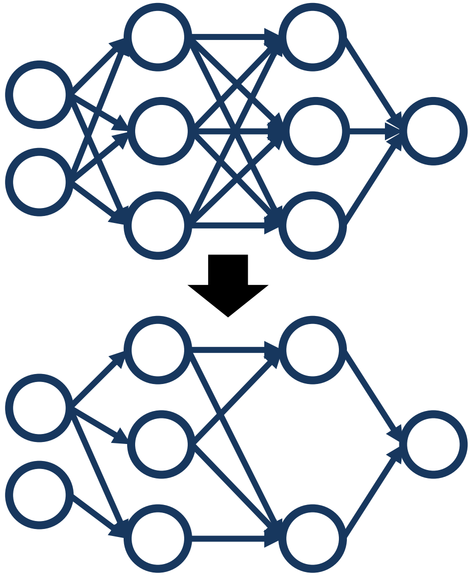
TECHNISCHE
UNIVERSITÄT
DARMSTADT

Proposal for a Bachelor Thesis

Artificial Neural Networks (NN) have seen widespread adoption in many fields, including control engineering, as they offer a more efficient alternative to classical modeling techniques in many control applications. This is especially important for embedded systems, where the computational resources are severely limited compared to general-purpose computers. Even though using a trained NN to derive predictions (inference) is less complex than training the NN in the first place, reducing the needed computational power and inference time is still important for embedded devices.

In this thesis, you will investigate existing techniques to improve the inference performance of the Neural Network. We will concentrate on the techniques that quantize the domain of possible outputs of the hidden layers, as well as NN pruning, which simplifies the structure of the NN based on its weights. You will summarize methods for NN compression in the form of a literature review. Afterwards, you will implement a simple NN in a controlled setting and then apply discussed compression techniques, comparing their impact on inference quality and computational burden (inference time, memory requirements, etc.)





Useful Skills:

Knowledge:	Neural Networks, Norms
Programming:	MatLab or Python
Language:	English or German

Contact:

Supervisors:	Email:
Hendrik Alsmeier, M. Sc	hendrik.alsmeier@iat.tu-darmstadt.de
Anton Savchenko, Dr.-Ing.	anton.savchenko@iat.tu-darmstadt.de