Chair for Integrated Systems
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Master's Thesis

Learning to Prune and Quantize Transformers

Advances in the deep learning architectures for computer vision applications have lead to new neural architectures such as vision transformers. These differentiate themselves from typical convolutional neural network-based implementations by decoupling the process of feature aggregation and transformation. Excellent performance is achieved through self-attention and self-supervision.

In this master thesis, visual transformers will be implemented in the first step. Following verification of state-of-the-art results, the transformers will be compressed through quantization and pruning to minimize their computational complexity on the inference hardware.

Prerequisites

Prerequisites

To successfully complete this project, you should have the following skills and experiences:

- Good programming skills in Python and Tensorflow
- Good knowledge of neural networks, basic knowledge of transformers

The student is expected to be highly motivated and independent.

Contact

Nael Fasfous
Department of Electrical and Computer Engineering
Chair of Integrated Systems

Phone: +49.89.289.23858
Building: N1 (Theresienstr. 90)
Room: N2116
Email: nael.fasfous@tum.de

Advisors

Alexander Frickenstein, Nael Yousef Abdullah Al-Fasfous