Associate Professorship of Coding and Cryptography
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Seminar

Neural Network Post-Training Quantization

Neural networks achieve state-of-the-art performance in many complex machine learning tasks (e.g., Object Detection, Image Classification, Audio Recognition etc.) In doing so, the respective models (weights and biases) size has exploded. This results in great power consumption, high inference latency and icreased memory complexity. It is therefore of interest to find ways to compress these models in order to achieve energy savings, inference speed and storing requirements. One very popular method to do so, is quantization. The task of the student is to explain how Post Training Quantization (PTQ) is applied, given that fixed-point representation is assumed [1].

[1] 2106.08295.pdf (arxiv.org)

Prerequisites

It is nice for the student to have some background knowledge on deep learning, e.g., what is a neural network, how it is represented, how it is trained etc. However, introductory material can be provided if a student is eager to learn, and questions on topics that are unclear to the student are always welcome.

In general, this is a student-driven task, therefore it is the student's job to plan and execute the review of the given paper. Support and guidance will be gladly provided if requested. There will also be a clear discussion of what is required in the final presentation as well as evaluation points, directly after the topic assignment.

Advisors

Paraskevi Papadopoulou