A simple approach for quantizing neural networks

Johannes Maly
LMU Munich, Germany
maly@math.lmu.de

In this talk, I propose a new method for quantizing the weights of a fully trained neural network. A simple deterministic pre-processing step allows to quantize network layers via memoryless scalar quantization while preserving the network performance on given training data. On one hand, the computational complexity of this pre-processing slightly exceeds that of state-of-the-art algorithms in the literature. On the other hand, the new approach does not require any hyper-parameter tuning and, in contrast to previous methods, allows a plain analysis. I provide rigorous theoretical guarantees in the case of quantizing single network layers and show that the relative error decays with the number of parameters in the network if the training data behaves well, e.g., if it is sampled from suitable random distributions. The developed method also readily allows the quantization of deep networks by consecutive application to single layers.

Joint work with Rayan Saab (UCSD).