
Contribution ID: **49** Type: **Standard Talk**

# Efficient Quantization of Deep Learning Models for Hardware Acceleration

*Wednesday 27 September 2023 13:45 (15 minutes)*

Today's deep learning models consume considerable computation and memory resources, leading to significant energy consumption. To address the computation and memory challenges, quantization is often used for storing and computing data as few as possible. However, exploiting efficient quantization for computing a given ML model is challenging, because it affects both the computation accuracy and hardware efficiency. In this work, we propose a fully automated toolflow, named Machine-learning Accelerator System Explorer (MASE), for exploration of efficient arithmetic for quantization and hardware mapping. MASE takes a deep learning model and represents it as a graph representation of both the software model and the hardware accelerator architecture. , This enables both coarse-grained and fine-grained optimization in both software and hardware. MASE implements a collection of arithmetic types, and supports mixed-arithmetic quantization search in mixed precisions. We evaluate our approach on OPT, an open-source version of the GPT model, and show that our approach achieves a 19× arithmetic density and a 5× memory density compared to the float32 baseline, surpassing the prior art 8-bit quantisation by 2.5× in arithmetic density and 1.2× in memory density.

**Authors:** ZHANG, Cheng (Imperial College London); Mr CHENG, Jianyi (University of Cambridge); CONSTANTINIDES, George (Imperial College London); ZHAO, Yiren

**Presenters:** ZHANG, Cheng (Imperial College London); Mr CHENG, Jianyi (University of Cambridge)

**Session Classification:** Contributed Talks

**Track Classification:** Contributed Talks