Speaker: Hengyi LI (Ritsumeikan University, Japan)

Biography: Hengyi LI received the Ph.D. degree of advanced electrical, electronic and computer systems from Ritsumeikan University, Japan in 2023. He is currently a senior researcher of Research Organization of Science and Technology, Ritsumeikan University. His research interests include High-performance computing for Artificial Intelligence (AI), Computer architecture, FPGA-based accelerator for AI, etc.

Title: Deep Learning Model Optimization and Acceleration

Abstract: Aiming for high-performance computing of DNNs, we conducts comprehensive and thorough studies on the layer-wise characteristics of typical DNN architectures, as well as the instruction-level deep workloads of DNN inference on Single Instruction Multiple Data (SIMD) CPUs; First and foremost, the data provides the fundamental theoretical basis for the following high-performance computing research. At the software level, the research puts forward two methods for compressing DNNs: the refined channel-level pruning method based on the layer-wise sparsity and channel-wise important indexes (SI-Pruning), and layer-level pruning (LL-Pruning) to optimize DNNs. As for the acceleration at the hardware level, we challenge to accelerate DNNs at the SIMD-instruction level at first; Furthermore, we implement the acceleration of DNNs on FPGA. To be exact, DNN optimization is to improve the efficiency of the feature extraction capability of DNNs. In view of this, we further propose an enhanced pooling function max-average pooling (FMAPooling) and an improved channel-attention mechanism (FMAttn) to enhance the feature extraction capability of DNNs.