Deep Model Compression and Acceleration towards On-Sensor AI

Changkyu Choi, Vice President, Samsung Advanced Institute of Technology

ML Inference on embedded hardware has been attracting attention for many applications. Limitations on computing power, memory usage, and power consumption are major bottleneck to deploy deep neural networks on resource-constrained devices. Existing techniques to mitigate the limitations often involve power-latency tradeoff and yield degraded accuracy.

This talk presents algorithm studies of deep model compression and computational acceleration on facial recognition tasks including face detection and anti-spoofing. A trainable quantizer is proposed to learn intervals to quantize activations and weights. This quantization-interval-learning allows the quantized networks to maintain the accuracy of the full precision (32-bit) networks with bit-width of activations and weights as low as 4-bit. The effectiveness of our trainable quantizer on ImageNet dataset will be demonstrated with various network architectures such as AlexNet, VGG-16, ResNet-50, Inception-V3, and so on.

Furthermore, this talk proposes ‘On-sensor AI’ computing architecture while exploring the advantages of 4-bit (or less) operations of activations and weights. 4-bit MAC operation can be implemented by utilizing AND and COUNT operations only. The speed of computation of AND/COUNT implementation is similar to the one of Multiplier/Adder implementation. However, it becomes way faster if the bit-width goes down to 3-bit or less. Moreover, replacing the Multiplier & Adder with logical AND & COUNT operators can reduce the number of transistor counts by over 40 times.

This ‘On-sensor AI’ is challenging but promising. This will pave the way to deploy highly accurate tiny ML models to things in everyday life.