Artificial Intelligence and Machine-Learning Hardware for Resource-Constrained Mobile Devices

Mingoo Seok, Associate Professor, Columbia University

Recent advances in Artificial Intelligence (AI) have unprecedentedly improved the accuracies in large-scale recognition and classification tasks. However, the computational complexity, memory access, and the associated energy cost limit the implementation of machine learning and post-learning operation in the systems and devices that have limited resources. The edge device in the Internet of Things (IoT) ecosystems is a good example of systems having limited resources. With the sought-after AI capabilities, however, the edge device can have more features, better accuracy in digital processing and autonomous decision, and lower system power dissipation (by scaling wireless data-rate). Here, we will discuss those challenges, namely in designing AI and ML hardware for resource-constrained mobile devices and present three of our recent works that span across algorithm, architecture, and circuits, and some combinations of those, to address the challenge.

In one project, we designed an in-memory computing SRAM macro that computes XNOR-and-accumulate in binary/ternary deep neural networks on the bitline. It uses the analog mixed-signal (AMS) resistive circuit for computing and does not require the row-by-row data access that is difficult to avoid in the conventional SRAM. The macro is prototyped in a 65-nm CMOS process. It demonstrates 33X better energy-efficiency and 300X better energy-delay product than digital ASIC of the same function, and also achieves significantly higher accuracy than prior AMS in-SRAM computing macro (e.g., 98.3% vs. 90% for MNIST) by being able to support the mainstream DNN/CNN algorithms.

In another project, we designed a sub-microwatt end-to-end Neural Signal Processing (NSP) systems for motor-intention decoding. This 96-channel Brain-Computer-Interface (BCI) implant features intercellular spike detection, sorting, and decoding, for prosthetics. We designed new algorithms, namely boundary based decision in sorting and ensemble-averaging-first computation in decoding. These algorithms achieve minimal computation complexity while matching or advancing the accuracy of state-of-art BCI algorithms. Based on our algorithms, we architect the VLSI hardware with the focus on hardware reuse and event-driven operation. The VLSI implementation of the proposed systems in a 180-nm CMOS shows more than 20X improvement in power efficiency over the prior arts.

Finally, we explored to combine analog and digital computing paradigms to improve energy efficiency in speech recognition. We designed analog and mixed-signal circuits that can directly extract relevant features from raw audio signals. These features enter the digital deep neural network with binary weights to perform classification for Voice Activity Detection (VAD). Our end-to-end systems consume less than 1 microwatt, more than 10X more power efficient than the previous techniques while achieving high accuracies for the input signals with various noise scenario and low SNR.

Building upon those prior efforts, currently, we are pursuing further research and prototype to enable more capable AI and ML in tiny devices. We are looking forward to further discussing the potential collaborations with the attendees during the tinyML summit.