A ½ mWatt, 128-MAC Sparsity Aware Neural Processing Unit for Classification and Semantic Segmentation

Joseph Hassoun, Sr. Director of Neural Processor Architecture, Samsung Semiconductor

This Presentation describes an energy-efficient neural processing unit for battery-operated devices. The architecture utilizing threefold of parallelisms for computing Convolutional and Fully Connected layers to achieve object detection for the at-the far-edge-computation. In this presentation, we will present the underlying technology of co-designing neural net models and neural net accelerators to achieve the right tradeoff for the highest energy efficiency. This 128-MAC structure is capable of running a low-precision modified 2-bit Group-Net Network that can perform Image classification and accurate semantic segmentation of 23 frames per seconds while operating at one-half of one mWatt.