Audio and Image Cross-Modal Intelligence via a 10TOPS/W 22nm SoC with Back-Propagation and Dynamic Power Gating

Zichen Fan, Hyochan An, Qiuru Zhang, Boxun Xu, Li Xu, Chien-wei Tseng, Yimai Peng, Ang Cao, Bowen Liu, Changwoo Lee, Zehong Wang, Fanghao Liu, Guanru Wang, Shenghao Jiang, Hun-Seok Kim, David Blaauw, Dennis Sylvester
University of Michigan, Ann Arbor, USA. zchen@umich.edu

Abstract We present an ultra-low-power multimedia signal processor (MSSP) SoC that integrates a versatile deep neural network (DNN) engine with audio and image signal processing accelerators for face recognition and deep voiceCMD. The proposed MSSP features 2MB MRAM to store all DNN weights on-chip with an energy-efficient dataflow using an MRAM-cache and dynamic power gating. The SoC achieves up to 3-10 TOPS/W peak energy efficiency and consumes only 0.25-3.84 mW. Being the first to demonstrate CNN, GAN, and back-propagation (BP) on a single accelerator SoC for cross-modal fusion, it outperforms state-of-the-art DNN processors by 1.4-4.5× in energy efficiency.

Introduction: DNN-based image and/or audio processing has been widely adopted in intelligent IoT systems. However, the traditional processing flow (Fig. 1 top) that offloads DNN processing to the cloud suffers from bandwidth, energy and privacy problems for resource-constrained IoT applications. To tackle these challenges, the proposed SoC adopts a fully-at-edge processing flow for audio and image multimodal intelligence with integrated pre-/post-processing accelerators (JPEG, H.264, FFT, and Mel-spectrum), a GAN/BP-capable neural engine, and 2MB on-chip MRAM for non-volatile weight/data/code storage. Fig. 2 depicts target application scenarios (a) DNN with a simple computational block, (b) MRAM for the computation logic and SRAM for the model parameters, and (c) MRAM macro is accompanied with an on-chip row/col buffer (8b×8×8 array, 12b per pixel VGA images and 8kHz 8b per sample audio signals using dedicated interfaces. The SoC can perform image- or audio-only applications such as face detection or keyword recognition, and also (conditionally) execute image/audio fusion applications such as cross-modal verification and active speaker detection to further increase the detection and recognition credibility. To accomplish this, the neural engine supports different types of DNN operations such as (depth-wise) convolution and fully connected layers (FCIs) for ANN, convolution back-propagation for GAN, convolution back-propagation for BP, and back-propagation for BPGAN [3], and on-chip weight retraining (transfer learning of the last layer).

Architecture: Fig. 3 shows the overall SoC architecture. All sub-blocks communicate with each other using an AHB bus. The audio interface performs audio feature extraction using frame buffers, a 256-point FFT unit, a 32/64 channel Mel-filter unit, and a power-to-dB log unit. The image interface features a change detection and on-the-fly JPEG compressed-memory [4] to temporarily store VGA frames in a compressed format. Only the change detected macro-block is stored and processed, creating a zero-leakage power path if MRAM access is not required (offstream). The H.264 engine performs image compression [4] on non-rectangular Rols for compressed storage in on/off-chip memory. The neural engine (NE) is a reconfigurable DNN accelerator that supports various operations including (depth-wise) convolution, deconvolution, FCL, and back-propagation. NE supports DNN weights both in uncompressed (8b/weight) and Huffman-compressed (~2b/weight) formats. The SoC integrates a 2MB MRAM macro to store all weights on-chip, and a 1.5MB multibank SRAM activation memory to store all feature maps for backpropagation. For simple applications that only require partial activation memory, unused SRAM macros are dynamically power gated to save leakage power. The main computation unit is an 8×8×8 processing element (PE, each with an 8-bit MAC) array, which enables activation and weight reuse via inter-PE connections. The top 1×8×8 PEs are multi-functional PEs (MPEs) that supports both MAC and max/average pooling operations. Furthermore, ping-pong memory structures for the local memristors and ping-pong non-blocking data movement and computation to increase PE utilization. The MRAM macro is accompanied with an MRAM cache to enable a dynamic power gating scheme (details in later section).

Power Domain: Fig. 4 shows the power domain design of the SoC. We implement power gating for each SRAM block and MRAM macro to decrease standby leakage power up to 83%. Because the MRAM can retain all data, SRAMs can be power gated when NE is inactive. Most of the computation logic can also be power gated while the always-on block (pads, headers, state registers, and power sequence control logic) consumes only 460nW.

NE Dataflow: Fig. 5 shows the energy-efficient, computation-skipping dataflow scheme implemented in NE. The base dataflow is output stationary, which is used in convolution, strided convolution, backpropagation, deconvolution, and other state-of-the-art DNN processors. Only the change detected or keyword recognized, processing to convolution and fully connected layers is integrated pre/post-processing flow. For convolution, only the convolution kernel is stored in on-chip memory, and consumed only 0.25-3.84 mW. For deconvolution, a zero-skipping dataflow exploits the deterministic and regular pattern of zero padding to increase throughput by 4× by only computing non-zero values in each PE (Fig. 5 left). For back-propagation through a ReLU layer, the zero activation positions are pre-recorded during the forward path to data-gate all computations in the backward path if they (back)propagate to a pre-recorded zero activation position (Fig. 5 right).

MRAM-cache Dynamic Power Gating: Although MRAM has the advantage of high density and non-volatility, its active leakage and readout power can be significant for ULP applications. To mitigate this issue, we propose an MRAM-cache architecture and dynamic power gating scheme (Fig. 6). In our scheme, MRAM is power gated until NE executes the load weight (LD WEIGHT) instruction. During LD WEIGHT operation, MRAM powers up and weights are read and loaded into the SRAM-based weight cache. After that, this MRAM block is power gated until NE supports either 1) sleep (SLP) mode where the MRAM array is powered off but peripherals remain on, or 2) power-down (PD) mode where the peripherals are also power gated. The optimal selection between SLP and PD depends on the reuse factor of the cached weight, as shown in Fig. 6 (bottom, right), which can be identified during NE programing. The measured MRAM VDIO current waveform shows the whole MRAM-cache sequence. During NN processing, weights are read from the weight cache (while MRAM is in either SLP or PD) for both convolution and fully connected layer operations. Based on measured results, the combined weight caching and power gating reduces weight readout power by 95.3% with only slightly increased (4.3%) operation time due to MRAM wake-up latency and cache loading time overhead.

The table in Fig. 6 summarizes the three different power modes: PD, SLP, and stand-by (STB). The analysis concludes that when each cached weight is reused for 353.4 MACs, the PD mode is preferable. Otherwise, SLP mode has an advantage because of the embedded SRAM (offstream) usage. Measurement Results: The SoC was fabricated in 22nm ULL technology and the die photo is shown in Fig. 9. The peak energy efficiency for various NN instructions is shown in Fig. 7 (left). The efficiency for convolution / deconvolution / stride-convolution-backpropagation (CONV / DECONV / S_CONV_BP) is 3.1 / 10 / 10 TOPS/W. The efficiency for DECONV and S_CONV BP is significantly higher because of the zero-skipping dataflow. The convolution backpropagation can achieve 3.7 TOPS/W with zero-gating dataflow. Deep-wise convolution (DWCONV) and FCL have lower efficiencies due to only 1/8 of the PE array being utilized. Fig. 7 (right) shows the voltage-frequency-efficiency tradeoff for CONV. The SoC achieves the highest energy efficiency at 0.46V (VDD_MAIN) and 1.2MHz, while the total system power is 387uW. Fig. 7 (right) shows that MRAM dynamic power gating enhances the energy efficiency especially at low voltage as it significantly reduces the leakage. Fig. 8 demonstrates a person-of-interest (PoI) tracking scenario and chip performance for that cross-modal intelligence scenario. This task first performs face detection (FD) on change-detected regions and then face recognition (FR) if faces are detected. The H.264 engine compresses the changed blocks if a PoI is recognized. In the meantime, the audio interface extracts the audio features (AFE), and cross-modal verification (CMV) is performed using both the audio features and face image. If the CMV model confirms the audio is matched with the target face, the audio signal is compressed (AC) by BPGAN. Fig. 8 (bottom) shows the power consumption and latency of each step in this process. Table 1 compares this SoC and other state-of-the-art prior designs in similar application spaces.
Fig. 2 Proposed multimedia signal processing SoC and its supported applications.

References:

Fig. 3 System overview architecture.

Convolution-RelU Back-propagation

Zero-skipping dataflow for Deconvolution forward-propagation

Fig. 5 Efficient dataflow for both forward and backward propagation.

Fig. 7 Measured peak efficiency and voltage-frequency-efficiency scaling

Fig. 9 Die photo and chip summary.

Table 1 Comparison table versus state-of-the-art.

Authorized license use limited to: University of Michigan Library. Downloaded on September 26, 2022 at 19:04:26 UTC from IEEE XplorE. Restrictions apply.