An In-Memory-Computing Charge-Domain Ternary CNN Classifier

Xiangxing Yang1, Keren Zhu1, Xiuyuan Tang1, Meizhi Wang1, Mingtao Zhan2, Nanshu Lu3, Jaydeep P. Kulkarni1, David Z. Pan1, Yongpan Liu2, Nan Sun1

1University of Texas at Austin, Austin, TX
2Tsinghua University, Beijing, China

Al edge devices require local intelligence for the concerns of latency and privacy. Given the accuracy and energy constraints, low-power convolutional neural networks (CNNs) are gaining popularity. To alleviate the high memory access energy and computational cost of large CNN models, prior works have proposed promising approaches including in-memory-computing (IMC) [1], mixed-signal multiply-and-accumulate (MAC) calculation [2], and reduced resolution network [3–4]. With weights and activations restricted to ±1, binary neural network (BNN) combining with IMC greatly improves sparsity and energy efficiency, which is well suited for edge-based applications, and has demonstrated state-of-the-art energy efficiency in image classification problems [5]. However, compared to full resolution network, BNN requires larger model thus more operations (OPs) per inference for a certain accuracy. To address such challenge, we propose a mixed-signal ternary CNN based processor featuring higher energy efficiency than BNN. It confers several key improvements: 1) the proposed ternary network provides 1.5-b resolution (0/1/+1), leading to 3.9x OPs/inference reduction than BNN for the same MNIST accuracy; 2) a 1.5b MAC is implemented by V\textsubscript{CM}-based capacitor accumulator in scheme, which inherently benefits from the reduced signal swing on the capacitive DAC (CDAC); 3) the V\textsubscript{CM}-based MAC introduces sparsity during training, resulting in lower switching rate. With a complete neural network on chip, the proposed design realizes 97.1% MNIST accuracy with only 0.18J per classification, presenting the highest power efficiency for comparable MNIST accuracy.

Fig. 1 shows the chip architecture and neural network topology of the proposed accelerator. The data path consists of 1 digital CNN layer at the input, 2 mixed-signal CNN layers followed by max-pooling layers, 1 SRAM bank to store image data, and 1 mixed-signal fully connected (FC) layer at the end. All weights/biases are trained on Tensorflow and loaded onto on-chip memory before classification. The weight memory is integrated with computations to mitigate the data movement cost. To exploit hardware parallelism and regularity, the number of channels for each CNN layer is 32 with 2×2 filters. Before feeding into the chip, the 8-bit pixel values from MNIST dataset will be quantized to a tri-level picture and zero-paddings to 30x30. The input layer CONV1 takes the ternarized data and compute the results digitally. The 128 pixels image data for the next convolution layer are generated in one clock cycle by stacking four 32-channel computational logics. Once the 256b data is ready at the CONV1 output, the 32-channel parallel switched-capacitor (SC) neuron MACs perform MAC with the 256b data. The results are stored in SRAM for CONV3, which is implemented the same way as CONV2. Then CONV3 outputs are accumulated in the FC layer 32 channels. Once data loading is completed, the final result is computed with the weights of all the digits.

Fig. 2 shows the comparison between the mixed-signal BNN and the proposed ternary neural network (TNN). The weighted sum of the compared results, C\textsubscript{1}/C\textsubscript{2}, is latched at comparator output, and then charged to a tri-level capacitor (C\textsubscript{3}). The result is ORed with Decoupled Read/Write and 1-

positive/negative thresholds V\textsubscript{REF}/-V\textsubscript{REF} perform ternary activation function. According to Monte-Carlo simulation, the comparator exhibits an 8.1mV standard deviation. One-time foreground calibration is performed to suppress the offsets to be within 1LSB. The calibration bits are loaded in SRAM during chip power-up. The convolution of a 32-channel image with a 2×2×32 filter requires 128 capacitors, and 32 capacitors are employed for the bias section.

Fig. 3 shows the data flow from the output of CONV1 to the input of CONV3. To boost the area efficiency, the 1.5b multiplier is designed with maximum density. The digital logic are routed with M1 to M3, while M4 to M6 above the transistors are used to implement the capacitive DACs. The channel of MAC calculation and activation is completed, the image pixel is latched at comparator output, and then stored in 1 of the 4 D flip-flops. When 4 of the 32-channel pixels are all computed and loaded into the registers, the max-pooling layer will be enabled to generate results. A 1352 bytes, 64b wide SRAM is placed at the output of max-pooling layer to store the entire frame of CONV2 output image with a size of 13×13×32. Before this image is picked up by CONV3 and apply sliding window convolution, an interchange multiplier is implemented to reduce memory access by 2x. For CONV3, the data flow is the same as CONV2, and the 12×12×32 output image will be downsampled to 6×6×32 for FC layer. Although demonstrated with 4-layer, 32-ch architecture for lightweight applications, the ternary neuron can be extended to fit in deeper neural network models.

Fig. 4 illustrates the architecture of FC layer. Each row represents one 32-channel image pixel. A total of 1152 pixels will be loaded after 36 activations of the previous max-pooling layer. All weights memory for number 0–9 is stored near multipliers and will be selected sequentially. Raw prediction logits are mapped to the charge on C\textsubscript{1}/C\textsubscript{2}. In the first cycle, the voltage representing number 0 will be redistributed and stored on C\textsubscript{1}, then the weights for number 1 are selected and the neurons acts again leaving the resulted voltage on C\textsubscript{2}. Based on the compared results, C\textsubscript{1}/C\textsubscript{2} with higher voltage will be kept, and the other one will be reset for storing the logit of the next number. After 9 comparisons, the final classification result is chosen as the number leaving highest voltage on C\textsubscript{1}/C\textsubscript{2}.

The prototype, fabricated in 40nm LP CMOS, occupies an active area of 0.96 mm\textsuperscript{2}. Measurement results, power breakdown and testing setup are shown in Fig. 5. The accuracy is evaluated on MNIST dataset. This chip operates at 549 FPS with 0.7V DVDD, 0.8V AVDD, and 0.9V V\textsubscript{REF}, leading to 0.18J/classification. The measured classification accuracy is 97.1%, which is 0.8% lowered than the ideal software model due to circuit noise, mismatch, and parasitic effects. This work constantly realizes the wide vector summation in charge domain, while [1][3][5] suffer from high switched capacitance of digital adders, and [4] consumes static current. Compared to [2] and [5] using BNN, it benefits from fewer OPs/inference and less switching activity. Moreover, this work performs all operations on chip, while [1][3][5] have only MAC operation. It consumes only 0.18J total energy for MNIST classification, which is the smallest to our best knowledge for comparable classification accuracy.

References:


**Fig. 1.** Architectural diagram of the proposed chip.

**Fig. 2.** Switch-capacitor ternary neuron architecture.

**Fig. 3.** Datapath from CONV1 output to CONV3 input.

**Fig. 4.** Fully-connect layer architecture (single-end shown)

**Fig. 5.** Measurement and power breakdown

**Fig. 6.** Comparison table