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

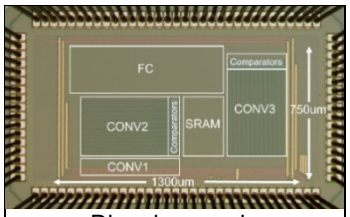
Xiangxing Yang[1], Keren Zhu[1], Xiyuan Tang[1], Meizhi Wang[1], Mingtao Zhan[2], Nanshu Lu[1], Jaydeep P. Kulkarni[1], David Z. Pan[1], Yongpan Liu[2], Nan Sun[1,2]

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

AI 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 the storage and computation efficiency, making it 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_{CM}-based capacitor switching scheme, which inherently benefits from the reduced signal swing on the capacitive DAC (CDAC); 3) the V_{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 0.18uJ 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 to 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-padding 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 CONV2 will process the data then pass it into the max-pooling logic. 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 multiplication results from filters and input pixels are computed by charge distribution, then the voltage at the charge conservation node passes into a comparator, which acts as the step type activation. In the binary case, +1/-1 are mapped as V_{REFP} and V_{REFN} in voltage domain, and the 2-level quantization is done by one comparator. In the proposed tri-level computation, 0/+1/-1 are represented by V_{CM}, V_{REFP}, V_{REFN}, respectively. Based on simulation, the introduction of V_{CM} reduces voltage swing on CDAC, thus providing 31% MAC power saving. In addition, extra sparsity can be introduced during training by enforcing more zero weights, resulting in further switching activity reduction. The tri-level quantization at the summing node is performed with a pair of differential comparators. The local 1.5b multiplier consists of 2 SRAM cells with stored weights, 2 logic gates as activation inputs, and a 1.5-b CDAC output. The 2 standard 6T SRAM cells are directly connected to computation logic and remain stationary during inference, amortizing the power from charging bit-line. With 0/+1/-1 coded as 0X/10/11, the 1.5b multiplication is performed efficiently by 1 AND and 1 XOR. Two comparators with positive/negative thresholds V_{REF+}/V_{REF-} perform ternary activation function. According to Monte-Carlo simulation, the comparator exhibits an offset with 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.



Die micrograph.

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 CDAC. Once 1 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 multiplexer 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 proposed 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_1 /C_2. In the first cycle, the voltage representing number 0 will be redistributed and stored on C_1, then the weights for number 1 are selected and the neurons acts again leaving the resulted voltage on C_2. Based on the compared results, C_1 or C_2 with higher voltage is 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_1/C_2.

The prototype, fabricated in 40nm LP CMOS, occupies an active area of 0.96 mm^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_{REFP}, leading to 0.18uJ/classification. The measured classification accuracy is 97.1%, which is 0.8% lowered than the ideal software model due to circuit noise, mismatch, and charge leakage. This work efficiently realizes the wide vector summation in charge domain, while [1][3] 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 0.18uJ total energy for MNIST classification, which is the smallest to our best knowledge for comparable classification accuracy.

References:

[1] K. Ando, et al., "BRein Memory: A Single-Chip Binary/Ternary Reconfigurable in-Memory Deep Neural Network Accelerator Achieving 1.4 TOPS at 0.6 W" JSSC, Apr. 2018.

[2] D. Bankman, et al., "An Always-On 3.8μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor with All Memory on Chip in 28nm CMOS" ISSCC, Feb. 2018.

[3] Y. Cheng, et al., "A 4-Kb 1-to-8-bit Configurable 6T SRAM-Based Computation-in-Memory Unit-Macro for CNN-Based AI Edge Processors" JSSC, Oct. 2020.

[4] C. Yu, et al., "A 16K Current-Based 8T SRAM Compute-In-Memory Macro with Decoupled Read/Write and 1-5bit Column ADC" CICC, April. 2020.

[5] H. Valavi, et al., "A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute" JSSC, Mar. 2019.

978-1-7281-7581-2/21/$31.00 ©2021 IEEE

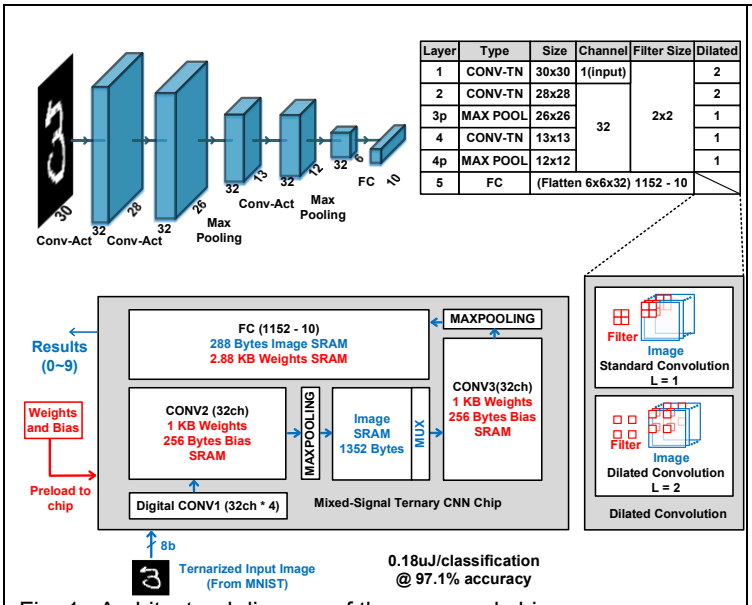


Fig. 1. Architectural diagram of the proposed chip.

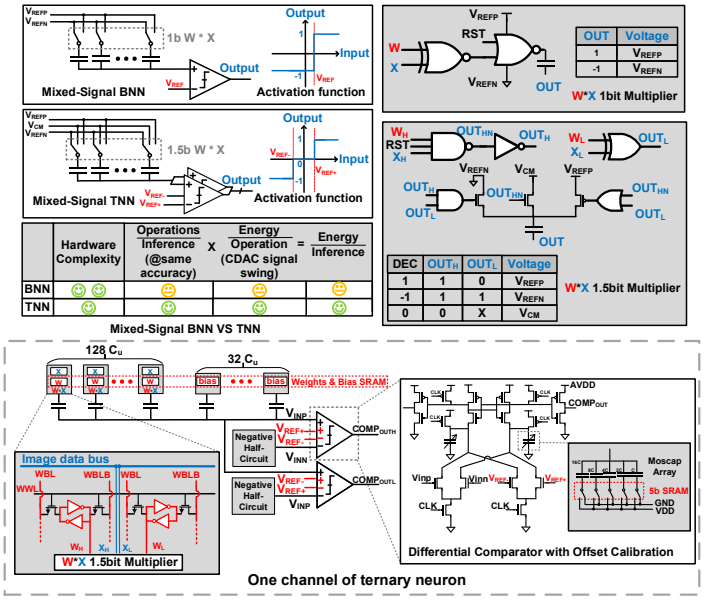


Fig. 2. Switch-capacitor ternary neuron architecture.

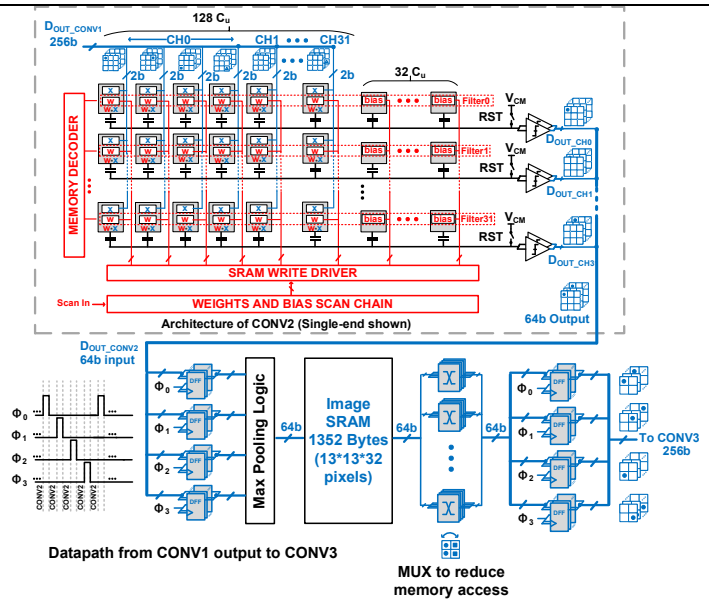


Fig. 3. Datapath from CONV1 output to CONV3 input.

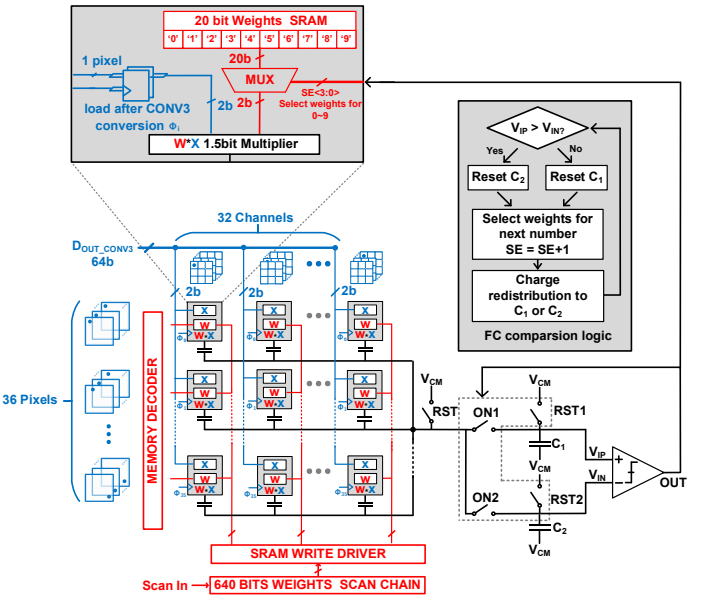
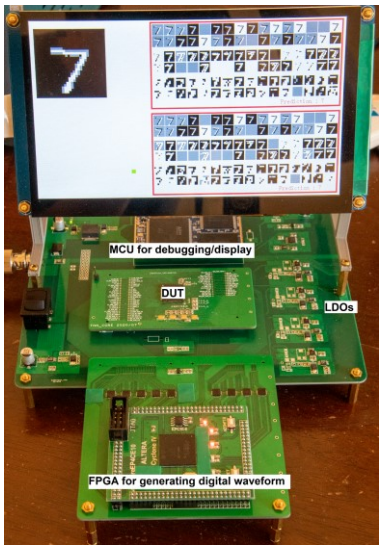
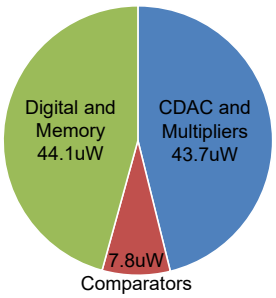


Fig. 4. Fully-connect layer architecture (single-end shown)





97.1% accuracy @ 549 FPS

Power domain	Voltage	Energy
DVDD	0.7 V	44.1 uW
AVDD	0.8 V	7.8 uW
V_{REFP}	0.9 V	37.8 uW
V_{CM}	0.45 V	5.9 uW

Fig. 5. Measurement and power breakdown

	This work	JSSC'18 [1]	ISSCC'18 [2]	JSSC'20 [3]	CICC'20 [4]	JSSC'19 [5]
Technology	**40nm**	65nm	28nm	55nm	65nm	65nm
Circuit Type	**Mixed-Signal**	Digital	Mixed-Signal	Mixed-Signal	Mixed-Signal	Mixed-Signal
Area(mm2)	**0.98**	3.9	4.6	5.85	0.055	12.6
Area Eff.(GOPS/mm2)	**469**[1]	105	67	N/A	N/A	1498
Operating VDD(V)	**0.8/0.7/0.9**	0.55-1.0	0.8/0.8	0.9	0.8/0.45	0.94/0.68/1.2
Energy Eff.(TOPS/W)	**556**[2]	2.3-6.0	532	40.2	490-15.8	866
Bit Precision	**1.5b**	1/1.5b	1b	1-8b	1-5b	1b
Dataset	**MNIST**	MNIST	CIFAR-10	MNIST	MNIST	MNIST
Accuracy	**97.1%**[3]	90.1%	86.05%	98.56%	96.2%	98.6%
FPS	**549**	N/A	237	N/A	N/A	651
Power(mW)	**0.096**	N/A	0.899	N/A	N/A	N/A
Operations / Classification (CL)	**3.57x10^7**	N/A	N/A	N/A	N/A	5.3x10^8
MACs Energy / CL	**0.09uJ**	N/A	N/A	N/A	N/A	0.8uJ
Total Energy / CL	**0.18uJ**	N/A	3.8uJ	N/A	N/A	N/A
All operations on chip	**Yes**	No	Yes	No	No	No

[1]Based on SC neuron
[2]Based on MACs energy efficiency
[3]10 runs average on 10,000 test set images.

Fig. 6. Comparison table