Considerations of Integrating Computing-In-Memory and Processing-In-Sensor into Convolutional Neural Network Accelerators for Low-Power Edge Devices

Kea-Tiong Tang[1], *Wei-Chen Wei[1], *Zuo-Wei Yeh[1], Tzu-Hsiang Hsu[1], Yen-Cheng Chiu[1], Cheng-Xin Xue[1], Yu-Chun Kuo[1], Tai-Hsing Wen[1], Mon-Shu Ho[2], Chung-Chuan Lo[1], Ren-Shuo Liu[1], Chih-Cheng Hsieh[1], Meng-Fan Chang[1]
[1]National Tsing Hua University, Hsinchu, Taiwan; [2]National Chung Hsin University, Taichung, Taiwan (*equal contribution)

Abstract

In quest to execute emerging deep learning algorithms at edge devices, developing low-power and low-latency deep learning accelerators (DLAs) have become top priority. To achieve this goal, data processing techniques in sensor and memory utilizing the array structure have drawn much attention. Processing-in-sensor (PIS) solutions could reduce data transfer; computing-in-memory (CIM) macros could reduce memory access and intermediate data movement. We propose a new architecture to integrate PIS and CIM to realize low-power DLA. The advantages of using these techniques and the challenges from system point-of-view are discussed.

Keywords: Deep learning accelerator, computing-in-memory, processing-in-sensor, artificial intelligence

Introduction

As convolutional neural network (CNN) use has expanded, demand for computing resources and consequent memory costs have grown significantly. Numerous works over the past few years have sought to develop a DLA [1–3], with the goal of effectively reducing the time complexity of matrix–vector multiplication [4–6]. To deploy CNNs in edge devices, high-throughput low-power DLA solutions that can execute emerging deep-learning algorithms has become mainstream.

A growing body of research seeks to utilize *in situ* analog computing to break the memory wall and reduce the communication overhead in DLA designs. Recent works have proven that computing-in-memory (CIM) macros, including in-memory computing (IMC) and near-memory computing (NMC), can break the von Neumann bottleneck to achieve high energy efficiency [7–13]. PIS shares the same design concept with CIM but focuses more on high-resolution computing, early detection, and image filtering [14, 15]. By conducting specific image processing before delivering data to the backend processor, bandwidth, power consumption, and latency of data transfer can be reduced effectively [14–17].

In this paper, we propose a novel architecture (Fig. 1) that integrates PIS and CIM into a low-power CNN accelerator for edge devices. The advantages and challenges are also outlined.

Computing-In-Memory

For edge devices, reducing the latency and reducing the energy consumption of multiply-and-accumulate (MAC) operations are two crucial challenges. CIM (Fig. 2) has been proposed to meet these requirements. Fig. 3 shows how to perform MAC operations using CIM. The multiply operation occurs in the memory cell when the word-line (WL) turns on/off and the product value is I_{MC}. Subsequently, the bit-line (BL) sums up every I_{MC} for the accumulation operation.

Both volatile and nonvolatile CIM macros have demonstrated MAC operations from binary [7, 18] to multibit precision [11, 12] (Fig. 4). To support more complex CNN algorithms in order to classify more complicated datasets, CIM should be capable of precise MAC operations through software/hardware co-design. Convolution is more complex in CIM than in a conventional digital design. The MAC of an input feature map convolution with a kernel must be decomposed to satisfy the CIM constraints. Moreover, the behavior of the multibit sensing amplifier must be considered.

Multibit CIM macros still face several challenges if they are to be integrated with high-accuracy CNN systems. First, multibit output precision reduces the sensing margin of each MAC value, which increases the error of the sensing amplifier output. Second, the computation of a MAC value in a memory array causes a large bit-line current, which increases not only power consumption but also the sensing amplifier input-offset. Finally, the CIM architecture computes the MAC in the analog domain while the peripheral circuits are digital. Consequently, the power bottleneck lies in the analog-to-digital conversion (ADC) and digital-to-analog conversion (DAC) at the I/O of the CIM macro. To effectively improve power efficiency, adaptive DAC and ADC must be considered [19].

Processing-In-Sensor

Recently, CMOS image sensors (CIS) with specific PIS functions for AI applications (see Fig. 5) have attracted attention. An architecture in [14] proposed an analog–digital hybrid image filtering in-sensor to reduce energy consumption. The imager demonstrated in [20] implemented GOPS column-parallel processing elements using 3D-stacked technology for high-speed spatiotemporal information extraction. A column-parallel computation scheme was presented in [15] for local binary pattern and edge extraction. Always-on motion detection cameras [16, 17] also benefit from PIS of feature extraction within subsampled or subarray segmentation. Research on next-generation AI vision chips is likely to emphasize hardware complexity to optimize data transfer and computation between the sensor and processor.

To achieve high speed and high resolution in the first computation layer for CNN applications, we propose a real-time analog-domain convolution technique, as presented in Fig. 6. We can use 4-bit or greater weights while maintaining the CIFAR-10 accuracy, as shown in Fig. 7. Therefore, programmable 3 × 3 kernels with 4-bit weights and tunable-resolutions for quantization of the convolution results were implemented for the general CNN model in this design. The convolution and readout were achieved in the proposed convolutional CIS (C^2IS) simultaneously. Furthermore, the required frame buffer, computation energy, and high-bit MAC operations in the following CNN processor could be mitigated.

While showing great potential, there are still challenges in the development of PIS. Due to the required extra circuits for in-sensor processing, the tradeoff between image quality and featuring function needs to be considered. Moreover, the achievable computation complexity of PIS is limited and suitable only for the well-defined application-driven architectures instead of general purpose.

Conclusion

CIM and PIS have shown great potential to increase energy efficiency while raising the resolution of CNNs for the future AI edge devices. Software/hardware co-designs such as adopting novel dataflow control and emerging crossbar-friendly algorithms are essential to the road to success.

Acknowledgment

This work was supported by the Ministry of Science and Technology, Taiwan (Contract MOST 107-2218-E-007-031).

References

[1] Y.-H. Chen, *ISSCC*, 2016. [2] K. Ueyoshi, *ISSCC,* 2018. [3] J. Lee, *ISSCC,* 2019. [4] A. Krizhevsky, *NIPS*, 2012. [5] K. Simonyan, *ICLR*, 2015. [6] K. He, *CVPR*, 2016. [7] A. Biswas, *ISSCC*, 2018. [8] J. Zhang, *JSSC*, 2017. [9] W.-S. Khwa, *ISSCC*, 2018. [10] D. Shin, *ISSCC*, 2017. [11] X. Si, *ISSCC*, 2019. [12] C.-X. Xue, *ISSCC*, 2019. [13] W.-H. Chen, *ISSCC*, 2018. [14] K. Bong, *ISSCC*, 2017. [15] X. Zhong, *VLSI*, 2018. [16] K.-D. Choo, *ISSCC*, 2019. [17] O. Kumagai, *ISSCC*, 2018. [18] R. Mochida, *VLSI*, 2018. [19] A. Nag, *ArXiv*, 2018. [20] T. Yamazaki, *ISSCC*, 2017.

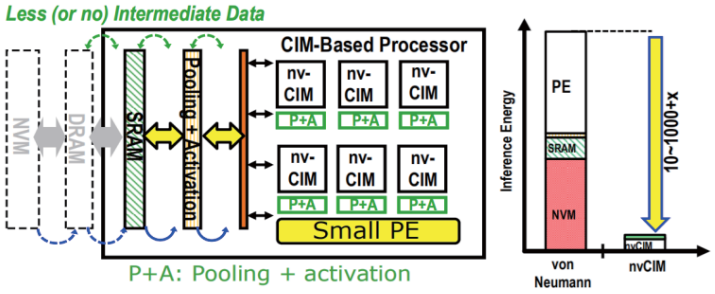

Fig. 2 CIM architecture [13]

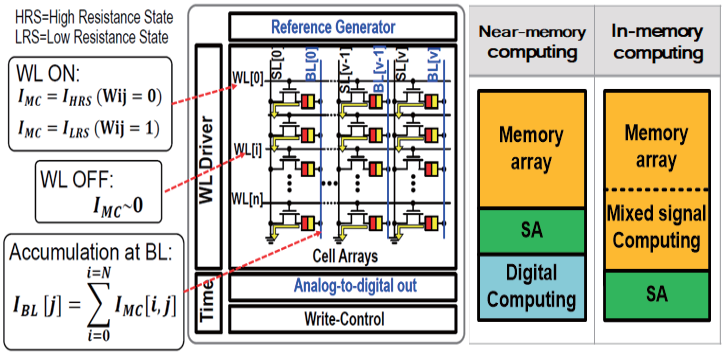

Fig. 3 MAC operation in CIM and the difference between IMC and NMC [12]

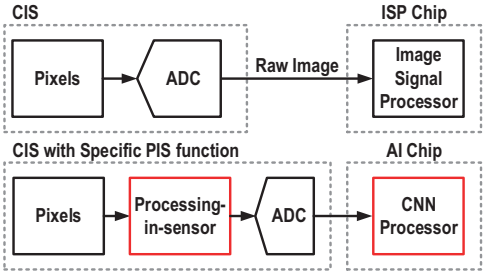

Fig. 5 Conventional image processing flow vs. CIS with specific PIS function for AI applications

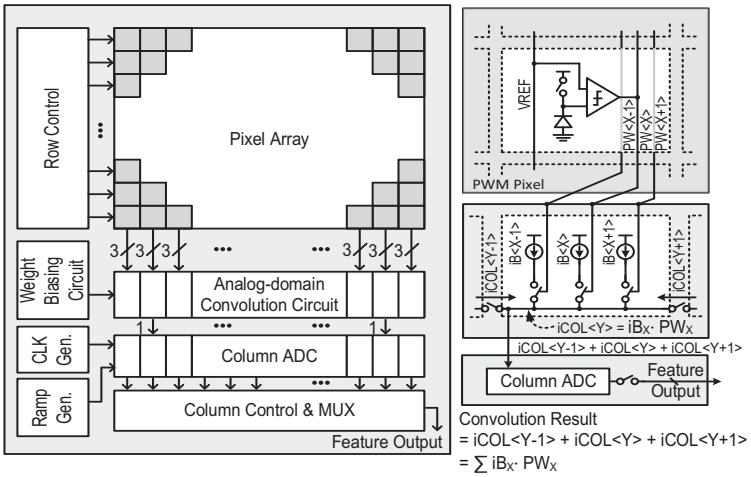

Fig. 6 Architecture of the real-time analog-domain convolution CIS

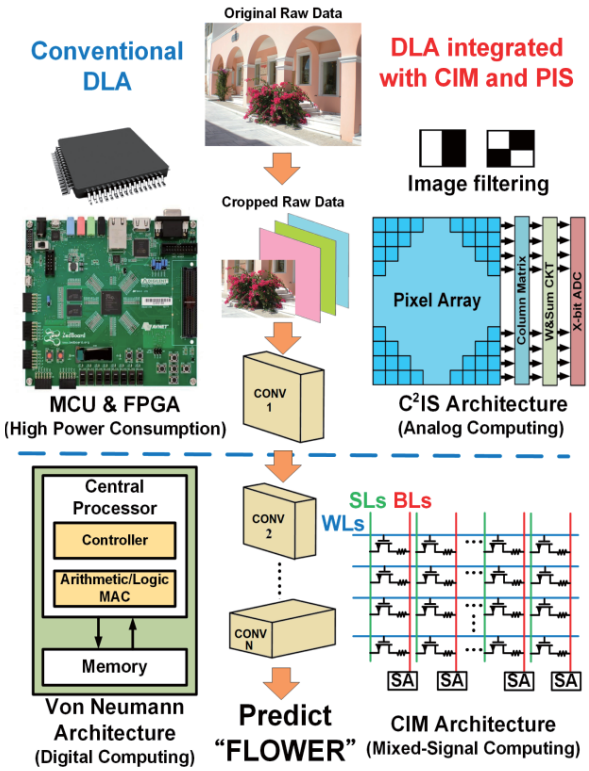

Fig.1 Comparison of conventional DLA with the proposed architecture integrated with CIM and PIS

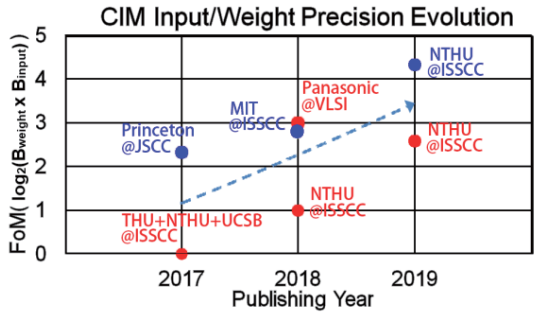

Fig. 4 CIM precision evolution

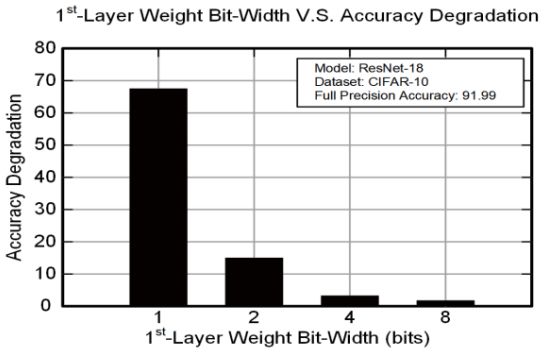

Fig. 7 Analysis of CIFAR-10 accuracy degradation with different weight resolution on ResNet-18