SCIENCE CHINA
Information Sciences

• LETTER •

April 2022, Vol. 65 149403:1–149403:2
https://doi.org/10.1007/s11432-021-3242-6

Self-compensation tensor multiplication unit for adaptive approximate computing in low-power CNN processing

Bo LIU[1†], Zilong ZHANG[1†], Hao CAI[1†], Reyuan ZHANG[1], Zhen WANG[2] & Jun YANG[1*]

[1]*National ASIC System Engineering Center, Southeast University, Nanjing* 210096, *China;*
[2]*Nanjing Prochip Electronic Technology Co. Ltd., Nanjing* 210001, *China*

Received 31 January 2021/Revised 23 March 2021/Accepted 31 March 2021/Published online 11 March 2022

Citation Liu B, Zhang Z L, Cai H, et al. Self-compensation tensor multiplication unit for adaptive approximate computing in low-power CNN processing. Sci China Inf Sci, 2022, 65(4): 149403, https://doi.org/10.1007/s11432-021-3242-6

Dear editor,

Approximate multiplication is an emerging circuit design technique for AI-based IoT devices using deep neural networks which can reduce energy consumption with acceptable accuracy loss. Digital signal processing with low power consumption is very important for battery-powered devices, such as real-time speech recognition [1] and cuffless blood pressure monitoring [2]. We propose a hybrid self-compensation approximate tensor multiplication unit for low-power convolutional neural network (CNN) processing. The proposed architecture encompasses three advantages: (a) a positive-negative hybrid compensation encoding scheme is utilized in partial products generation and the partial products are reorganized; (b) a self-compensation addition tree structure with a staged configuration method is proposed to process the accumulation of reorganized partial products; (c) the proposed unit is further optimized by using a lower supply voltage in the imprecision parts of the addition tree. With the implementation of 22 nm process technology, we evaluated the error metrics and the power reduction. Compared to the state-of-the-art designs, the mean error distance (MED) is reduced by 15.3%–57.1% without reducing the probability of relative error distance smaller than 2% (PRED), while the CNN processing power consumption can be reduced by 19.7%–40.4%.

Architecture of the proposed tensor multiplication unit. The tensor multiplication unit proposed in this study is shown in Figure 1(a). The calculation process is divided into four steps, partial product (PP) generation, PP merge, PP accumulation, and final addition. In the PP generation stage, we use a positive-negative hybrid compensation encoding scheme to reduce error. Half of the input is approximate encoded using the radix-8 approximate booth encode (R8ABE1) method, and the other half is approximate encoded using the R8ABE2 method [3, 4]. The approximate encoding methods using R8ABE1 and R8ABE2 can produce positive and negative errors respectively, and the two methods can compensate each other. PP merge can reduce the additional shift and sign expansion. The multiple partial product arrays obtained from a set of data are rearranged and merged by the same position. PP accumulation uses a self-compensation addition tree structure with a staged configuration method as shown in Figure 1(b). We set the approximate adder structure in each stage of the addition tree with different configurations using the segmented configuration method. The error generated by the approximate adder of the current stage can be compensated by the next stage. The optimal configuration can be obtained by experimental comparisons as shown in Figure 1(b). The full adder (FA), mirror approximate adder (AMA1), and transmission gate-based approximate adder (TGA) are functional with the power supply of VDD (voltage drain-to-drain), while lower bits of low-part OR gate adder (LOA) are functional with a power supply of VDDL (voltage drain-to-drain (low)), which is much lower than the VDD. There is no carry propagation in the adders with LOA, therefore we can use a dual-VDD method to further reduce the power consumption of the addition tree without an increase of the circuit timing delay. The configuration of VDDL can be obtained as shown in Figure 1(c). The dual-VDD method can be implemented by adding an extra power line in the layout of each cell, requiring only an area increase of less than 5% under 22 nm technology.

Evaluation and comparisons. In this study, we use 2×10^5 groups of random input vectors, which is similar to the study [5], to test the proposed tensor multiplication unit. The MED is selected as error metrics [6]. This study is implemented and evaluated under TSMC 22 nm ULL (ultra-low-leakage) process technology. We choose the LOA, AMA1, and TGA2 to build the proposed addition tree. For

* Corresponding author (email: dragon@seu.edu.cn)
† Liu B, Zhang Z L, and Cai H have the same contribution to this work.

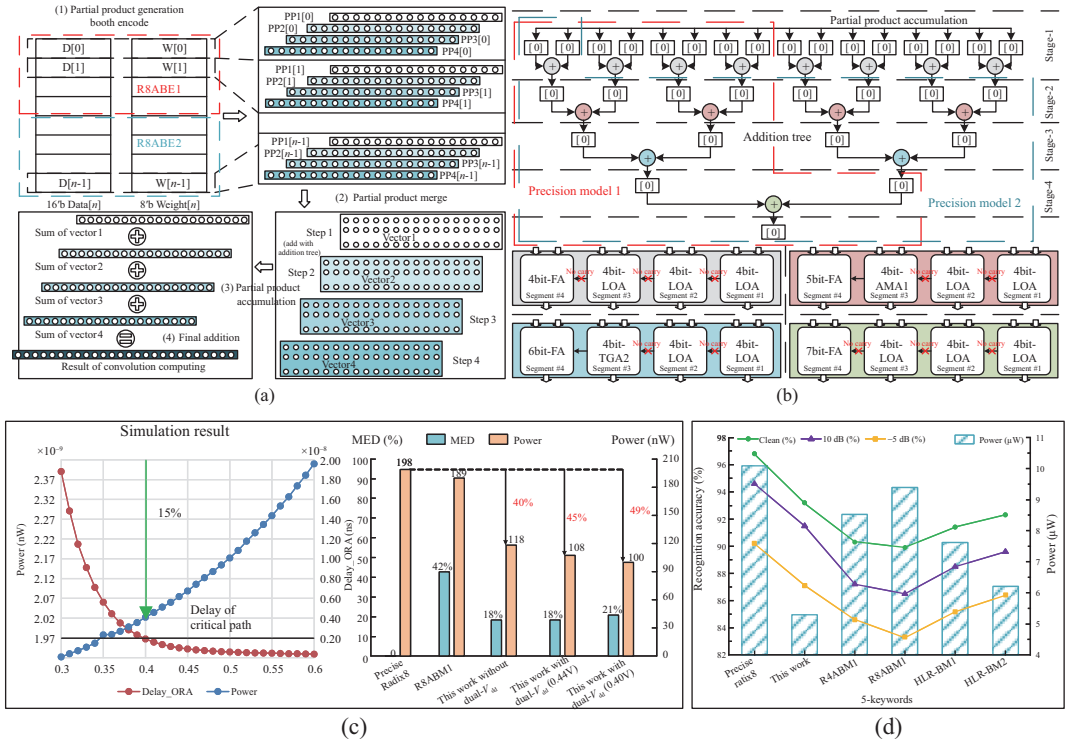

Figure 1 (Color online) (a) Overall architecture of proposed hybrid self-compensation approximate tensor multiplication unit; (b) architecture of the self-compensation approximate addition tree; (c) dual-VDD method utilized to the addition tree; (d) comparisons of power consumption and CNN computing accuracy with different approximate multiplication architectures.

each stage, the low 8-bit LOA is configured as VDDL, and the high-bit adder is configured as VDD of 0.6 V. As shown in Figure 1(c), in order to achieve the maximum power reduction, VDDL is configured as 0.4 V. The comparison results show that with the dual-VDD method, power consumption is reduced by 49% and 15% compared to the precise Radix8 multiplier and the proposed tensor multiplication unit without the dual-VDD method, respectively.

Figure 1(d) shows the comparisons of power consumption and CNN computing accuracy in different signal-noise ratios (SNR) [7] with different approximate multiplication architectures. The CNN adopted consists of 4 convolution layers and 2 fully connected layers with the data/weight bit width of 16/8 bits and is used for speech recognition. The power consumption of this work is 5.3 μW, and the accuracy of the CNN processing for speech recognition is 93.2%. Compared with the standard multiplier of full computing precision, the power consumption is reduced by 40.4%, with an accuracy loss of 3.6%. Compared with the approximate multiplier HLR-BM1 [3], the power consumption is reduced by 19.7%, while the accuracy is increased by 1.8%. Compared to the CNN accelerator proposed for speech recognition in study [1], the CNN accelerator with our proposed tensor multiplication unit can reduce the power consumption from 41.3 to 5.3 μW, while maintaining the real-time speech recognition (response latency < 20 ms, with the clock frequency of 250 kHz).

Conclusion. We propose a hybrid self-compensation tensor multiplication unit to explore the advantages of energy-efficient approximate computing for low-power CNN processing. To make it energy efficient while maintaining the high processing accuracy for CNN processing, the positive-negative hybrid compensation encoding scheme and the self-compensation addition tree structure with the dual-VDD method are utilized to build the proposed tensor multiplication unit. Compared to the state-of-the-art architectures, the proposed tensor multiplication unit can reduce the power consumption of CNN processing by 19.7%, while the accuracy of CNN processing is increased by 1.8%.

Acknowledgements This work was supported by the National Science and Technology Major Project (Grant No. 2018ZX01031101-005) and National Natural Science Foundation of China (Grant No. 61904028).

References

1 Liu B, Wang Z, Zhu W T, et al. An ultra-low power always-on keyword spotting accelerator using quantized convolutional neural network and voltage-domain analog switching network-based approximate computing. IEEE Access, 2019, 7: 186456–186469
2 Zhang Q R, Xie Q S, Duan K F, et al. A digital signal processor (DSP)-based system for embedded continuous-time cuffless blood pressure monitoring using single-channel PPG signal. Sci China Inf Sci, 2020, 63: 149402
3 Waris H, Wang C H, Liu W Q. Hybrid low radix encoding-based approximate booth multipliers. IEEE Trans Circ Syst II, 2020, 67: 3367–3371
4 Boro B, Reddy K M, Kumar Y B N, et al. Approximate radix-8 Booth multiplier for low power and high speed applications. Microelectron J, 2020, 101: 104816
5 Farshchi F, Abrishami M S, Fakhraie S M. New approximate multiplier for low power digital signal processing. In: Proceedings of the 17th CSI International Symposium on Computer Architecture & Digital Systems, 2013
6 Jiang H L, Santiago F J H, Mo H, et al. Approximate arithmetic circuits: a survey, characterization, and recent applications. Proc IEEE, 2020, 108: 2108–2135
7 Varga A, Steeneken H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun, 1993, 12: 247–251