A Test Methodology for Neural Computing Unit

Minho Cheong, Ingeol Lee and Sungho Kang
Electrical and electronic engineering
Yonsei University
Seoul, Korea
{cmh9292, keor}@soc.yonsei.ac.kr, and shkang@yonsei.ac.kr

Abstract— **As convolutional neural networks (CNN) has been widely employed in deep learning applications, the accelerator for CNN has been proposed. Neural computing unit (NCU), which is an accelerator for CNN, includes thousands of identical cores named multiplier and accumulate (MAC), so testing NCU with the conventional methods are inefficient. This paper proposes a novel method to test NCU by applying test patterns for a MAC to all MACs in NCU. The experimental results indicate that the new method reduces test time to 1.38% and test data volume to 0.03%.**

Keywords; neural computing unit (NCU); convolutional neural networks (CNN); identical cores; testing;

I. INTRODUCTION

Recently, applications based on deep learning algorithms have rapidly grew in various field such as image recognition, machine translation and speech recognition. Especially, deep convolutional neural networks (CNN) has been widely employed because it can achieve high accuracy. CNN includes specific computation pattern such as matrix multiplication, so general purpose processors are not efficient for CNN. Thus, various accelerators have been proposed to improve the performance of CNN [1].

An accelerator for CNN named neural computing unit (NCU) includes multiple identical cores named multiplier and accumulate (MAC) [2]. CNN includes thousands of identical cores, so the general test methods are not efficient in testing application time and test data volume. Testing a MAC is simple because it is a simple circuit which performs multiplying, adding and storing, but testing NCU is difficult because it includes thousands of MACs.

This paper proposes an efficient test method for NCU to reduce testing application time and test data volume. By applying the test pattern for a MAC to all MACs in NCU, they can be tested, then the test results can be achieved by comparing the test results of MACs.

II. BACKGROUND

Fig. 1 shows NCU which includes N×N MACs to calculate N×N matrix multiplication. $A_1 - A_N$ represent input signals of one side and $B_1 - B_N$ represent input signals of the other side in NCU. The structure of NCU is a form of systolic array which means that the MACs in the first row and column which receive signal directly from input signals pass signals to the MACs in the second row and column at next clock and so on. As a result,

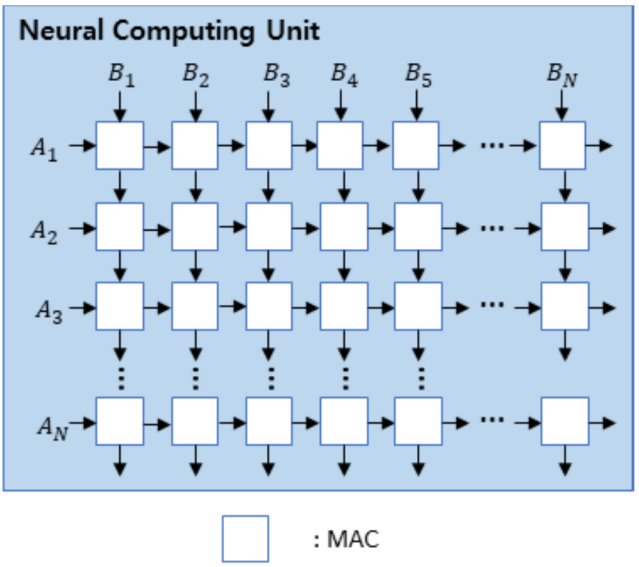

Figure 1. NCU for N×N matrix multiplication

the input signal can be transmitted to the MACs in south-side and east-side. The input A_i and B_i need to be 0 for i-1 clocks because the input of other side reaches to the MAC which receives the input signal directly from A_i or B_i signal at i-1 clocks, so it takes 3N clocks for NCU to end the calculation.

III. PROPOSED TEST METHOD

Fig. 2 shows proposed test method for NCU which contains 4×4 MACs. In the proposed method, the test pattern for only one MAC is applied to all MACs in NCU. The primary inputs of test patterns are assigned to MACs through $A_1 - A_4$ and $B_1 - B_4$. The test patterns assigned to MAC through $A_2 - A_4$ and $B_2 - B_4$ are delayed as the functional mode of NCU. Also, the scan inputs of test patterns are broadcasted to all MACs and the scan inputs of MACs located in the same dotted diagonal line are assigned simultaneously. As a result, the MACs located in the same diagonal line are tested simultaneously where the MACs in different diagonal line are tested gradually.

The outputs of MACs are compared with two other neighboring MACs in the same diagonal line except for the MACs in the first row, the first column, the last row and the last column. If the outputs from one MAC are different with two other MACs, the MAC can be assumed to be defected. The MACs in first row and first column can compared with one other

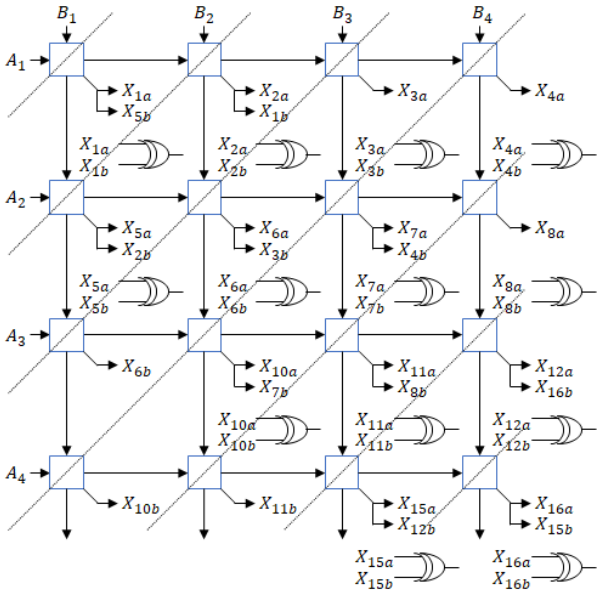

Figure 2. Proposed test structure for NCU

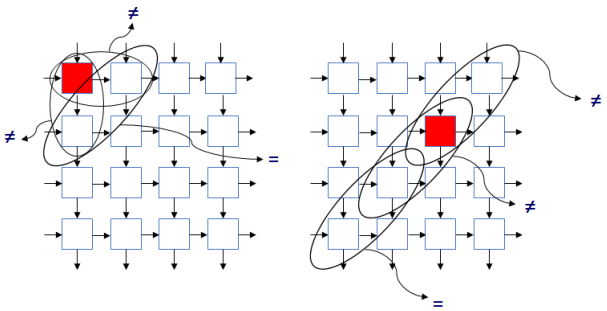

Figure 3. Example of comparison result when a MAC is failed

neighboring MAC in the same diagonal line except for the MAC in the MAC in left upper side and the MAC in right bottom side. The MAC in left upper side is compared with two neighboring MACs in the next diagonal line and the MAC in the right bottom side is compared with the neighboring MACs in the previous diagonal line with timing delay. Those MACs are assumed to be defected if the comparison result between two other MACs are different, too. The other MACs in first row, first column, last row and last column are considered as a defected MAC if the comparison result between itself and neighboring MAC is different, and the comparison result between neighboring MAC and the next MAC is the same.

Fig. 3 shows an example of comparison results when there is a fault in a MAC. When there is a fault in the MAC located at the first row and the first column, the comparison results between neighboring MACs are different while the comparison result with two other MACs is the same. Similarly, when there is a fault in the MAC located not in the first row, the first column, the last row and the last column, the comparison results between two other neighboring MACs are different.

IV. Experimental Result

Table I shows test clock and test data volume of the conventional method and the proposed method. The

TABLE I. Test Clock and Test Data Volume of Conventional Method Solution and Proposed Method

N	Input bit	Conventional method		Proposed method	
		Test clock	*Test data volume*	*Test clock*	*Test data volume*
4	8	4,543	83,074	1,027	3,572
8	8	10,650	410,402	1,035	3,572
16	8	31,510	2,480,322	1,051	3,572
32	8	78,548	12,502,274	1,083	3,572

conventional method assumes that there are $2N - 1$ scan chains in NCU. The test pattern can test stuck-at faults and transition faults. Test clock of the conventional method increases very fast while that of the proposed method increases linearly. This is because the scan chain length of conventional method increases when N increases, while that of proposed method does not increase. Also, the test data volume of the conventional method increases proportional to the square of N while that of the proposed method stays the same. This is because the circuit to be tested of the conventional method becomes larger when N increases, while that of the proposed method is fixed to a MAC. As a result, the test clock of the proposed method is 1.38% of that of conventional method and the test data volume of the proposed method is 0.03% of that of the conventional method.

V. Conclusion

This paper proposes a novel test method for NCU. In the proposed technique, test pattern of only one MAC is used by utilizing the characteristic that the input signal of NCU transmitted to MAC without modifying. The test results of MACs are obtained by comparing with two other MACs which output the same value at the same time, or with the neighboring MACs with some timing delays. As a result, the test clock of proposed method achieves only 1.38% clock and the test data volume shows 0.03% data compared with conventional method.

Acknowledgment

This work was supported by Samsung Electronics Company, Ltd., Hwasung, Korea.

References

[1] S. Ji, W. Xu, M. Yang, and K. Yu. "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, Jan. 2013.

[2] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, J. Cong, "Optimizing FPGA-based accelerator design for deep convolutional neural networks", *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, pp. 161-170, Feb. 2015.

[3] T. T. Hoang, M. Sjalander and P. L. Edefors, "A High-Speed, Energy-Efficient Two-Cycle Multiply-Accumulate (MAC) Architecture and Its Application to a Double-Throughput MAC Unit," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol.57, no.12, pp.3073,3081, Dec. 2010.