2018 IEEE International Conference on
Integrated Circuits, Technologies and Applications

Optimization for Efficient Hardware Implementation of CNN on FPGA

Fasih Ud Din Farrukh
Institute of Microelectronics
Tsinghua University
Beijing 100084, China
fa-s17@mails.tsinghua.edu.cn

Tuo Xie
Institute of Microelectronics
Tsinghua University
Beijing 100084, China
xie-t12@mails.tsinghua.edu.cn

Chun Zhang
Institute of Microelectronics
Tsinghua University
Beijing 100084, China
zhangchun@tsinghua.edu.cn

Zhihua Wang, *Fellow, IEEE*
Institute of Microelectronics
Tsinghua University
Beijing 100084, China
zhihua@tsinghua.edu.cn

***Abstract*—Deep neural networks (DNN) have been a hot research topic in recent years. The key element of DNN is to explore the real time hardware implementation. However, it requires a complete knowledge of hardware where the DNN is going to be implemented. The computational complexity and resource consumption of DNN is increasing by the time. Convolutional Neural Network (CNN) is the popular architecture of DNN especially for image classification. One requires an efficient implementation strategy of CNN to incorporate more computations in real time. Field Programmable Gate Array (FPGA) is considered to be the energy efficient choice for CNN as compared to Graphical Processing Units (GPUs). In this paper, new idea is explored and implemented for basic Processing Element (PE) of CNN. FPGA has limited built-in multiplier accumulator (MAC) units. In this work, MAC units are replaced by Wallace Tree based Multiplier which belongs to the family of *log time array multipliers*. The resources are saved in terms of MAC units and we can implement more processing elements on FPGA.**

Keywords—Convolutional Neural Network, WALLACE Tree Multiplier, FPGA, MAC Unit

I. Introduction

In past years, deep neural networks (DNNs) have achieved a remarkable progress and attention as a strong candidate for a wide range of applications. This is all because of increase in the availability of training set of data and powerful computing resources. However, for real time implementation, DNNs have a choice of GPUs/CPUs and FPGAs. CNN is currently the most popular architecture of DNN for visual recognition tasks. FPGA is becoming an essential part for implementation of CNN accelerator due to its hardware flexibility and it can also provide high performance per unit power. On the other hand, GPUs give us more resources in terms of memory and computational units. However, for inference phase, GPUs have long execution time due to sequential logic design. FPGA is becoming hot choice for hardware designers in inference phase of CNN accelerator design due to its parallel architecture. Still FPGA has limitations due to limited number of hardware resources. Therefore, problem is to find the efficient way of CNN accelerator implementation on FPGA to accommodate more computations. In this paper, main emphasis is to find the alternate solution to implement CNN accelerator on FPGA and overcome the problem of limited hardware resources. Basic unit of CNN accelerator is processing element and new PE is proposed and implemented using a new multiplication strategy by replacing the accumulation process of partial products with WALLACE tree reduction scheme. It belongs to the parallel array multipliers and reduction is performed in parallel.

The rest of the paper is organized as follows. In Section II, the main architecture of new proposed PE based on WALLACE tree is illustrated. Section III is for experimental and simulation results followed by conclusion in Section IV.

II. Proposed PE Based On WALLACE Tree

For CNN accelerator design, there are many potential solutions that can be explored to reduce the hardware implementation cost. In this way, we can accommodate more computational complexity onto real time hardware processing. The essential part of CNN is PE. The basic PE architecture is based on the convolution of input feature map with the shifted window of $K \times K$ kernel to generate one pixel in one output feature map. There could be 90% performance difference between two different approaches of implementation considering the same logic utilization on FPGA [1]. If the CNN accelerator design is going to be implemented on FPGA, it may require hundreds to thousands of hardware MAC units. This is going to increase the computational cost on FPGA because MAC units are limited in numbers on FPGA [2].

In previous work, problem was the limited number of resources available on FPGA in terms of MAC units [3]. This limitation creates a problem not to implement the whole layer convolution once for all. Parallel structure of convolution and PE unit is shown in Fig. 1. The given design is using the idea of loop tiling and implemented a parallel structure with 16 groups of inputs and 16 groups of outputs [3]. This architecture is based on 16-bit fixed point operations and the PE unit requires 16-bit fixed point multiplication and additions for convolution. In each PE unit, calculation order is optimized using three stage pipelines. For convolution, MAC units are required to do the convolution task. However, this comes at the expense of using MAC units. There is a problem with MAC unit that it consumes more area and power. It contains a multiplier and each multiplier costs a large number of logic gates and high power [4]. A new PE is proposed in this paper to replace the MAC unit with WALLACE tree multiplier as shown in Fig. 2 (a).

As discussed, convolution involves the multiplication of input feature map with the kernel. If multiplication is going to be 16-bit then it requires 16-bit MAC unit on FPGA to perform this task. Alternate solution is to replace the MAC unit with efficient multiplier. In literature, lots of architectures have been proposed to perform efficient multiplication. WALLACE tree based multiplier is proposed in this article to replace the MAC unit. Multiplication using MAC unit can be replaced with simple ANDing operation to generate partial products (PPs) and then WALLACE tree can be applied for the reduction of partial products. Wallace tree reduction is an efficient methodology for multiplication and its hardware implementation is relatively easy. WALLACE tree works in a group of three PPs. We can arrange the bits in a group of three bits and add them together in each stage using carry save reduction scheme. Each group reduces its

978-1-5386-6551-0/18/$31.00 ©2018 IEEE

three layers of bits to two layers and this reduction is called (3:2) compressor and these two layers are regrouped to three in next stage. Within each group of three bits, full adder (FA) or (3:2) compressor is applied and half adder (HA) or (2:2) compressor is applied for group of three containing two bits. Depending on the multiplier width, the number of levels is determined consequently. Dot diagram of simple 4-bit WALLACE tree based multiplier is shown in Fig. 2 (b) and last stage consists of simple two input adder and any available fast adder can be used to do addition. Basic elements of WALLACE tree architecture are full adder and half adder. Therefore, FA and HA is designed using simple combinational logic for binary inputs. In this way, it further simplifies the proposed design because no built-in adder is going to be used in this architecture. It is also well known that tree multiplier realizes savings of hardware for larger multipliers and the propagation delay is also reduced [5]. Each adder level will be facing one full adder delay in its critical path. The propagation delay of a multiplier can be calculated using $O(\log_{3/2}(N))$ [5]. The proposed design will be facing six full adder delay because it is a 16-bit multiplier.

III. Experimental Results

In order to test the proposed architecture of PE of CNN accelerator, Virtex-6 of part number XC6VLX130T-2 FPGA of Xilinx family is used. To test the behavior of architecture, ModelSim SE 10.5 is used as a simulator for behavioral testing of proposed PE. Xilinx ISE 14.4 is used for synthesis and Vivado 2012.4 is used for implementation.

The target FPGA has 480 MAC units which comes as DSP48 slices. The already implemented PE unit was using 16 MAC units in number and overall 256 MACs due to limited resources available in terms of DSP slices [3]. This MAC unit, which is used for multiplication of input feature map and kernel, is replaced with proposed architecture for new PE as shown in Fig. 2 (a). Simulation is performed using ModelSim SE 10.5 to test the behavior of both the previously implemented PE and the proposed PE using WALLACE tree based multiplier. The proposed PE design is then synthesized using Xilinx ISE 14.4 and implemented by Vivado 2012.4. In comparison with the previous PE, proposed system does not use MAC units and this architecture is free from the limitations of hardware resources in FPGA. Using the proposed design, we can fully utilize FPGA logic to implement more PE because there are no limitations of MACs. As discussed in previous section, for simplicity full adder and half adder is implemented using simple combinational logic in proposed multiplier for addition of binary inputs within each group of layer of PPs.

New PE gives a synthesis frequency of ~167 MHz, utilization of DSP slice is zero, and it uses 339 slice registers. Fig. 3 (a) shows the RTL schematic of new PE based on WALLACE tree multiplier and it is evident that the implemented design will be facing six full adder delay in its critical path. Fig. 3 (b) is graphical representation of resource utilization summary of two designs and it can be observed that zero DSP slices are used in proposed PE unit design.

IV. Conclusion

In this paper, a new approach is proposed to optimize the basic structure of PE. Instead of using the MAC units for multiplication, WALLACE tree based multiplier is proposed and used for the efficient implementation of multiplication in PE unit of CNN accelerator. Implementation results show that proposed architecture is free from the limitations of hardware resources in FPGA. Since MAC unit is replaced with WALLACE tree based multiplier, more processing elements can be implemented on FPGA and we can accommodate more computations.

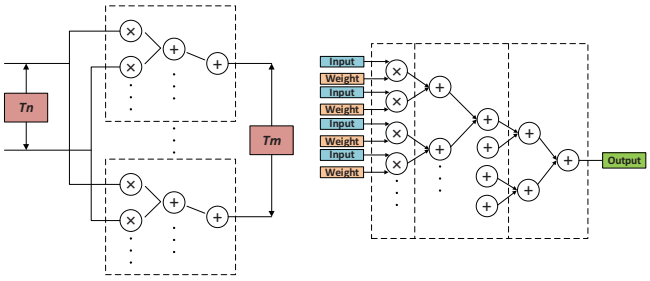

Fig. 1. Parallel structure of Convolution and one PE unit [3].

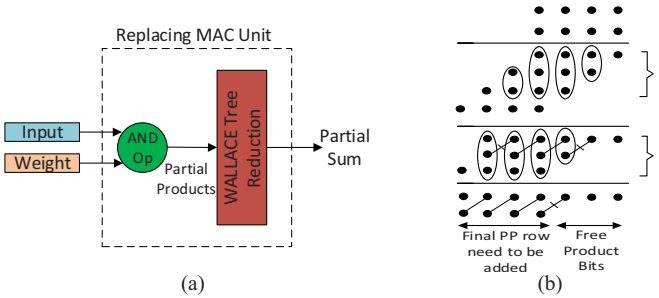

Fig. 2. (a) Proposed WALLACE tree based multiplier for CNN. (b) Dot notation of a WALLACE tree reduction for a 4-bit multiplier.

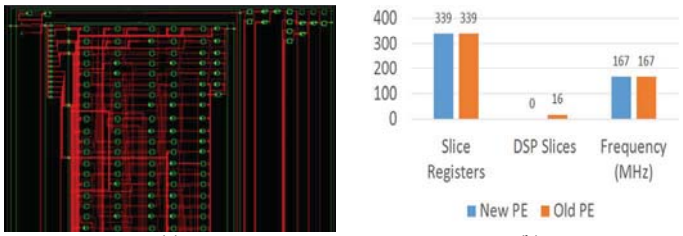

Fig. 3. (a) RTL schematic of WALLACE tree based multiplier. (b) Resource utilization summary of one processing element unit.

Acknowledgment

This work was supported by the National High-tech R&D Program of China (863 Program) No. 2015AA016701.

References

[1] Zhang C, Li P, Sun G, et al, "Optimizing fpga-based accelerator design for deep convolutional neural networks," Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 2015: 161-170.

[2] S. Sabeetha, J. Ajayan, S. Shriram, K. Vivek, and V. Rajesh, "A study of performance comparison of digital multipliers using 22nm strained silicon technology," In 2015 2nd International Conference on Electronics and Communication Systems (ICECS). 180–184.

[3] Wenao Xie, Chun Zhang, Yuanhang Zhang, Chuanbo Hu, Hanjun Jiang, Zhihua Wang, "An Energy-Efficient FPGA-Based Embedded System for CNN Application," In Proceedings of the 14th IEEE International Conference on Electron Devices and Solid State Circuits (EDSSC), China, June 6-8, 2018.

[4] James Garland and David Gregg, "Low Complexity Multiply-Accumulate Units for Convolutional Neural Networks with Weight-Sharing," ACM Transactions on Architecture and Code Optimization, Vol. 15, No. 3, Article 31 (August 2018): 1-24.

[5] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, Digital Integrated Circuits: A Design Perspective, 2nd ed., Prentice Hall Electronics and VLSI Series, Upper Saddle River, NJ: Pearson Education, 2003.