합성곱 신경망 병렬 연산처리를 지원하는 저전력 곱셈 프로세싱 엘리먼트 설계

[1]박은평, [2*]박종수

Low-Power Multiplication Processing Element Hardware to Support Parallel Convolutional Neural Network Processing

[1]Eunpyoung Park, [2*]Jongsu Park

요 약

CNN은 이미지 인식분야에서 높은 성능을 보이지만 반복적인 학습이 진행될 경우 많은 데이터 연산처리로 인한 시스템 자원부족으로 학습 시간이 오래 걸리고 많은 전력을 소모한다는 단점이 있다. 이에 본 논문에서는 합성곱 신경망 연산처리의 핵심 요소인 곱셈 프로세싱 엘리먼트에서 곱셈연산을 수행할 때 발생되는 스위칭 엑티비티를 줄이기 위해 승수와 피승수의 교환율을 늘리는 저전력 부스 곱셈기를 기반으로 하는 프로세싱 엘리먼트를 제안한다. 합성곱 신경망 병렬 연산처리를 지원하는 저전력 곱셈 프로세싱 엘리먼트는 Verilog-HDL을 사용하여 설계되었고, Intel DE1-SoC FPGA Board에 구현하였다. 실험은 성능평가에 대표적으로 MNIST의 숫자 이미지 데이터베이스를 대상으로 기존 제안된 곱셈기의 교환율과 비교하여 성능을 검증하였다.

Abstract

CNNs tend to take a long time to learn and consume a lot of power due to lack of system resources with many data processing units when there are repetitive handles that do not have high performance in the image field. In this paper, we propose a handling method based on a low-power bus that can increase the exchange rate of multipliers and multiplicands within the convolution mixer, which is a tendency activity that occurs when a convolution mixer has multiplication, which is the core element of combination. Convolutional neural networks have proprietary low-power shared processor support and the design was implemented on an Intel DE1-SoC FPGA board using Verilog-HDL. The experiments validated the performance by comparing it with the exchange rate of the multiplier originally proposed by Shen on MNIST's numeric image database.

Keywords: *CNN, Low-power multiplication, Processing element, Multiplier, FPGA*

[1] *목원대학교 지능정보융합학과 석사과정 (eppark1996@naver.com)*
[2*교신저자] *목원대학교 전기전자공학과 조교수 (jspark@mokwon.ac.kr)*

Received: Mar. 29, 2024, Revised: Apr. 25, 2024, Accepted: Apr. 26, 2024

I. 서론

CNN(Convolutional Neural Network)은 시각적 이미지를 분석 및 분류에 있어 높은 인식률을 보여주며, 머신러닝의 한 유형인 딥러닝에서 가장 많이 사용되어지는 알고리즘이다. CNN 은 크게 합성곱층과(convolution layer)와 풀링층(pooling layer)으로 구성되어있고, 합성곱층은 곱셈기를 사용한 합성곱 연산을 통해서 이미지의 특징을 추출하는 역할을 한다. 그러나 반복적인 학습이 진행될 경우 많은 데이터 연산처리로 인한 시스템 자원부족으로 학습 시간이 오래 걸리고 많은 전력을 소모한다는 단점이 있기에, 저전력 곱셈기는 합성곱 신경망 설계 시 고려해야 할 핵심 사항 중 하나이다 [1-3].

CMOS 회로에서, 스위칭 전력 소모에 관한 식은 (1)과 같다.

$$P_{Switching} = aCV_{dd}^2 f_{clk} \quad (1)$$

여기서 a 는 스위칭 액티비티, C 는 부하 캐패시턴스, V_{dd} 는 공급 전압, f_{clk} 는 동작 클럭 주파수를 나타낸다. 수식 (1)을 통해 a(스위칭 액티비티)를 줄임으로써 CMOS 회에서의 스위칭 전력의 소모가 감소됨을 알 수 있다. 즉 곱셈 프로세싱 엘리먼트에서 연산 수행 중 알고리즘 수준에서 스위칭 액티비티를 줄이는 것은 저전력 설계기법을 고려하기 전에 가장 먼저 고려되어야 할 사항이다. 기존의 곱셈 알고리즘을 수정하여 소모전력을 줄이기 위하여 다양한 곱셈 방식이 제안되었고, 그 중 Shen 은 곱셈을 할 경우 승수와 피승수간의 교환율을 증가시켜 전체적인 곱셈연산의 스위칭 액티비티를 감소시킬 수 있는 저전력 곱셈기를 제안하였다 [4]. 또한 이미지 데이터에 대해서 곱셈 연산을 수행할 경우 데이터의 상위 비트 영역에서는 곱셈연산이 불필요한 경우가 많다는 것에 착안한 저전력 곱셈 알고리즘이 연구되었다 [5].

본 논문에서는 Shen 이 제안한 곱셈기보다 승수와 피승수의 데이터를 더 작은 비트를 갖는 여러 개의 곱셈식으로 변화시켰다. 또한 부스 인코딩 결과가 0 이 되는 확률을 더 높인 9 개의 곱셈기를 병렬로 연결하여 기존 합성곱 신경망의 곱셈 프로세싱 엘리먼트보다 전력소모를 줄일 수 있는 곱셈 프로세싱 엘리먼트를 제안한다.

II. 합성곱 연산

합성곱 연산은 그림 1 과 같이 커널(kernel) 또는 필터(filter)라는 N x M 크기의 행렬로 높이 X 너비 크기로 이루어진 이미지를 처음부터 끝까지 겹치며 훑으며 N x M 크기의 겹쳐지는 부분의 각 이미지와 커널의 원소의 값을 곱해서 모두 더한 값을 출력으로 하는 것을 말한다. 커널은 일반적으로 3 x 3 또는 5 x 5 를 사용하며, 그림 1 과 같은 5 x 5 입력으로부터 3 x 3 커널을 사용하여 합성곱 연산을 통해 나온 결과를 특성 맵(feature map)이라고 한다 [6].

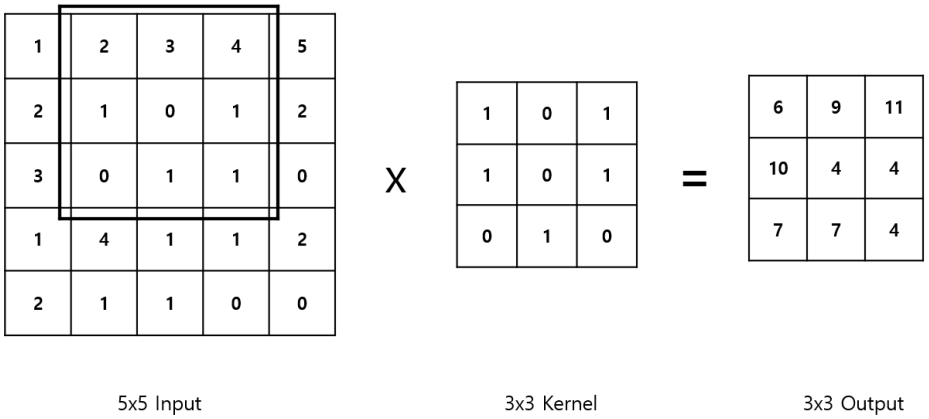

Figure 1. Multiplication of input and kernel
그림 1. 입력과 커널간의 곱셈

III. 곱셈 알고리즘 및 저전력 곱셈기 설계

일반적으로 곱셈 연산은 add-and-shift 의 알고리즘을 적용하여 X 는 피승수, Y 는 승수로 정의하고 수식 (2)를 기반으로 설계된다. 그러나 add-and-shift 알고리즘은 많은 연산 시간이 소요되기 때문에 연산 시간을 단축시키기 위하여 수식 (3)의 Radix-4 modified Booth 알고리즘을 사용한다. Radix-4 modified Booth 알고리즘을 하드웨어로 구현시 그림 2 와 같은 구조를 갖게 된다.

$$P(m+n) = X(m) \cdot Y(n) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} a_i b_j 2^{i+j} \tag{2}$$

$$Y = \sum_{i=0,\, i=even}^{N-1} (y_{i-1} + y_i - 2y_{i+1}) \cdot 2^i = \sum_{i=0,\, i=even}^{N-1} y \cdot 2$$

$$N = even, \quad y = y_{i-1} + y_i - 2y_{i+1} \in \{-2, -1, 0, 1, 2\} \tag{3}$$

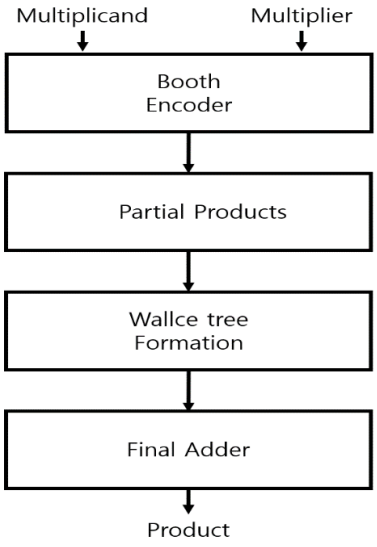

Figure 2. Block diagram of modified Booth multiplier
그림 2. Modified Booth 곱셈기의 블록다이어그램

Shen 의 곱셈 알고리즘은 곱셈을 수행할 때 피승수와 승수간의 교환을 가능하게 함으로써 스위칭 액티비티를 줄였다. 스위칭 액티비티가 줄어들면 곱셈 연산에 필요한 dynamic power 또한 감소한다. 즉 더 작은 활성영역을 갖는 값을 승수로 채택한다면 부스 인코딩 결과로써 더 많은 '0'을 발생시킬 확률이 높아진다. 따라서 Shen 곱셈 알고리즘은 저전력 곱셈기 구현을 위해 두개의 입력 중 더 작은 활성영역을 갖는 입력을 승수로 채택한다.

Shen 곱셈 알고리즘을 발전시켜서 주어진 곱셈식을 더 작은 여러 개의 곱셈식으로 나누는 저전력 부스 곱셈 알고리즘이 제안되었다 [5]. Shen 의 방식은 피승수와 승수의 전체 비트를 기준으로 활성영역의 크기를 비교하므로 실제로 피승수와 승수간 교환이 이루어지는 경우가 많지 않다. 따라서 이러한 피승수와 승수간의 교환율을 증가시키기 위하여, 저전력 부스 곱셈 알고리즘은 승수와 피승수를 더 작은 여러 개의 곱셈식으로 나누었다. 그림 3 과 같이 '10101000 x 10110011'의 곱셈의 경우, Shen 의 곱셈 알고리즘을 이용하면 두 입력 승수가 피승수보다 더 작은 활성영역을 가지고 있기 때문에 두 입력이 교환되지 않는다. 반면 저전력 부스 곱셈기

구조에서는 분리된 4 개의 비트를 갖는 곱셈식을 이용하여 Shen 의 곱셈 알고리즘 입력보다 더 작은 여러 개의 곱셈식의 활성영역을 비교하기 때문에 그림 3 의 상황에서도 두 입력이 교환되는 것을 확인할 수 있다. 즉 저전력 부스 곱셈기 알고리즘은 승수로 채택된 입력의 부스 인코딩 결과가 '0'이 되는 확률을 증가시킴으로써 스위칭 액티비티를 줄여 적은 하드웨어의 추가만으로 곱셈 연산에 필요한 전력 소모를 줄 일 수 있다 [7].

```
10101000 x 10110011 =  1010 x 1011 x 100000000
                     + 1010 x 0001 x 10000
                     +  1000 x 1011 x 10000
                     +  1000 x 0011
```

Figure 3. Proposed multiplication method
그림 3. 제안된 곱셈방식

IV. 저전력 곱셈 프로세싱 엘리먼트

그림 4 는 제안하는 저전력 곱셈 프로세싱 엘리먼트의 구조를 보여준다. 5 x 5 로 이루어진 입력과 3 x 3 커널의 합성곱을 통해 이미지 특징을 추출하기 위하여 두 입력 데이터 각각의 활성영역 크기를 비교하여 교환율을 증가시킴으로써 CMOS 회로에서의 스위칭 액티비티를 낮추는 저전력 부스 알고리즘을 이용한 곱셈기 9 개로 이루어져 있으며, 한 개의 저전력 곱셈기 입력으로는 9 개의 8 비트 고정 소수점 입력과 9 개의 가중치 값이 사용된다. 제안하는 곱셈 프로세싱 엘리먼트의 출력은 16 비트이며 9 개의 곱셈기의 16 비트 결과값을 프로세싱 엘리먼트의 최종 결과값을 구하기 위한 덧셈기의 입력으로 사용하게 된다. 또한 프로세싱 엘리먼트를 구성하는 각각의 저전력 곱셈기들을 병렬로 구성하여 동작시킴으로써 하나의 클럭 주기에 연산이 가능하도록 구현하였다.

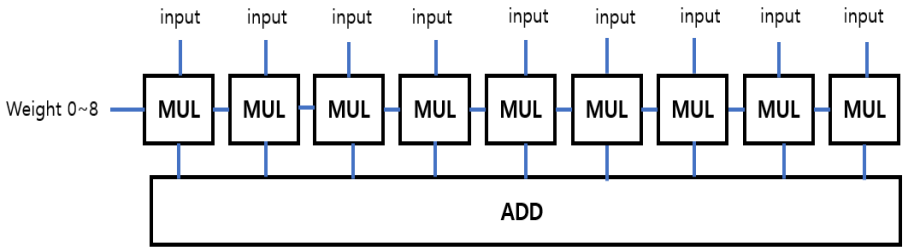

Figure 4. Multiplication processing element
그림 4. 곱셈 프로세싱 엘리먼트

V. 실험 결과

합성곱 신경망 병렬 연산처리를 지원하는 저전력 곱셈 프로세싱 엘리먼트를 구성하는 요소로서 어떠한 곱셈 알고리즘을 적용하는 것이 타당한지 실험하였다. 실험을 위해 Verilog-HDL 을 사용하여 기존의 Radix-4 Booth 곱셈기, Shen 곱셈기, 그리고 제안하는 곱셈 프로세싱 엘리먼트에 사용된 저전력 곱셈기를 각각 설계 및 적용하였다. 그 후 Quartus Prime 18.1 를 사용하여 Intel DE1-SoC FPGA Board 에 각 곱셈기들을 구현하였다. FPGA 구현 후 각각의 곱셈기에 사용되어진 ALMs(Adaptive Logic Modules)를 비교하였을 때, 사용되어진 저전력 곱셈기의 ALMs 개수가 약간 증가된 것을 볼 수 있다. 이는 피승수와 승수를 교환하기 위한 추가적인 하드웨어 로직이 필요하기 때문이다.

Table 1. Multiplication processing element synthesis result
표 1. 곱셈 프로세싱 엘리먼트 합성결과

Applied Multiplier	ALMs Used	Ratio (%)
Conventional Radix-4 Booth Multiplier	1593	100%
Shen's Multiplier [4]	1630	102.32%
Low-power Multiplier [7]	1638	102.82%

그림 5 는 MNIST 데이터베이스의 데이터 셋이다 [8]. 이는 손으로 쓴 숫자들로 이루어진 대형 데이터베이스이며 다양한 화상 처리 시스템을 트레이닝하기 위해 일반적으로 사용된다. 즉 필기 숫자의 분류를 위한 학습 데이터 집합인 MNIST 의 데이터셋 일부를 가지고 실험을 진행하였다. 실험 환경은 표 1 의 실험에 사용된 Shen 곱셈기가 내장된 프로세싱 엘리먼트와 저전력 곱셈기가 내장된 프로세싱 엘리먼트에 각각 MNIST 데이터셋을 적용하여 연산하였다. 각각의 프로세싱 엘리먼트는 Quartus Prime 18.1 을 통해 DE1-SoC FPGA Board 에 탑재하여 실험하였다.

표 2 는 MNIST 데이터셋를 기반으로 연산을 수행했을 때 발생하는 피승수과 승수 간의 교환횟수 및 교환율을 보여준다. Shen 곱셈기를 내장한 프로세싱 엘리먼트는 총 324 회의 곱셈연산에서 14%의 교환율을 보여주는 반면, 저전력 곱셈기를 내장한 제안된 프로세싱 엘리먼트는 32%의 교환율을 보여준다.

Figure 5. MNIST data set
그림 5. MNIST 데이터 셋 [8]

Table 2. Input data exchange rate (Total # of multiplication: 324)
표 2. 입력데이터 교환율 (총 곱셈 횟수: 324)

	Processing Element with Shen's Multiplier	Proposed Processing Element
Exchange Ratio	45	104
Ratio (%)	13.88%	32.09%

VI. 결론

본 논문에서는 부스 곱셈 알고리즘 사용하여 합성곱 신경망 병렬 연산처리를 지원하는 저전력 곱셈 프로세싱 엘리먼트 제안하였다. 전력 소모를 줄이기 위하여 입력된 피승수와 승수 각각의 전체 비트를 기준으로 활성영역의 크기를 비교하는 Shen 의 곱셈기 보다 더 전력소모를 줄이기 위하여 저전력 부스 곱셈기는 하나의 곱셈연산을 작은 비트를 갖는 여러 개의 곱셈연산으로 변화시켰다. 그 결과 피승수와 승수 간의 교환율을 증가시킴으로서 채택된 승수의 부스 인코딩 결과가 '0'이 되는 확률을 더 높여 곱셈연산 시 발생되는 스위칭 액티비티를 감소시켰다. 이러한

저전력 부스 곱셈기 9 개를 병렬로 연결하여 곱셈 프로세싱 엘리먼트를 구현할 수 있었다. 제안한 저전력 곱셈 프로세싱 엘리먼트에 사용되어진 곱셈기는 기존 부스 곱셈기를 적용했을 때 보다 약 2.82% 증가하였고, Shen 곱셈기를 적용했을 때보다 약 0.4% 증가하였다. 이는 피승수와 승수의 교환에 따른 전력 소모 감소를 감안한다면 미미한 증가로 볼 수 있다. 따라서 제안된 저전력 곱셈 엘리먼트는 합성곱 신경망 연산처리를 위한 연산기를 개발하는데 유용하게 사용될 수 있을 것이다.

VII. 참고문헌

[1] D. Perdios, M. Vonlanthen, F. Martinez, M. Arditi and J.-P. Thiran, "CNN-Based Ultrasound Image Reconstruction for Ultrafast Displacement Tracking," IEEE Transactions on Medical Imaging, vol. 40, no. 3, pp. 1078-1089, March 2021.

[2] N. A. S, V. Chaturvedi and M. Shafique, "FRNet: A Feature-Rich CNN Architecture to Defend Against Adversarial Attacks," IEEE Access, vol. 12, pp. 26943-26956, 2024.

[3] G. Gao, Z. Xu, J. Li, J. Yang, T. Zeng and G. -J. Qi, "CTCNet: A CNN-Transformer Cooperation Network for Face Image Super-Resolution," IEEE Transactions on Image Processing, vol. 32, pp. 1978-1991, 2023.

[4] Nan-Ying Shen and O.T.-C. Chen, "Low-power multipliers by minimizing switching activities of partial products," 2002 IEEE International Symposium on Circuits and Systems (ISCAS), Phoenix-Scottsdale, AZ, USA, 2002, pp. IV-IV, doi: 10.1109/ISCAS.2002.1010397.

[5] Chang-Young Han, Hyoung-Joon Park and Lee-Sup Kim, "A low-power array multiplier using separated multiplication technique," IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 48, no. 9, pp. 866-871, Sept. 2001.

[6] Yilun He, "Image Processing System Based on Convolutional Neural Networks," Ph.D. dissertation, Department of Computer Engineering, Paichai Univ., Daejeon, Korea, 2018.

[7] Jongsu Park, Jinsang Kim, and Won-kyung Cho, "Low-Power Multiplier Using Input Data Partition", Journal of Korean Institute of Communications and Information Sciences, vol. 30, no. 11A, pp. 1093-1097, 2005.

[8] Akmaljon Palvanov, and Young Im Cho,"Comparisons of Deep Learning Algorithms for MNIST in Real-Time Environment," International Journal of Fuzzy Logic and Intelligent Systems, vol. 18, no. 2, pp. 126-134, 2018.

저자소개

박은평(*Eunpyoung Park*)

2022 년 2 월 : 목원대학교 IT 융합전자공학과 학사
2022 년 8 월 ~ 현재 : 목원대학교 일반대학원 지능정보융합학과 석사과정

관심분야 : AI 반도체, VLSI 설계, ASIC 설계, 저전력설계, 컴퓨터시스템 등

박종수(*Jongsu Park*)

2017 년 2 월 : 연세대학교 전기전자공학과 박사
2009 년 9 월 ~ 2020 년 2 월 : 삼성전자 SystemLSI 사업부
2020 년 3 월 ~ 현재 : 목원대학교 전기전자공학과 조교수

관심분야 : AI 반도체, 반도체설계, ASIC 설계, 컴퓨터시스템, 멀티미디어통신 등