심층 신경망 가속을 위한 유연한 곱셈 누산 연산기 기반 가변 필터 처리 시스톨릭 어레이 하드웨어

임현서, 신동엽, 권대길, 임용석
한국전자기술연구원 스마트네트워크연구센터

hihi981025@keti.re.kr, dongyeob.shin@keti.re.kr, tgkwon@keti.re.kr, busytom@keti.re.kr

Reconfigurable Multiply Accumulate Unit Based Variable Filter Processing Systolic Array Hardware for Deep Neural Network Acceleration

Hyunseo Lim, Dongyeob Shin, Tai-Gil Kwon, Yong-Seok Lim

Smart Network Research Center
Korea Electronics Technology Institute

요 약

본 논문은 비트-정밀도와 곱셈-누산(Multiply-Accumulate) 연산의 출력 개수를 조절할 수 있는 유연한 곱셈-누산기 및 이를 적용하여 한 번에 처리하는 심층 신경망 필터 개수를 가변 가능한 시스톨릭 어레이(Systolic Array) 하드웨어 구조를 제안한다. 제안하는 구조는 빠르게 많은 필터 처리를 요구하는 레이어에서 기존 구조보다 효율적이다.

Ⅰ. 서 론

심층 신경망(Deep Neural Networks)은 수백만 개 이상의 가중치(Weight) 파라미터 수와 그에 대응되는 입력 데이터와 가중치 값 사이의 반복적인 곱셈-누산(Multiply-Accumulate) 연산을 통해 컴퓨터 비전, 자연어 처리 등 다양한 분야에서 높은 정확도를 보여주고 있다. 하지만 수많은 곱셈-누산 연산으로 인해 하드웨어 구현 시 많은 전력 및 에너지를 소모하는 문제점이 있다. 따라서 곱셈-누산 연산의 하드웨어 에너지 소모를 줄이기 위해, 곱셈-누산 연산의 피연산자(Operand)인 입력 데이터와 가중치의 비트-정밀도를 감소시키되 심층 신경망 정확도는 유지하는 양자화(Quantization) 기법이 많이 사용되고 있다[1]. 예를 들어, 양자화를 통해 입력 데이터와 가중치의 비트-정밀도를 절반씩 줄이면 하나의 단위 곱셈-누산 연산에 소모하는 하드웨어의 전력/에너지/면적 소모를 1/4 로 줄일 수 있다. 정확도를 유지하기 위해서, 양자화를 통한 비트-정밀도는 심층신경망 모델별로 혹은 같은 모델 내에서도 레이어(Layer)별로 최적의 값이 달라질 수 있으며, 최근 관련 연구들이 급증하고 있다[2]. 따라서, 심층 신경망 모델 및 레이어에 따라 비트-정밀도를 가변할 수 있는 가변 비트-정밀도 곱셈-누산기[3] 및 이를 적용한 심층신경망 연산 하드웨어 설계가 중요해지고 있다.

본 논문에서는 비트-정밀도와 출력의 개수를 조절 가능한 곱셈-누산기 및 이를 적용한 심층 신경망 연산용 시스톨릭 어레이(Systolic Array) 하드웨어를 제안한다. 제안하는 하드웨어는 비트-정밀도 가변 시 모드(Mode) 변경을 통해 곱셈-누산 연산의 출력 개수를 가변하여, 처리하는 심층신경망 필터 개수를 조절할 수 있다.

Ⅱ. 본론

2.1. 기존 가변 비트-정밀도 곱셈-누산기 구조

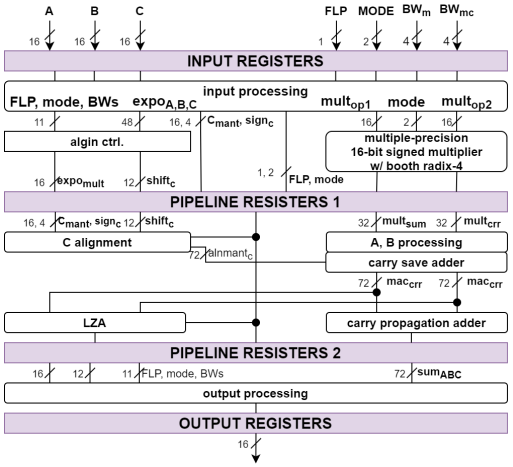

그림 1. 기존 가변 비트-정밀도 곱셈-누산기 구조

그림 1 은 기존의 가변 비트-정밀도 곱셈-누산기 구조이다. A, B, C 를 입력 데이터로 하여 A*B+C 의 곱셈-누산 연산을 수행한다. 이때, A 와 B 의 비트-정밀도에 따라 여러 개의 곱셈을 동시에 처리하는 것이 가능하다. 예를 들어 A, B 가 16bit 일 때 1 개의 곱셈

연산을 처리할 수 있지만, A, B 가 8bit 일 때는 2 개의 곱셈 연산을 동시에 처리하여 A0*B0 + A1*B1 + C 의 연산을 수행할 수 있다. 그림 1 에서 볼 수 있듯이, 기존의 가변 비트-정밀도 곱셈-누산기는 항상 모든 곱셈기 결과를 전부 더해서 하나의 결괏값을 출력한다.

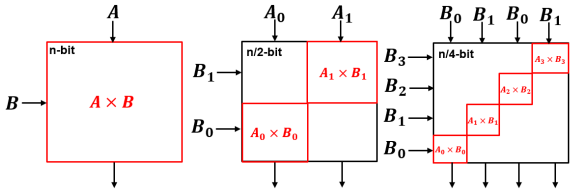

그림 2. 비트-정밀도 변화에 따른 곱셈기 동작

그림 2는 비트-정밀도 변화에 따른 곱셈기의 동작 변화를 나타낸다. 입력 A, B가 n-bit, n/2-bit, n/4-bit인 경우 1, 2, 4개의 곱셈 연산이 동시에 가능하다. 하지만 기존 가변 비트-정밀도 곱셈-누산기는 입력으로 받는 A, B는 가변이 가능하지만 누산 연산의 출력은 고정되어 있다. 대부분의 심층신경망 연산 하드웨어는 곱셈-누산기 여러 개를 시스톨릭 어레이와 같은 배열로 구성하여 처리한다. 곱셈-누산기가 여러 개의 곱셈-누산 연산을 처리하더라도 전부 누산하여 하나의 곱셈-누산 연산 결과만 출력한다면, 전체 연산 측면에서 좀 더 다양한 배열 처리를 할 수 있는 가능성을 제한하게 된다.

2.2. 제안하는 유연한 곱셈-누산기 구조 및 적용

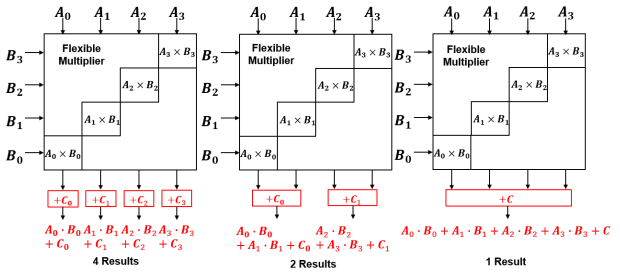

그림 3. 곱셈- 누산 연산 출력 수의 가변이 가능한 누산기 구조

그림 3은 n/4-bit의 곱셈 연산 시, 출력의 가변 형태를 보여준다. 기존에 고정된 개수 만을 출력했던 구조와 달리, 다양한 형태의 누산 결과를 출력할 수 있다. 이러한 구조는 곱셈 결과 누산에 사용되는 누산기의 구조에 간단한 MUX 만을 추가하여 사용 가능하므로, 기존과 하드웨어 면적의 차이가 거의 없다.

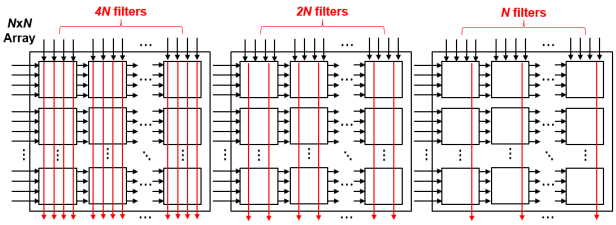

그림 4. 연산 출력의 가변에 따른 필터 처리 수 변화

그림 4 는 본 논문에서 제안한 가변 비트-정밀도 곱셈-누산기를 기본 연산 단위로 하는 시스톨릭 어레이 형태의 하드웨어 가속기이다. 그림 4 와 같이 NxN 배열의 가속기에서 제안하는 곱셈-누산기 1 개가 2 개 혹은 4 개의 곱셈-누산 연산을 처리하는 경우 2N, 4N 개의 필터 연산도 동시에 처리가 가능해진다. 심층 신경망의 레이어 특성에 따라 동시에 처리하는 필터 수가 많을수록 좋은 경우에는 고정된 개수의 필터 연산만을 수행하는 구조보다 더 많은 필터 연산을 동시에 처리할 수 있으므로 효율적이다.

2.3. 구현 결과

표 1 은 본 논문에서 제안하는 곱셈-누산기(MAC)와 제안하는 MAC 에 기반한 8x8 시스톨릭 어레이의 하드웨어 구현 결과이다. 제안하는 하드웨어의 구현을 위해서 시놉시스(Synopsys) 디자인 컴파일러(Design Compiler) 툴과 130nm CMOS 공정 라이브러리를 사용하여 10MHz 동작 주파수에서 합성을 진행하였다. 제안하는 MAC 은 148,305 um^2 의 면적을 가지며, 제안하는 8x8 시스톨릭 어레이는 9,492,205 um^2 의 면적을 갖는다.

	1 MAC	8x8 MAC Array
공정 [nm]	130	130
동작 주파수 [MHz]	10	10
면적 [um^2]	148,305	9,492,205

표 1. 제안한 하드웨어 구조의 합성 결과

Ⅲ. 결론

본 논문에서는 비트-정밀도의 가변과 곱셈-누산 출력의 개수 가변이 가능한 새로운 곱셈-누산기 하드웨어 구조를 제안하였다. 비트-정밀도의 가변을 통해 여러 곱셈 연산을 동시에 처리할 수 있고, 출력의 개수를 조절하여 빠르게 많은 필터 처리를 요구하는 레이어에서 기존 고정된 출력 개수를 처리하는 구조보다 효율적이다. 기존 대비 하드웨어 면적 오버헤드는 5% 미만이며, 130nm 공정, 10MHz 동작 주파수에서 8x8 시스톨릭 어레이로 구현 시 면적은 9,492,205 um^2 이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부의 재원으로 정보통신기획평가원의 정보통신방송기술개발지원사업의 연구결과로 수행되었습니다 (IITP- 2020-0-01077, IoT 다중 인터페이스 기반의 데이터센싱, 엣지컴퓨팅 분석 및 데이터공유 지능형 반도체 기술 개발)

참 고 문 헌

[1] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2704-2713.

[2] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, “HAQ: hardware-aware automated quantization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 8612-8620.

[3] H. Zhang, D. Chen, and S.-B. Ko, ‘‘New flexible multiple-precision multiply-accumulate unit for deep neural network training and inference,’’ IEEE Trans. Comput., vol. 69, no. 1, pp. 26- 38, Jan. 2020.