



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년03월07일
(11) 등록번호 10-2371451
(24) 등록일자 2022년03월02일

(51) 국제특허분류(Int. Cl.)
G06F 7/544 (2017.01) G06F 15/78 (2006.01)
G06F 7/53 (2006.01) G06N 3/063 (2006.01)
(52) CPC특허분류
G06F 7/5443 (2013.01)
G06F 15/786 (2013.01)
(21) 출원번호 10-2021-0068613
(22) 출원일자 2021년05월27일
심사청구일자 2021년05월27일
(56) 선행기술조사문헌
KR1020210002495 A
KR1020190055608 A
KR1020180052063 A
KR1020180034853 A

(73) 특허권자
충남대학교 산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
남병규
대전광역시 서구 도안북로 136 한라비발디 111동 1504호
조수환
충청남도 천안시 서북구 월봉4로 19
(뒷면에 계속)
(74) 대리인
이은철, 김재문

전체 청구항 수 : 총 2 항

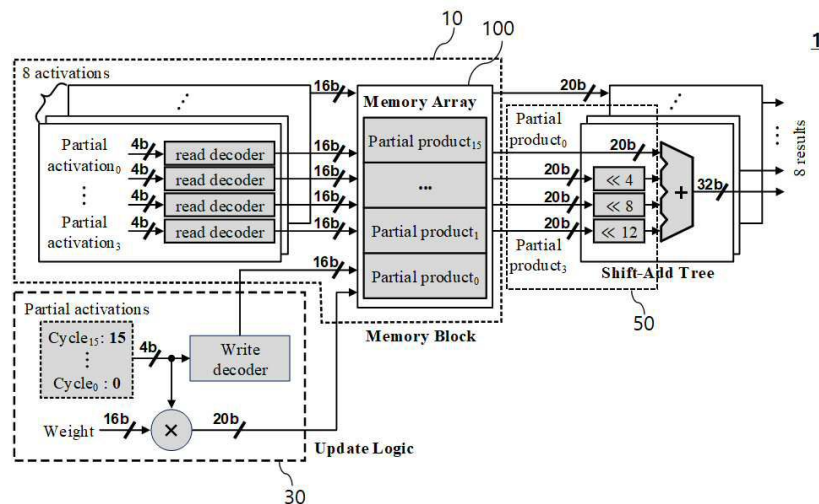
심사관 : 지정훈

(54) 발명의 명칭 멀티포트 메모리를 이용한 병렬 곱셈장치

(57) 요약

본 발명은 병렬 연산 처리가 가능한 곱셈기에 있어서, 매개 변수와 복수의 입력값에 대한 곱셈 결과를 저장하며 하기 멀티포트와 다중 연결하는 단일 메모리; 및 상기 단일 메모리에 저장된 상기 곱셈 결과를 읽어오는 멀티포트를 포함하며, 상기 멀티포트를 구성하는 복수개의 단일 포트가 상기 단일 메모리와 각각 연결됨으로써 저면적 구조로 복수의 곱셈 연산 수행이 가능한 것을 특징으로 한다.

대표도 - 도2



(52) CPC특허분류

G06F 7/53 (2013.01)

G06N 3/063 (2013.01)

최성립

대전광역시 서구 청사로 65 황실타운 108동 1002호

(72) 발명자

도현구

경기도 화성시 동탄하나1길 61

이 발명을 지원한 국가연구개발사업

과제고유번호 1711133806

과제번호 10080568

부처명 과학기술정보통신부

과제관리(전문)기관명 한국산업기술평가관리원

연구사업명 전자정보디바이스산업원천기술개발(R&D, 정보화)

연구과제명 딥러닝 기반 상황인식용 임베디드 GP-GPU기술 개발

기 여 율 1/1

과제수행기관명 서경대학교 산학협력단

연구기간 2021.01.01 ~ 2021.12.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

병렬 연산 처리가 가능한 곱셈기에 있어서,

매개 변수와 복수의 입력값에 대한 곱셈 결과를 저장하며 하기 멀티포트와 다중 연결하는 단일 메모리; 및

상기 단일 메모리에 저장된 곱셈 결과를 읽어오는 멀티포트를 포함하며,

상기 멀티포트를 구성하는 복수개의 단일 포트가 상기 단일 메모리와 각각 연결됨으로써 저면적 구조로 복수의 곱셈 연산 수행이 가능한 것을 특징으로 하는 곱셈기.

청구항 2

제 1 항에 있어서,

상기 단일 메모리에 저장된 곱셈 결과를 갱신하고, 갱신된 곱셈 결과를 상기 단일 메모리로 전송하는 단일 제어부를 더 포함하며,

상기 단일 제어부는 상기 단일 메모리의 개수와 대응되는 개수로 마련되어 단수 개인 것을 특징으로 하는 곱셈기.

발명의 설명

기술 분야

[0001] 본 발명은 컨볼루션 신경망(Convolutional Neural Network, CNN)과 같이 면적 효율적인 설계가 요구되는 응용에서의 곱셈기 부에 대한 면적을 줄이기 위한 방법에 관한 것으로 보다 상세하게는, 다수의 곱셈을 병렬적으로 처리할 시에 곱셈기를 읽기포트로 대체하여 면적 효율을 높이는 것을 특징으로 하는 멀티포트 메모리를 이용한 병렬 곱셈 장치에 관한 것이다.

배경 기술

[0003] 딥러닝 알고리즘 중 CNN은 최고 수준의 정확도로 인해 각광받고 있는 기계학습 알고리즘이며, 근래에는 수십억의 MAC(Multiply-Accumulate) 연산을 통해 사람의 인식 정확도를 뛰어넘는 수준에 이르렀다. 다만, 이때 대량의 MAC 유닛은 차지하는 면적이 넓기 때문에 하나의 칩에 집적시키기 위해서는 면적 측면에서 MAC 유닛에 대한 효율적인 설계가 요구된다. 특히, MAC 유닛 중 곱셈기가 면적의 대부분을 차지하므로, 곱셈기의 면적을 줄이는 연구가 중요하다.

[0004] 한편, CNN은 컨볼루션 연산 시 하나의 가중치(weight)를 다수의 활성화(activation) 값이 공유하므로 가중치에 대한 재사용 빈도가 높다. 이에 종래의 메모리 기반 곱셈기는 재사용 빈도가 높은 복수의 입력값에 대한 곱셈 결과를 LUT(lookup table) 또는 단일 메모리에 저장하고, 저장된 곱셈 결과를 재사용하는 방식을 통해 곱셈기의 전력소모를 낮추고 면적을 줄였다. 그러나, 종래의 메모리 기반 곱셈기 기술은 MAC 유닛에 포함되는 복수의 곱셈기의 수만큼 복수개의 메모리 블록을 요구하므로, 종래의 메모리 기반 곱셈기는 전체적인 회로의 면적을 줄이는 데 한계가 있다. 특히, CNN과 같이 수십억 개의 MAC 연산이 요구되는 응용에서는 더욱 문제가 된다.

[0005] 반면, 본 출원인은 CNN의 가중치와 같이 재사용 빈도가 높은 데이터에 대한 연산 결과를 하나의 단일 메모리에 저장하고, 복수의 곱셈 연산을 수행함에 있어서도 해당 단일 메모리에서 여러 연산 결과값을 읽어오는 연구를 진행하였다. 해당 연구는 읽기 포트(단일포트 또는 멀티포트)가 메모리에 저장된 곱셈 결과를 읽어오는 점에 착안하였고, 이에 곱셈기를 읽기 포트(단일포트 또는 멀티포트)로 대체하여 읽기 포트를 추가하는 것만으로 다수의 곱셈을 병렬적으로 처리할 수 있도록 하였다. 이를 통해 동일한 곱셈 연산이 다수회 반복하여 진행하지 않으며, 곱셈 결과값을 읽어오

는 과정에서도 추가적인 단일 메모리 없이 전술한 하나의 단일 메모리만을 이용함으로써 상기 선행특허보다 복수의 곱셈 연산이 면적 효율적으로 수행될 수 있음을 확인하게 되었다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) 한국공개특허 제10-2021-0002495호
(특허문헌 0002) 한국공개특허 제10-2019-0055608호

발명의 내용

해결하려는 과제

- [0008] 본 발명은 다수의 곱셈기를 면적 효율적으로 설계하여 하나의 칩에 대량의 MAC 유닛을 집적가능하게 하고자 한다.
- [0009] 본 발명이 해결하려는 과제들은 앞에서 언급한 과제들로 제한되지 않는다. 본 발명의 다른 과제 및 장점들은 아래 설명에 의해 더욱 분명하게 이해될 것이다.

과제의 해결 수단

- [0012] 상기 목적을 달성하기 위하여 본 발명은, 병렬 연산 처리가 가능한 곱셈기에 있어서, 매개 변수와 복수의 입력 값에 대한 곱셈 결과를 저장하며 하기 멀티포트와 다중 연결하는 단일 메모리; 및 상기 단일 메모리에 저장된 곱셈 결과를 읽어오는 멀티포트를 포함하며, 상기 멀티포트를 구성하는 복수개의 단일 포트가 상기 단일 메모리와 각각 연결됨으로써 저면적 구조로 복수의 곱셈 연산 수행이 가능한 것을 특징으로 한다.
- [0013] 바람직하게, 상기 단일 메모리에 저장된 곱셈 결과를 갱신하고, 갱신된 곱셈 결과를 상기 단일 메모리로 전송하는 단일 제어부를 더 포함하며, 상기 단일 제어부는 상기 단일 메모리의 개수와 대응되는 개수로 마련되어 단수 개일 수 있다.

발명의 효과

- [0015] 본 발명에 따르면, 다수의 곱셈 연산을 수행할 수 있는 곱셈기에 있어서 단일 메모리에 재사용 빈도가 높은 데이터에 대한 곱셈 결과를 저장하고, 곱셈기를 읽기 포트에 대체하여 읽기 포트를 추가하는 것만으로 다수의 곱셈을 병렬적으로 처리할 수 있어 회로의 면적을 줄일 수 있다는 장점을 갖는다.
- [0016] 또한, 동일한 곱셈 연산 과정을 수행하지 않으므로 전력소비를 감소할 수 있으며, 단일 메모리의 곱셈 결과를 읽어오는데 있어서 멀티포트를 통해 동시에 병렬적으로 처리하므로 효율적으로 연산을 처리할 수 있는 효과가 있다.

도면의 간단한 설명

- [0018] 도 1은 컨볼루션 신경망에서 가중치를 공유하는 컨볼루션 연산이 수행되는 모습을 구조화하여 나타낸 것이다.
- 도 2는 본 발명의 실시 예에 따른 곱셈기의 개략도이다.
- 도 3은 단일 메모리를 공유하는 곱셈 당 면적을 나타낸 그래프이다.
- 도 4는 본 발명의 실시 예에 따른 단일 제어부를 통한 단일 메모리의 데이터가 갱신되는 과정과 멀티포트를 통해 곱셈 결과를 읽어오는 과정을 나타낸 것이다.
- 도 5는 본 발명의 실시예에 따른 단일 메모리(10)의 메모리 어레이가 32 포트 비트 셀(bitcell)의 20×16 행렬로 구성된 모습이다.
- 도 6은 도 5의 32 포트 비트 셀의 개략도이다.

발명을 실시하기 위한 구체적인 내용

- [0019] 이하, 첨부된 도면들에 기재된 내용들을 참조하여 본 발명을 상세히 설명한다. 다만, 본 발명이 예시적 실시 예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일 참조부호는 실질적으로 동일한 기능을 수행하는 부재를 나타낸다.
- [0020] 본 발명의 목적 및 효과는 하기의 설명에 의해서 자연스럽게 이해되거나 보다 분명해질 수 있으며, 하기의 기재만으로 본 발명의 목적 및 효과가 제한되는 것은 아니다. 또한, 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이, 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다.
- [0021] 도 1은 컨볼루션 신경망에서 가중치를 공유하는 컨볼루션 연산이 수행되는 모습을 구조화하여 나타낸 것이다. 도 1과 같이, 필터의 단일 가중치 매개 변수(도 1에서 W_0 내지 W_8)는 컨볼루션 연산이 수행될 때, 여러 활성화 값에 의해 공유될 수 있다. 이에 본 발명에서는 컨볼루션 연산 수행 시, 연산 및 면적 측면에서의 효율적인 설계를 위해 가중치 매개 변수를 재사용할 수 있다.
- [0022] 도 2는 본 발명의 실시 예에 따른 곱셈기(1)의 개략도이다. 곱셈기(1)는 다수의 곱셈 연산을 병렬적으로 처리가 가능하며, 단일 메모리(10), 단일 제어부(30) 및 멀티포트(50)를 포함할 수 있다.
- [0023] 단일 메모리(10)는 매개 변수와 복수의 입력값에 대한 곱셈 결과를 저장할 수 있다. 매개 변수는 재사용성이 높은 데이터로서 가중치(weight) 또는 활성화(activation) 값일 수 있다. 이하에서는 매개 변수가 가중치인 것으로 설명한다. 도 1에서 전술한 바와 같이 곱셈기(1)에서 가중치는 연산 수행 시 공유되므로 연산에 반복적으로 이용된다. 또한, 동일한 입출력 값 및 가중치에 대해 동일한 연산이 반복 수행되는 비경제성을 낮추기 위해 단일 메모리(10)는 다수의 곱셈 결과를 저장할 수 있다.
- [0024] 단일 메모리(10)는 멀티포트(50)와 다중 연결될 수 있다. 단일 포트 기반의 메모리는 동시에 다양한 기능을 수행할 수 없는 반면, 본 발명의 실시예에 따른 단일 메모리(10)는 멀티포트(50)와 다중 연결됨으로써 복수의 곱셈 연산값을 불러올 수 있어 효율을 높일 수 있다. 또한, 단일 메모리(10)는 다수의 곱셈 연산을 처리할 때에도 연산 수 또는 멀티포트(50)를 구성하는 포트의 개수에 관계없이 하나만으로 구성된다.
- [0025] 단일 제어부(30)는 단일 메모리(10)에 저장된 곱셈 결과를 갱신할 수 있다. 단일 메모리(10)는 이전에 수행되었던 연산에 대한 곱셈 결과만을 저장하므로 새로운 입력값 또는 새로운 가중치에 대한 곱셈 결과에 대해서는 단일 제어부(30)를 통해 갱신될 수 있다. 단일 제어부(30)에서 갱신된 곱셈 결과는 다시 단일 메모리(10)로 전송되어 단일 메모리(10)에 저장될 수 있다. 단일 제어부(30)를 통한 단일 메모리(10)의 곱셈 결과 갱신 과정은 이하 도 4에서 후술한다.
- [0026] 단일 제어부(30)는 단일 메모리(10)의 개수에 대응되는 개수로 곱셈기(1)에 마련될 수 있으며, 본 발명은 복수 개가 아닌 하나의 단일 메모리(10)에서만 곱산 연산이 수행되므로 단일 제어부(30) 또한 하나만 마련될 수 있음에 주목한다.
- [0027] 멀티포트(50)는 복수 개의 개별 단일 포트(50)로 구성된 형태일 수 있으며, 멀티포트(50)에 포함된 복수 개의 단일 포트는 모두 단일 메모리(10)와 연결될 수 있다. 멀티포트(50)는 단일 메모리(10)에 연결되어, 단일 메모리(10)에 저장된 곱셈 결과를 읽어 단일 메모리(10)로부터 출력할 수 있다.
- [0028] 멀티포트(50)를 구성하는 단일 포트의 개수는 제한되지 않으며, 단일 메모리(10)로부터 동시에 더 많은 곱셈 결과를 읽어오기 위해 멀티포트(50)에 단일 포트가 추가될 수 있다. 또한, 멀티포트(50)는 단일 포트 상호 간 병렬적으로 배치될 수 있고, 이를 통해 곱셈기(1)는 저면적 구조로 집적되어 복수의 곱셈 연산 수행이 가능할 수 있다.
- [0029] 도 3은 단일 메모리(10)를 공유하는 곱셈 당 면적을 나타낸 그래프이다. 본 발명의 실시예에서는 멀티포트(50)가 단일 메모리(10)에 연결됨에 있어서 단일 메모리(10)를 공유하는 곱셈의 수가 문제될 수 있다. 이에 도 3을 참고하면, 단일 메모리(10)를 공유하는 곱셈 수(도 3의 x축)가 많을수록 곱셈당 유효 면적(도 3의 y축)은 급격히 감소하는 것을 확인할 수 있다. 다만, 이익 또한 곱셈 수가 8개 근처에서 감소하므로, 본 발명의 실시 예에서는 단일 메모리(10)를 공유하는 곱셈의 수를 8개로 설정하여 최대 영역의 효율성을 높일 수 있다.
- [0030] 도 4는 본 발명의 실시 예에 따른 단일 제어부(30)를 통한 단일 메모리(10)의 데이터가 갱신되는 과정과 멀티포트(50)를 통해 곱셈 결과를 읽어오는 과정을 나타낸 것이다. 도 4를 참고하면 갱신 단계 동안 곱셈 결과값이 매 클럭 주기마다 순차적으로 갱신됨을 확인할 수 있다. 곱셈 결과값 갱신 시간은 전체 작업주기에 비해 미비하므로 전체 작업주기에서 단일 제어부(30)의 갱신으로 인한 대기 시간은 영향이 거의 없을 수 있다. 또한 다수의

곱셈기가 공유하는 메모리가 하나이므로 곱셈 결과를 갱신하는 단일 제어부(30)가 하나만 필요하여 보다 면적 효율적으로 설계할 수 있다.

[0031] 도 5는 본 발명의 실시예에 따른 단일 메모리(10)의 메모리 어레이(100)가 32 포트 비트 셀(bitcell)의 20×16 행렬로 구성된 모습이다. 도 5에서 도시한 메모리 어레이(100)는 16개 행의 32개 읽기 워드 라인(RWL; Read Wordline)과 20개 열의 32개 읽기 비트 라인(RBL; Read Bitline)으로 구성된다. 도 6은 도 5의 32 포트 비트 셀의 개략도이다. 도 6을 참고하면, 32 포트 비트 셀은 크게 데이터 저장소(data storage), 쓰기 포트(write port), 그리고 32개의 RBL을 구동하는 32개의 읽기 포트(read port)로 구성된다. 도 5 및 도 6에서의 읽기 포트는 본 발명의 멀티포트(50)를 의미한다. 2단계 F04 버퍼링 구조는 커패시턴스가 큰 32개 읽기 포트를 구동하기 위해 포함될 수 있다. 도 6을 참고하면, 본 발명의 실시예에 따른 읽기 포트는 4개 단위로 묶여 총 8개의 그룹으로 구분될 수 있다.

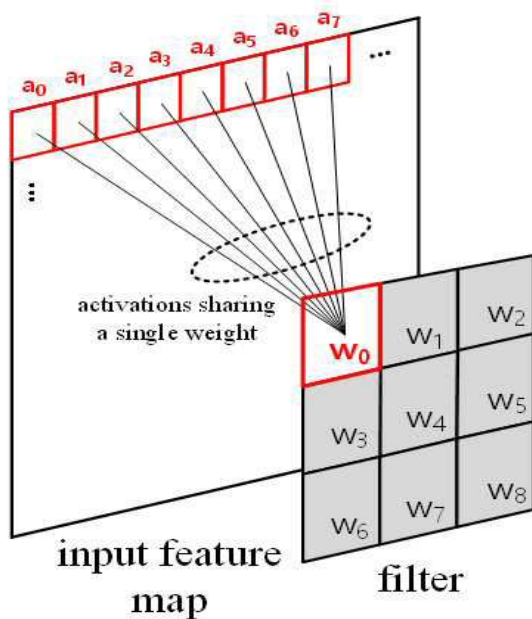
[0032] 이상에서 대표적인 실시예를 통하여 본 발명을 상세하게 설명하였으나, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리 범위는 설명한 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허청구범위뿐만 아니라 특허청구범위와 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태에 의하여 정해져야 한다.

부호의 설명

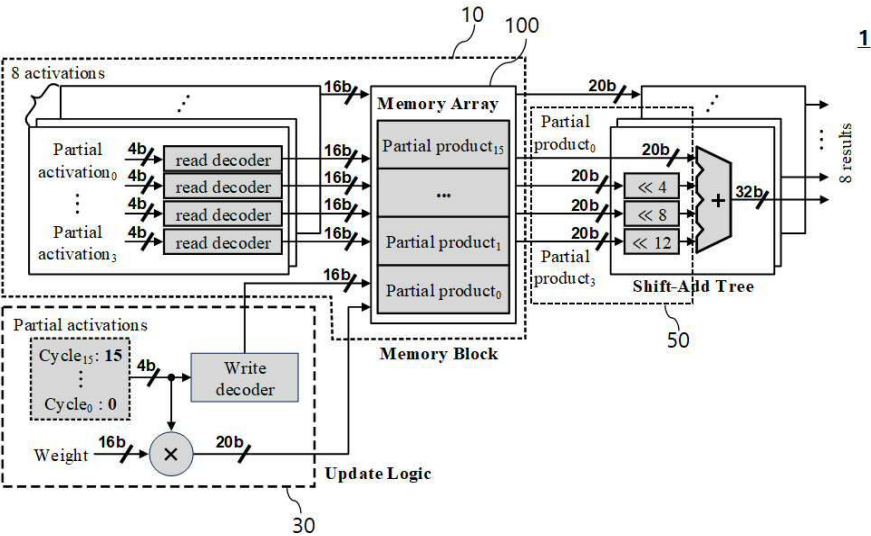
[0035] 1: 곱셈기
10: 단일 메모리
100: 메모리 어레이
30: 단일 제어부
50: 멀티포트

도면

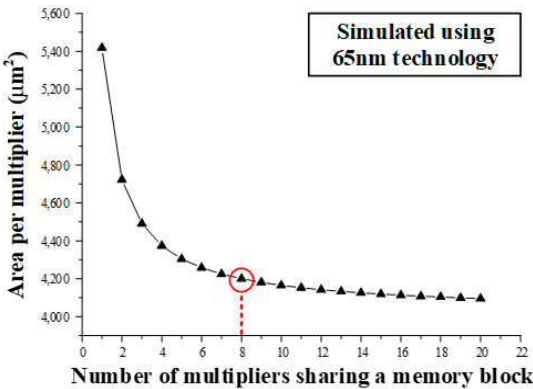
도면1



도면2



도면3



도면4

