



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2022년10월27일

(11) 등록번호 10-2459985

(24) 등록일자 2022년10월24일

(51) 국제특허분류(Int. Cl.)

G06N 3/063 (2006.01) G06F 11/20 (2006.01)

G06F 9/30 (2018.01) G06N 3/04 (2006.01)

(52) CPC특허분류

G06N 3/063 (2013.01)

G06F 11/202 (2013.01)

(21) 출원번호 10-2021-0161147

(22) 출원일자 2021년11월22일

심사청구일자 2021년11월22일

(56) 선행기술조사문헌

KR1020030030624 A*

KR1020190094679 A*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

인하대학교 산학협력단

인천광역시 미추홀구 인하로 100(용현동, 인하대학교)

(72) 발명자

서영교

경기도 부천시 길주로77번길 19-45, 920호(상동, 디아프갤러리4)

최수름

인천광역시 연수구 선학로 100, 5동 1204호(선학동, 금호아파트)

전영준

서울특별시 서대문구 홍제내2길 67, 101동 110호(홍제동, 중앙하이츠아파트)

(74) 대리인

양성보

전체 청구항 수 : 총 6 항

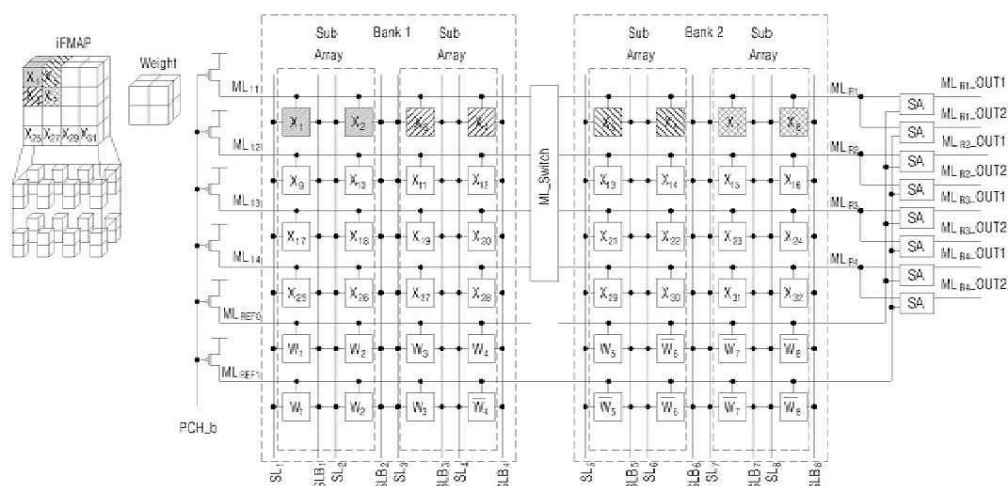
심사관 : 이준상

(54) 발명의 명칭 콘텐츠 주소화 메모리 기반 BNN 가속기의 오류 정정 기법

(57) 요약

콘텐츠 주소화 메모리 기반 BNN 가속기의 오류 정정 기법이 제시된다. 본 발명에서 제안하는 콘텐츠 주소화 메모리 기반 BNN 가속기는 미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 제1 기준 ML(Match Line) 및 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 제2 기준 ML을 포함하고, 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행한다.

대표도



(52) CPC특허분류

G06F 9/30029 (2013.01)

G06N 3/04 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711139247

과제번호 2021-0-02052-001

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 대학 ICT연구센터 육성·지원사업

연구과제명 [Ezbaro][정부] 스마트 모빌리티를 위한 인공지능 시스템반도체 핵심 기술 개발 및
인력 양성 (1차년도)

기 여 율 1/2

과제수행기관명 인하대학교 산학협력단

연구기간 2021.07.01 ~ 2021.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1415172978

과제번호 20009972

부처명 산업통상자원부

과제관리(전문)기관명 한국산업기술평가관리원

연구사업명 차세대지능형반도체기술개발

연구과제명 고효율 초저전력 경량 엣지 디바이스용 소자회로 및 SoC 개발 (2차년도)

기 여 율 1/2

과제수행기관명 한국전자기술연구원

연구기간 2021.01.01 ~ 2021.12.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 제1 기준 ML(Match Line); 및
매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 제2 기준 ML
을 포함하고,
연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행하고,
상기 제1 기준 ML 및 제2 기준 ML은,
입력 특징맵(input feature map; iFMAP) 연결을 만들기 위해 ML 스위치에 의해 연결되고,
하향 필터 슬라이딩의 경우 iFMAP 쌍을 재배열하기 위해 추가 읽기 및 쓰기 작업을 수행하는
BNN 가속기.

청구항 2

삭제

청구항 3

제1항에 있어서,
상기 복수의 추가 CAM 셀은,
상기 복수의 추가 CAM 셀을 포함하는 CAM 어레이에서 XNOR-popcount 연산을 수행함으로써 BNN의 컨볼루션 연산을 추가 디지털 논리 회로 없이 CAM 어레이로 대체 가능하며,
공정 변화로 인한 하드웨어 오류율을 감소시키기 위해 추가 클럭 주기와 함께 SRAM 읽기 작업이 뒤따르는 XNOR-popcount 연산을 수행하는
BNN 가속기.

청구항 4

제3항에 있어서,
상기 복수의 추가 CAM 셀은,
상기 복수의 추가 CAM 셀을 포함하는 CAM 어레이의 병렬 ML 작업을 활용하여 데이터 처리를 활성화하는
BNN 가속기.

청구항 5

제1 기준 ML(Match Line)을 통해 미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 단계;
제2 기준 ML을 통해 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 단계; 및
연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행하는 단계
를 포함하고,
상기 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행하는 단계는,

상기 제1 기준 ML 및 제2 기준 ML가 입력 특징맵(input feature map; iFMAP) 연결을 만들기 위해 ML 스위치에 의해 연결되고, 하향 필터 슬라이딩의 경우 iFMAP 쌍을 재배열하기 위해 추가 읽기 및 쓰기 작업을 수행하는

BNN 가속기의 오류 정정 방법.

청구항 6

삭제

청구항 7

제5항에 있어서,

상기 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행하는 단계는,

상기 복수의 추가 CAM 셀을 포함하는 CAM 어레이에서 XNOR-popcount 연산을 수행함으로써 BNN의 컨볼루션 연산을 추가 디지털 논리 회로 없이 CAM 어레이로 대체 가능하며,

공정 변화로 인한 하드웨어 오류율을 감소시키기 위해 추가 클럭 주기와 함께 SRAM 읽기 작업이 뒤따르는 XNOR-popcount 연산을 수행하는

BNN 가속기의 오류 정정 방법.

청구항 8

제7항에 있어서,

상기 복수의 추가 CAM 셀을 포함하는 CAM 어레이의 병렬 ML 작업을 활용하여 데이터 처리를 활성화하는

BNN 가속기의 오류 정정 방법.

발명의 설명

기술 분야

[0001] 본 발명은 콘텐츠 주소화 메모리 기반 BNN 가속기의 오류 정정 방법 및 장치에 관한 것이다.

배경 기술

[0002] 이진화 신경망(Binarized Neural Network; BNN)은 가중치와 활성화가 데이터의 양을 줄여 하드웨어 측면에서 유리한 두 가지 값으로 제한되기 때문에 경량 신경망에 있어서 많은 관심을 받았다[1], [2]. 하지만, 별도의 처리 장치 및 데이터 저장 요소를 채택하는 기존 컴퓨팅 시스템은 처리 요소와 메모리 간의 데이터 전송으로 인해 여전히 많은 전력 및 성능 오버헤드를 필요로 한다.

[0003] 따라서, 메모리 회로 내에서 직접적으로 동작하는 고성능 및 강력한 곱셈 누적기를 수행하기 위한 새로운 BNN 가속기를 필요로 한다.

선행기술문헌

비특허문헌

[0004] (비특허문헌 0001) [1] Y. H. Chen et al. 2017. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. IEEE JSSC, 52, 1, (2017), 127-138.

(비특허문헌 0002) [2] S. Han et al. 2016. EIE: Efficient Inference Engine on Compressed Deep Neural Network. (2016). arXiv:1602.01528.

(비특허문헌 0003) [3] M. Courbariaux et al. 2016. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1. (2016). arXiv:1602.02830.

(비특허문헌 0004) [4] M. Rastegari et al. 2016. XNOR-Net: ImageNet Classification Using Binary

Convolutional Neural Networks. In 2016 Computer Vision, 525-542.

(비특허문헌 0005) [5] W. Tang et al. 2017. Wang, "How to Train a Compact Binary Neural Network with High Accuracy?," in 2017 AAAI, 2625- 2631.

(비특허문헌 0006) [6] D. Miyashita et al. 2017. A Neuromorphic Chip Optimized for Deep Learning and CMOS Technology With Time-Domain Analog and Digital Mixed-Signal Processing," IEEE JSSC. 52, 10, (2017), 2679-2689

(비특허문헌 0007) [7] K. Pagiamtzis et al. 2006. Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey. IEEE JSSC, (2006). 41, 3, 712-727.

(비특허문헌 0008) [8] W. Choi, K. Jeong, K. Choi, K. Lee, and J. Park, "Content addressable memory based binarized neural network accelerator using time-domain signal processing," in ACM/IEEE Design Automation Conference (DAC), 2018.

(비특허문헌 0009) [9] H. Yonekawa et al. 2017. On-Chip Memory Based Binarized Convolutional Deep Neural Network Applying Batch Normalization Free Technique on an FPGA. in 2017 IPDPSW, 98-105.

발명의 내용

해결하려는 과제

[0005] 본 발명이 이루고자 하는 기술적 과제는 하드웨어 오류 확률을 줄이기 위해 완전 병렬 검색 작업을 활용하여 컨볼루션 작업을 효율적으로 수행하고, 이진화된 데이터를 사용하여 메모리 크기와 컨볼루션 연산의 복잡성을 줄일 수 있는 CAM(Content Addressable Memory) 기반 BNN 가속기 및 그 동작 방법을 제공하는데 있다. 제안하는 CAM 기반 BNN 가속기를 통해 ML(Match Line) 센싱 작동 중 공정 변화로 인한 작동 신뢰성의 저하 및 하드웨어 오류 확률을 줄이고자 한다.

과제의 해결 수단

[0006] 일 측면에 있어서, 본 발명에서 제안하는 콘텐츠 주소화 메모리 기반 BNN 가속기는 미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 제1 기준 ML(Match Line) 및 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 제2 기준 ML을 포함하고, 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행한다.

[0007] 상기 제1 기준 ML 및 제2 기준 ML은 입력 특징맵(input feature map; iFMAP) 연결을 만들기 위해 ML 스위치에 의해 연결되고, 하향 필터 슬라이딩의 경우 iFMAP 쌍을 재배열하기 위해 추가 읽기 및 쓰기 작업을 수행한다.

[0008] 상기 복수의 추가 CAM 셀은 상기 복수의 추가 CAM 셀을 포함하는 CAM 어레이에서 XNOR-popcount 연산을 수행함으로써 BNN의 컨볼루션 연산을 추가 디지털 논리 회로 없이 CAM 어레이로 대체 가능하며, 공정 변화로 인한 하드웨어 오류율을 감소시키기 위해 추가 클럭 주기와 함께 SRAM 읽기 작업이 뒤따르는 XNOR-popcount 연산을 수행한다.

[0009] 상기 복수의 추가 CAM 셀은 상기 복수의 추가 CAM 셀을 포함하는 CAM 어레이의 병렬 ML 작업을 활용하여 데이터 처리를 활성화한다.

[0010] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 콘텐츠 주소화 메모리 기반 BNN 가속기의 오류 정정 방법은 제1 기준 ML(Match Line)을 통해 미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 단계, 제2 기준 ML을 통해 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 단계 및 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행하는 단계를 포함한다.

발명의 효과

[0011] 본 발명의 실시예들에 따른 CAM(Content Addressable Memory) 기반 BNN 가속기 및 그 동작 방법을 통해 하드웨어 오류 확률을 줄이기 위한 완전 병렬 검색 작업을 활용하여 컨볼루션 작업을 효율적으로 수행하고, 이진화된

데이터를 사용하여 메모리 크기와 컨볼루션 연산의 복잡성을 줄일 수 있다. 제안하는 CAM 기반 BNN 가속기를 통해 ML(Match Line) 센싱 작동 중 공정 변화로 인한 작동 신뢰성의 저하 및 하드웨어 오류 확률을 줄일 수 있다.

도면의 간단한 설명

- [0012] 도 1은 본 발명의 일 실시예에 따른 BNN 연산 진리표를 나타내는 도면이다.
- 도 2는 본 발명의 일 실시예에 따른 BNN 활성화 및 XNOR-popcount를 설명하기 위한 도면이다.
- 도 3은 본 발명의 일 실시예에 따른 CAM 메모리 기반 XNOR 연산을 설명하기 위한 도면이다.
- 도 4는 본 발명의 일 실시예에 따른 CAM 메모리 기반 BNN 가속기의 회로를 나타내는 도면이다.
- 도 5는 본 발명의 일 실시예에 따른 ML 스위치를 사용한 ML 연결을 설명하기 위한 도면이다.
- 도 6은 본 발명의 일 실시예에 따른 다중 기준 센싱 방법을 설명하기 위한 도면이다.
- 도 7은 본 발명의 일 실시예에 따른 CAM 메모리 기반 BNN 가속기의 오류 정정 방법을 설명하기 위한 흐름도이다.
- 도 8은 본 발명의 일 실시예에 따른 CAM 메모리 기반 BNN 가속기와 종래기술의 오류 정정 결과를 비교한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0013] 이진화 신경망(Binarized Neural Network; BNN)은 저비용 컨볼루션 연산을 위한 가장 효율적인 신경망 중 하나이다. 이러한 BNN에서는 이진화된 데이터를 사용하여 메모리 크기와 컨볼루션 연산의 복잡성을 줄일 수 있다.
- [0014] CAM(Content Addressable Memory) 기반 BNN 가속기는 CAM에서 완전 병렬 검색 작업을 활용하여 컨볼루션 작업을 효율적으로 수행할 수 있다. 하지만, CAM 기반 BNN 하드웨어의 중요한 문제 중 하나는 ML(Match Line) 센싱 작동 중 공정 변화로 인해 작동 신뢰성이 심각하게 저하된다는 것이다.
- [0015] 따라서, 본 발명에서는 하드웨어 오류 확률을 줄일 수 있는 새로운 CAM 어레이 설계를 제안한다. 제안하는 CAM 기반 가속기는 XNOR-popcount 연산에서 62% 감소를 달성하고 Fashion-MNIST 데이터 셋의 분류 정확도 하락은 2.33%에서 1.26%로 감소한다. 이하, 본 발명의 실시 예를 첨부된 도면을 기준으로 상세하게 설명한다.
- [0017] 도 1은 본 발명의 일 실시예에 따른 BNN 연산 진리표를 나타내는 도면이다.
- [0018] 도 1(a)는 +1과 -1의 이진화된 비트를 기반으로 한 BNN의 곱셈 연산의 진리표를 나타낸다[3]. 하지만, 단순화된 BNN 하드웨어 구현의 경우 -1을 사용하는 대신 1 또는 0의 이진화된 비트를 처리하는 것을 선호한다[4], [5]. -1을 0으로 변환하면 곱셈 연산을 위한 진리표가 도 1(b)와 같이 바뀌고 XNOR 게이트도 같은 연산을 수행할 수 있다.
- [0020] 도 2는 본 발명의 일 실시예에 따른 BNN 활성화 및 XNOR-popcount를 설명하기 위한 도면이다.
- [0021] 도 2(a)의 BNN에서 post-accumulate signum 함수는 도 2(b)와 같이 $n/2$ (누적 크기의 절반) 임계값의 비교기를 사용하여 비트 카운트 연산에 매핑될 수 있다.
- [0023] 도 3은 본 발명의 일 실시예에 따른 CAM 메모리 기반 XNOR 연산을 설명하기 위한 도면이다.
- [0024] CAM은 입력된 검색 데이터와 저장된 데이터를 비교하여 일치하는 위치의 주소를 반환한다. 도 3은 6T SRAM과 양쪽에 2-스택 NMOS 트랜지스터로 구성된 CAM 셀을 나타낸다[6], [7].
- [0025] CAM 어레이의 ML 센싱 작동은 다음과 같이 설명할 수 있다. ML 사전 충전 단계 이전에 SL(Search-Line) 쌍은 의도하지 않은 ML 방전을 방지하기 위해 접지로 비활성화된다. ML 사전 충전 단계에서 어레이의 모든 ML은 전압을 공급하기 위해 사전 충전된다. SL 쌍이 검색 데이터로 설정되면 2-스택 NMOS 트랜지스터가 검색 데이터와 저장된 데이터를 비교한다.
- [0026] 도 3의 진리표와 같이 검색 데이터가 저장된 데이터와 일치하지 않을 경우 풀다운 경로가 형성되어 ML 노드를 방전시킨다. 반면, 검색 데이터와 저장된 데이터가 일치하지 않으면 두 개의 적층된 NMOS 트랜지스터가 꺼지고 ML에서 접지까지 전류 경로는 없어진다[8].
- [0028] 도 4는 본 발명의 일 실시예에 따른 CAM 메모리 기반 BNN 가속기의 회로를 나타내는 도면이다.

- [0029] 제안하는 CAM 메모리 기반 BNN 가속기는 제1 기준 ML(Match Line) (ML_{REF0}) 및 제2 기준 ML(ML_{REF1})을 포함한다.
- [0030] 제1 기준 ML은 미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교한다. 제2 기준 ML은 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교한다. 이후, 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행할 수 있다.
- [0031] 본 발명의 실시예에 따른 BNN MAC(Multiply Accumulate) 연산에서 입력 X는 CAM 셀의 저장 노드에 매핑되고 검색 라인 노드는 가중치(Weight) 입력에 의해 결정된다.
- [0032] 다수결 원칙을 고려하여 평가된 ML과 기준 ML의 방전 지연을 비교하여 XNOR-popcount 연산을 수행할 수 있다 [4], [8]. 도 4와 같이 필터 가중치가 SL에 연결되고 입력 특징 맵(iFMAP)이 CAM 배열에 저장되어 XNOR-popcount 연산에 사용할 수 있다.
- [0033] 예를 들어, XNOR 연산은 CAM 셀에서 iFMAP과 가중치를 비교하여 수행된다. ML 방전 지연은 미스매치(mismatch) CAM 셀의 수에 따라 달라지므로 센싱 증폭기를 통해 ML의 전압과 기준 ML을 비교하여 주로 '1' 인지 주로 '0' 인지 알 수 있다. बैंक 1(Bank 1)과 बैंक 2(Band 2) 사이에 위치한 ML 스위치(ML_{Switch})는 필터 슬라이딩에 사용된다.
- [0035] 도 5는 본 발명의 일 실시예에 따른 ML 스위치를 사용한 ML 연결을 설명하기 위한 도면이다.
- [0036] बैंक 1(Bank 1)과 बैंक 2(Band 2) 사이에 위치한 ML 스위치는 필터 슬라이딩에 사용된다. 먼저 도 5(a)와 같이 XNOR-popcount 연산을 4개의 ML에서 동시에 수행한다. 다음 오른쪽 필터 슬라이딩에서 ML 연결은 필요한 iFMAP 연결을 만들기 위해 ML 스위치에 의해 재구성되어야 한다. 그렇지 않으면 하향 필터 슬라이딩의 경우 적절한 iFMAP 쌍을 재배열하기 위해 추가 읽기 및 쓰기 작업이 필요하다. 수정된 CAM 어레이에서 XNOR-popcount 연산을 수행함으로써 BNN의 강력한 컨볼루션 연산을 추가 디지털 논리 회로 없이 CAM 어레이로 대체할 수 있다. 또한 CAM 어레이의 병렬 ML 작업을 활용하여 CAM 기반 BNN 가속기에서 높은 데이터 처리량을 활성화할 수 있다.
- [0037] 하지만, CAM 기반 BNN 가속기[8]의 주요 단점은 XNOR-popcount 연산의 신뢰성이 공정 변화에 의해 심하게 영향을 받아 큰 센싱 오류를 초래한다는 것이다. 이것은 작은 미스매치 차이의 경우 XNOR-popcount 연산의 연산 신뢰성의 향상을 필요로 한다.
- [0038] 본 발명에서는 추가적인 로직 기반 XNOR-popcount 연산을 선택적으로 수행하여 연산 오류를 줄일 수 있는 새로운 설계 기법을 제안한다. XNOR-popcount 연산 오류는 평가된 ML과 기준 ML이 작은 경우에 주로 발생한다는 사실에 근거한다. 이는 미스매치 차이가 작을 때 ML 방전 속도가 가까울수록(close) ML 센싱 신뢰도가 공정 변화에 영향을 받지 않기 때문이다.
- [0039] 본 발명의 실시예에 따르면, ML 센싱 신뢰도 문제를 해결하기 위해 도 4와 같이 다중 기준을 사용하는 새로운 센싱 방식을 제안한다. 두 개의 개별 기준은 다음과 같이 결정된다. 하나의 기준 ML($REF0$)에서, 미스매치된 3개의 추가 CAM 셀은 기준 단일 기준 방식과 비교되어 방전 속도가 빨라진다. 반면 다른 기준 ML($REF1$)은 3개의 추가 CAM 셀이 매칭되어, 기준 기준 방식보다 방전 속도가 느려진다.
- [0041] 도 6은 본 발명의 일 실시예에 따른 다중 기준 센싱 방법을 설명하기 위한 도면이다.
- [0042] 기준 ML($REF0$) 및 기준 ML($REF1$)은 각각 5비트 미스매치 및 3비트 미스매치의 경우이다. 미스매치 입력이 없는 경우에는 ML이 방전되지 않아 평가된 ML 전압이 기준 ML($REF0$)과 기준 ML($REF1$)의 전압보다 높게 된다. 따라서, XNOR-popcount 연산의 오류 가능성은 거의 0 이다.
- [0043] 마찬가지로 모든 미스매치 입력의 평가된 ML은 기준 ML($REF0$) 및 기준 ML($REF1$)보다 훨씬 빠르게 방전되므로 이 경우 XNOR-popcount 연산에서 높은 신뢰성을 가질 수 있다. 반면, 4비트 미스매치(4bit mismatch)의 경우 미스매치 차이가 작기 때문에 XNOR-popcount 연산 신뢰도가 낮다. ML 방전 속도가 두 기준 사이에 위치하는 경우 높은 분류 정확도를 달성하려면 수동으로 디지털 논리 기반 XNOR-popcount 연산을 수행하는 것이 바람직하다.
- [0044] 따라서, 본 발명에서 제안하는 기법은 공정 변화로 인한 하드웨어 오류율을 완화하기 위해 추가 클럭 주기와 함께 SRAM 읽기 작업이 뒤따르는 디지털 논리 기반 XNOR-popcount 연산을 수행한다.
- [0046] 도 7은 본 발명의 일 실시예에 따른 CAM 메모리 기반 BNN 가속기의 오류 정정 방법을 설명하기 위한 흐름도이다.
- [0047] 제안하는 CAM 메모리 기반 BNN 가속기의 오류 정정 방법은 제1 기준 ML(Match Line)을 통해 미스매치되는 복수

의 추가 CAM 셀을 단일 기준 ML과 비교하는 단계(710), 제2 기준 ML을 통해 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교하는 단계(720) 및 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행하는 단계(730)를 포함한다.

[0048] 단계(710)에서, 제1 기준 ML(Match Line)을 통해 미스매치되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교한다.

[0049] 단계(720)에서, 제2 기준 ML을 통해 매칭되는 복수의 추가 CAM 셀을 단일 기준 ML과 비교한다.

[0050] 단계(730)에서, 연산 오류를 감소시키기 위하여 상기 복수의 추가 CAM 셀, 제1 기준 ML 및 제2 기준 ML을 통해 XNOR-popcount 연산을 선택적으로 수행한다.

[0051] 본 발명의 실시예에 따르면, 제1 기준 ML 및 제2 기준 ML가 입력 특징맵(input feature map; iFMAP) 연결을 만들기 위해 ML 스위치에 의해 연결되고, 하향 필터 슬라이딩의 경우 iFMAP 쌍을 재배열하기 위해 추가 읽기 및 쓰기 작업을 수행할 수 있다.

[0052] 본 발명의 실시예에 따른 복수의 추가 CAM 셀을 포함하는 CAM 어레이에서 XNOR-popcount 연산을 수행함으로써 BNN의 컨볼루션 연산을 추가 디지털 논리 회로 없이 CAM 어레이로 대체 가능하다. 본 발명에서는 공정 변화로 인한 하드웨어 오류율을 감소시키기 위해 추가 클럭 주기와 함께 SRAM 읽기 작업이 뒤따르는 XNOR-popcount 연산을 수행할 수 있다. 또한, 복수의 추가 CAM 셀을 포함하는 CAM 어레이의 병렬 ML 작업을 활용하여 데이터 처리를 활성화할 수 있다.

[0054] 도 8은 본 발명의 일 실시예에 따른 CAM 메모리 기반 BNN 가속기와 종래기술의 오류 정정 결과를 비교한 도면이다.

[0055] 본 발명의 실시예에 따른 CAM 기반 BNN 가속기의 효율성을 검증하기 위해 상용 28nm CMOS 기술을 사용하여 LeNet-5 모델의 두 번째 컨볼루션 레이어를 구현하였다. HSPICE 시뮬레이션을 수행하기 위해 5개의 뱅크(ML당 150셀)가 CAM 어레이 아키텍처에 구현되었다.

[0056] 도 8(a)는 기준 ML과 평가된 ML의 미스매치 수의 차이(Mismatch Difference)와 단일 기준 기반 CAM 어레이에서 XNOR-popcount 연산의 연산 실패율(Error Probability) 간의 관계를 보여준다.

[0057] 상술된 바와 같이 미스매치 차이의 절대값이 증가하면 연산 실패율이 감소한다. 제안하는 CAM 어레이(Prop.)는 동작 신뢰성에 취약한 경우에 디지털 논리 회로 기반 XNOR-popcount 연산을 수행하기 때문에 도 8(b)와 같이 단일 기준 방식(1Ref.)보다 훨씬 낮은 동작 실패율을 보인다.

[0058] 표 1은 BNN 가속기의 오류 확률 및 분류 정확도를 비교한 표이다.

[0059] <표 1>

	<i>XNOR-popcount error probability</i>	<i>Fashion MNIST classification accuracy</i>
TOP-1 Accuracy	-	84.37%
1Ref.	7.51%	82.04%
Prop.	2.88%	83.11%

[0060]

[0061] 표 1과 같이 원래 XNOR-popcount 결과에 ML 센싱 오류 확률을 적용할 때 Fashion MNIST 테스트 이미지에 대한 전체 출력 활성화의 7.51%가 반전되어 LeNet-5에서 상위 1개 정확도의 2.33%가 저하된다. 반면에 제안된 메모리 아키텍처는 XNOR-popcount 연산 오류의 2.88%만 달성할 수 있으며 이는 하드웨어 오류에서 62% 감소를 나타낸다. 이것은 상위 1 정확도와 비교하여 1.26%의 분류 정확도 감소로 이어진다. 또한 제안된 BNN 가속기는 [9]에 표시된 기준 XNOR-popcount 연산에 비해 입력 이미지에 대해 51.72% 감소된 연산 주기를 달성할 수 있다.

[0062] 이와 같이, 본 발명에서는 빠르고 강력한 CAM 기반 BNN 가속기를 위한 새로운 신뢰성 향상 기법을 제안한다. BNN의 거대한 컨볼루션 연산은 수정된 CAM 배열에서 XNOR-popcount 연산으로 대체될 수 있다. 병렬 컨볼루션 연산은 CAM 어레이의 완전 병렬 검색으로 수행할 수 있어 연산 성능이 50% 이상 향상된다. 또한, 다중 기준이 있는 제안된 센싱 방식은 XNOR-popcount 오류 확률 문제를 완화하여 Fashion MNIST 분류 정확도를 1% 이상 향상시킬 수 있다.

[0064] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어

구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0065] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

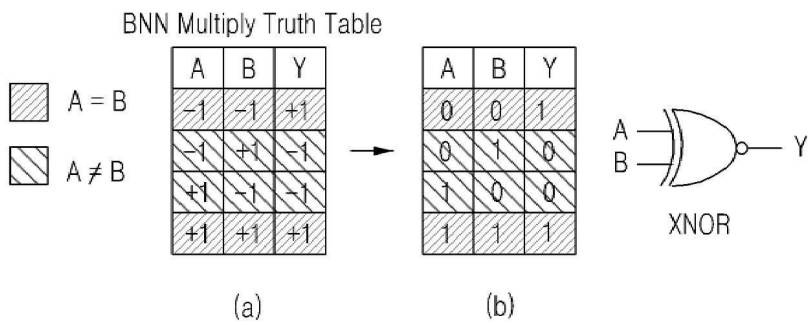
[0066] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0067] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

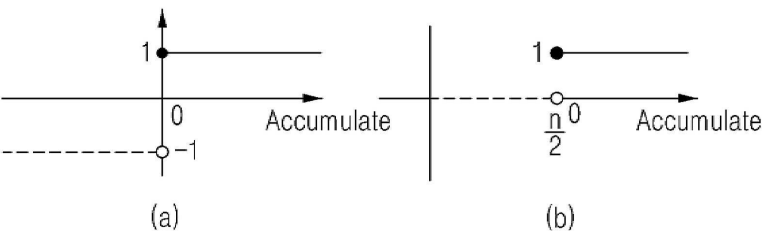
[0068] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

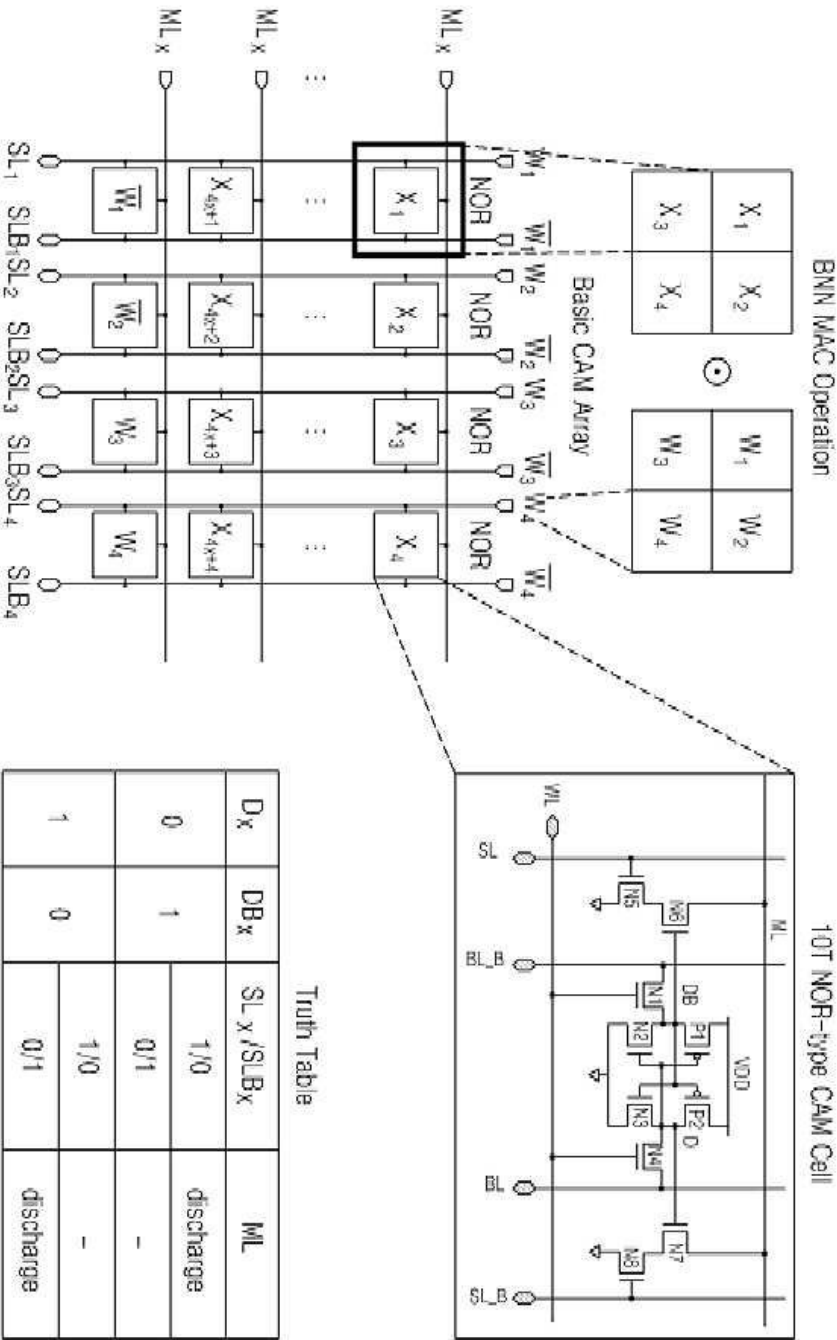
도면1



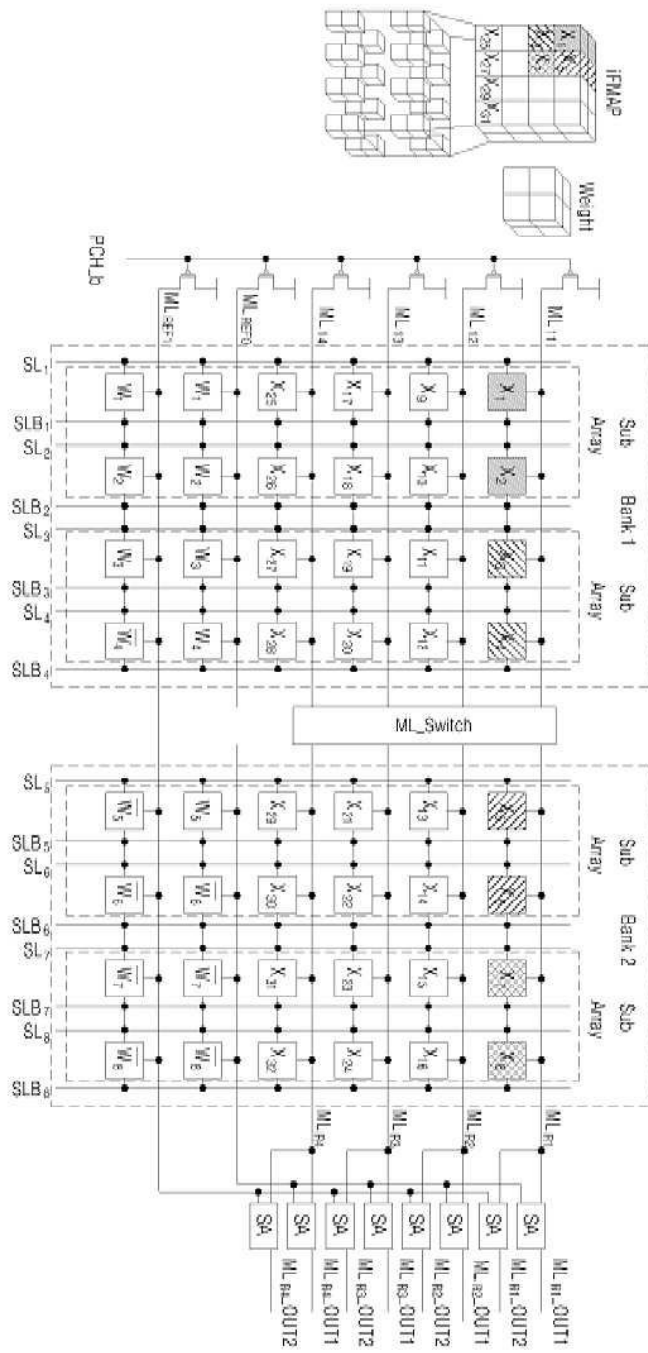
도면2



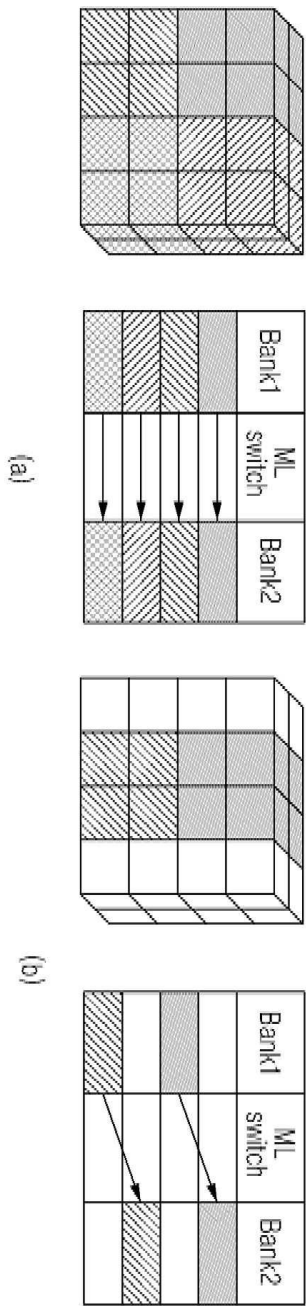
도면3



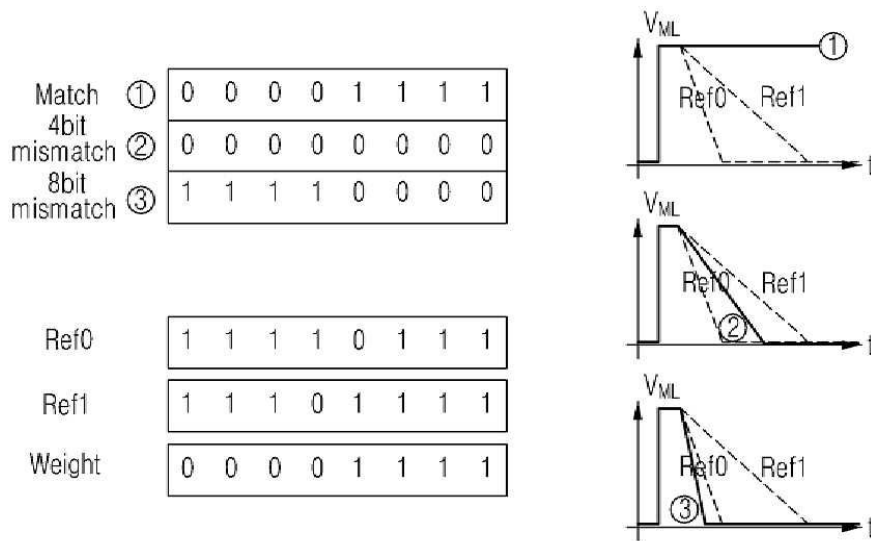
도면4



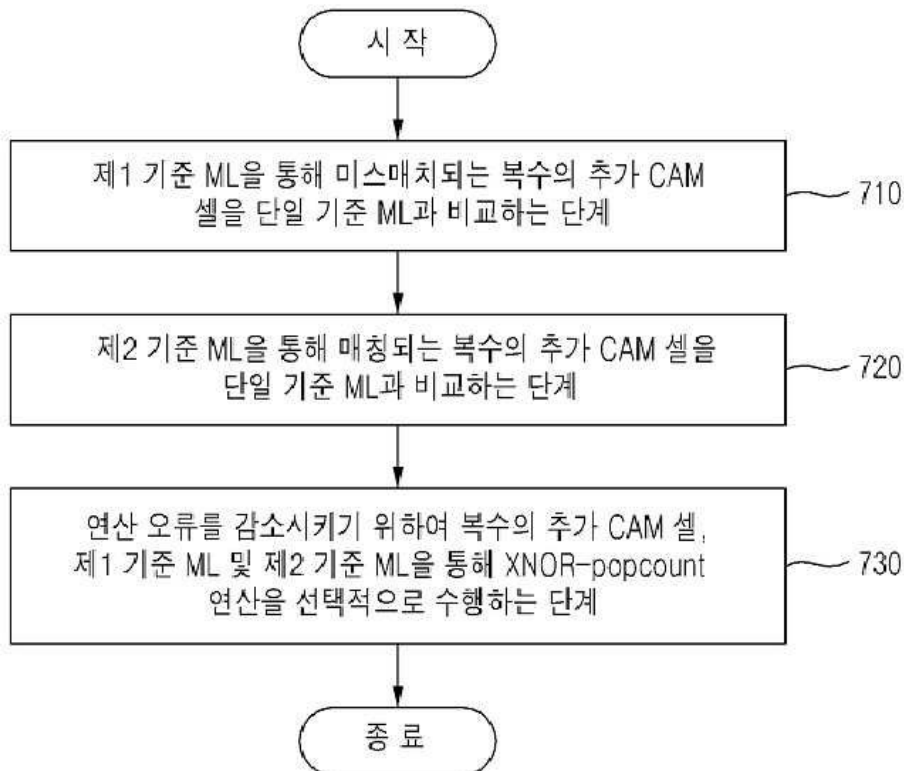
도면5



도면6



도면7



도면8

