



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0094679
(43) 공개일자 2019년08월14일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 3/04 (2006.01)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2018-0014121
(22) 출원일자 2018년02월05일
심사청구일자 2018년06월04일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암
동5가)
(72) 발명자
박중선
서울시 서초구 반포동 반포아파트 85동 403호
최웅
경기도 부천시 소사구 소사본3동 한신아파트 108
동 607호
정광효
경기도 성남시 중원구 도촌남로 101, 707동 1102
호
(74) 대리인
이기성

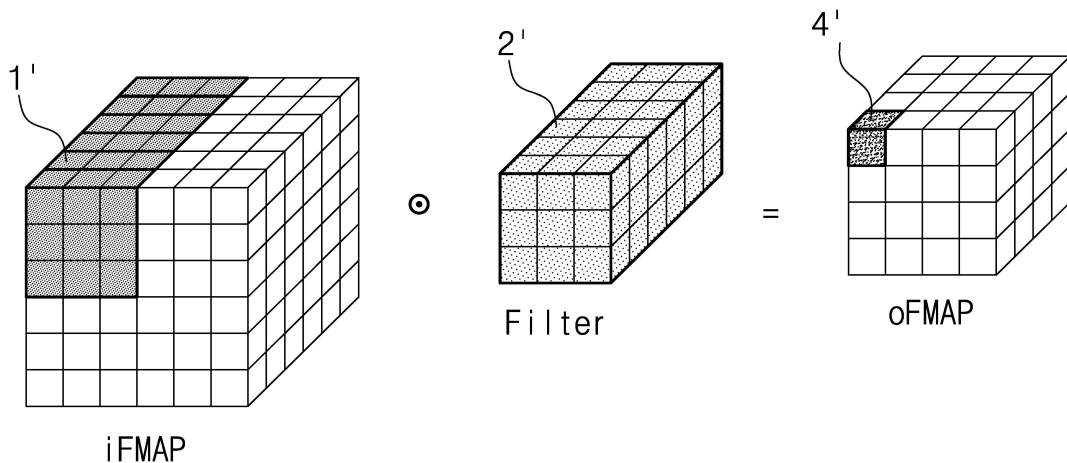
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 이진화 컨볼루션 연산을 지원하는 프로세서 인 메모리 및 그 동작 방법

(57) 요약

본 출원은 인공 신경망 장치 및 그 동작 방법에 관한 것이다. 본 출원의 실시 예에 따른 인공 신경망 장치는 입력 특징맵에 대한 인공 신경망 연산 동작을 수행하는 연산부; 및 상기 연산부의 인공 신경망 연산 동작에 기초하여, 상기 입력 특징맵에 대한 분류 동작을 수행하는 분류부를 포함하며, 상기 연산부는 상기 입력 특징맵과 필터에 대한 XNOR 연산 동작을 수행하는 XNOR 연산 회로; 및 상기 XNOR 연산 회로의 XNOR 연산 결과에 기초하여, 이진화 동작을 수행하는 이진화 회로를 포함한다. 본 출원의 실시 예에 따른 인공 신경망 장치는 소형화 가능하면서도 빠른 속도로 연산 동작을 수행할 수 있다.

대표도 - 도3



(52) CPC특허분류

G06N 3/063 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711076106
부처명 과학기술정보통신부
연구관리전문기관 한국산업기술평가관리원
연구사업명 전자정보디바이스산업원천기술개발(정보화)
연구과제명 스마트 센서 SoC용 초저전압 회로 및 IP 설계 기술 개발
기 여 율 3/10
주관기관 연세대학교 산학협력단
연구기간 2018.04.01 ~ 2019.01.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711071760
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 사물인터넷용 저면적/저전력 전용 프로세서 설계를 위한 맞춤형 임베디드 메모리 개발
기 여 율 4/10
주관기관 고려대학교
연구기간 2018.04.01 ~ 2019.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711070699
부처명 과학기술정보통신부
연구관리전문기관 한국산업기술평가관리원
연구사업명 전자정보디바이스산업원천기술개발(정보화)
연구과제명 스마트 모바일 및 IoT 디바이스를 위한 뉴럴셀(Spiking Neural Cell)기반 SNP SoC 원천기술 개발
기 여 율 3/10
주관기관 전자부품연구원
연구기간 2018.01.01 ~ 2018.12.31

명세서

청구범위

청구항 1

입력 특징맵에 대한 인공 신경망 연산 동작을 수행하는 연산부; 및

상기 연산부의 인공 신경망 연산 동작에 기초하여, 상기 입력 특징맵에 대한 분류 동작을 수행하는 분류부를 포함하며,

상기 연산부는

상기 입력 특징맵과 필터에 대한 XNOR 연산 동작을 수행하는 XNOR 연산 회로; 및

상기 XNOR 연산 회로의 XNOR 연산 결과에 기초하여, 이진화 동작을 수행하는 이진화 회로를 포함하는, 인공 신경망 장치.

청구항 2

제1 항에 있어서,

상기 XNOR 연산 회로는

복수의 매치라인들에 연결되며, 상기 입력 특징맵의 데이터를 저장하는 복수의 메모리 셀들을 포함하는 메모리 셀 어레이; 및

상기 복수의 매치라인들 중 적어도 하나의 매치라인을 프리차지하는 프리차지 회로를 포함하는, 인공 신경망 장치.

청구항 3

제2 항에 있어서,

상기 메모리 셀 어레이는 복수의 소스라인들에 연결되며, 상기 복수의 소스라인들을 통하여 상기 필터의 웨이트 데이터가 제공되는, 인공 신경망 장치.

청구항 4

제3 항에 있어서,

상기 복수의 메모리 셀들 중 적어도 하나의 메모리 셀은,

상기 복수의 매치라인들 중 선택된 매치라인에 연결되며, 저장된 데이터의 상보 값에 응답하여 턴-온 또는 턴-오프 되는 제1 트랜지스터;

상기 선택된 매치라인에 연결되며, 상기 저장된 데이터에 응답하여 턴-온 또는 턴-오프 되는 제2 트랜지스터;

상기 제1 트랜지스터에 연결되며, 상기 복수의 소스라인들 중 선택된 소스라인에 제공되는 데이터에 응답하여 턴-온 또는 턴-오프 되는 제3 트랜지스터; 및

상기 제2 트랜지스터에 연결되며, 상기 선택된 소스라인에 제공되는 데이터의 상보 값에 응답하여 턴-온 또는 턴-오프 되는 제4 트랜지스터를 포함하는, 인공 신경망 장치.

청구항 5

제4 항에 있어서,

상기 저장된 데이터와 상기 선택된 소스라인을 통하여 제공되는 데이터가 서로 동일한 경우에, 상기 적어도 하나의 메모리 셀은 상기 선택된 매치라인에 대한 디스차지 경로를 차단하고,

상기 저장된 데이터와 상기 선택된 소스라인을 통하여 제공되는 데이터가 서로 다른 경우에, 상기 적어도 하나

의 메모리 셀은 상기 선택된 매치라인에 대한 디스차지 경로를 생성하는, 인공 신경망 장치.

청구항 6

제3 항에 있어서,

상기 메모리 셀 어레이는,

기준 매치라인에 연결되며, 상기 기준 매치라인에 대한 디스차지 경로를 생성하거나 차단하는 복수의 기준 셀들을 더 포함하는, 인공 신경망 장치.

청구항 7

제6 항에 있어서,

상기 복수의 기준 셀들 중 적어도 하나의 기준 셀은,

상기 기준 매치라인에 연결되며, 디스차지 경로를 차단하는 제1 트랜지스터;

상기 기준 매치라인에 연결되며, 상기 필터의 웨이트 값에 응답하여 턴-온 또는 턴-오프 되는 제2 트랜지스터;

상기 제1 트랜지스터에 연결되며, 상기 복수의 소스라인들 중 선택된 소스라인에 제공되는 데이터에 응답하여 턴-온 또는 턴-오프 되는 제3 트랜지스터; 및

상기 제2 트랜지스터에 연결되며, 상기 선택된 소스라인에 제공되는 데이터의 상보 값에 응답하여 턴-온 또는 턴-오프 되는 제4 트랜지스터를 포함하는, 인공 신경망 장치.

청구항 8

제7 항에 있어서,

인공 신경망 연산 동작이 수행될 때에, 상기 복수의 기준 셀들 중 일부는 상기 기준 매치라인에 대한 디스차지 경로를 생성하고, 상기 복수의 기준 셀들 중 다른 일부는 상기 기준 매치라인에 대한 디스차지 경로를 차단하는, 인공 신경망 장치.

청구항 9

제6 항에 있어서,

상기 복수의 기준 셀들은 데이터를 저장하지 않는, 인공 신경망 장치.

청구항 10

제6 항에 있어서,

상기 이진화 회로는

복수의 TDA들; 및

상기 복수의 매치라인들과 상기 복수의 TDA들 사이에 배치된 딜레이 버퍼를 포함하며,

상기 복수의 TDA들은 상기 복수의 매치라인들 중 대응하는 매치라인에 각각 연결되고, 상기 기준 매치라인에는 공통적으로 연결되는, 인공 신경망 장치.

청구항 11

제10 항에 있어서,

상기 복수의 TDA들 중 적어도 하나의 TDA는,

상기 복수의 매치라인들 중 대응하는 매치라인의 디스차지 속도와 상기 기준 매치라인의 디스차지 속도의 비교 결과에 기초하여, 출력 특징맵의 데이터를 출력하는, 인공 신경망 장치.

청구항 12

제2 항에 있어서,

상기 복수의 메모리 셀들은

상기 복수의 매치라인들 중 제1 매치라인에 대응하는 제1 그룹의 메모리 셀들; 및

상기 복수의 매치라인들 중 상기 제1 매치라인과 다른 제2 매치라인에 대응하는 제2 그룹의 메모리 셀들을 포함하며,

상기 입력 특징맵의 데이터는 상기 제1 그룹의 메모리 셀들과 상기 제2 그룹의 메모리 셀들에 중첩하여 저장되는, 인공 신경망 장치.

청구항 13

제12 항에 있어서,

상기 제1 그룹에 저장된 데이터에 대한 인공 신경망 연산 동작과 상기 제2 그룹에 저장된 데이터에 대한 인공 신경망 연산 동작은 동시에 수행되는, 인공 신경망 장치.

청구항 14

제2 항에 있어서,

상기 복수의 메모리 셀들은

상기 복수의 매치라인들 중 제1 매치라인에 대응하는 제1 그룹의 메모리 셀들; 및

상기 복수의 매치라인들 중 상기 제1 매치라인과 다른 제2 매치라인에 대응하는 제2 그룹의 메모리 셀들을 포함하며,

상기 입력 특징맵의 데이터는 상기 제1 그룹의 메모리 셀들과 상기 제2 그룹의 메모리 셀들에 중첩되지 않고 저장되는, 인공 신경망 장치.

청구항 15

제2 항에 있어서,

상기 메모리 셀 어레이는,

상기 복수의 메모리 셀들을 저장하는 복수의 बैं크들; 및

상기 복수의 बैं크들 사이에 배치된 매치라인 스위치 회로를 포함하는, 인공 신경망 장치.

청구항 16

제15 항에 있어서,

상기 매치라인 스위치 회로는,

상기 필터의 슬라이딩 동작에 기초하여, 상기 복수의 बैं크들 중 제1 बैं크의 매치라인을 선택하고, 상기 제1 बैं크와 인접한 제2 बैं크의 매치라인을 선택하며, 상기 제1 बैं크의 선택된 매치라인과 상기 제2 बैं크의 선택된 매치라인을 전기적으로 서로 연결시키는, 인공 신경망 장치.

청구항 17

메모리 셀 어레이에 입력 특징맵의 데이터를 프로그램하는 단계;

상기 메모리 셀 어레이에 연결된 복수의 매치라인들 중 적어도 하나의 매치라인을 프리차지하는 단계;

상기 메모리 셀 어레이에 연결된 복수의 소스라인들 중 적어도 하나의 소스라인을 통하여 필터의 웨이트 데이터를 제공하는 단계; 및

상기 메모리 셀 어레이에 저장된 데이터와 상기 필터의 웨이트 데이터에 기초하여, 상기 적어도 하나의 매치라인에 대한 이벤트루이션 동작이 수행되는 단계를 포함하는, 인공 신경망 장치의 동작 방법.

청구항 18

제17 항에 있어서,

상기 적어도 하나의 매치라인에 대한 디스차지 속도와 기준 매치라인에 대한 디스차지 속도의 비교 결과에 기초하여, 출력 특징맵의 데이터를 출력하는 단계를 더 포함하는, 인공 신경망 연산 장치의 동작 방법.

청구항 19

제17 항에 있어서,

상기 복수의 매치라인들 중 제1 매치라인 및 상기 제1 매치라인과 다른 제2 매치라인에 대한 이벨루에이션 동작은 동시에 수행되는, 인공 신경망 장치의 동작 방법.

청구항 20

제17 항에 있어서,

상기 메모리 셀 어레이에는 상기 입력 특징맵의 데이터가 중첩하여 저장되는, 인공 신경망 장치의 동작 방법.

발명의 설명

기술 분야

[0001] 본 출원은 인공 신경망 장치 및 그 동작 방법에 관한 것이다.

배경 기술

[0002] 최근 인공 신경망에 관한 기술이 각광을 받고 있다. 딥러닝(deep learning) 기술의 핵심 계산 모형인 컨볼루션 뉴럴 네트워크(CNN : Convolutional Neural Network, 이하에서 "CNN"이라고 지칭함)은, 각각의 뉴런이 인간의 시각 신경계에서 중복 영역의 응답 특성과 유사한 특성을 갖도록 다중 계층으로 배치된 인공신경망의 한 종류이다.

[0003] CNN은 계층의 수가 많아질수록 복잡한 패턴의 영상 및 음성을 높은 인식률로 인식할 수 있다는 장점이 있다. 그에 따라, 오랫동안 정체 상태에 머물러 있던 영상 및 음성 인식률이 최근 들어 CNN 기술에 의해 극적으로 높아지고, ImageNet과 같은 각종 영상인식 벤치마킹시험에서 CNN 시스템들이 최고기록을 경신하고 있다. 그 결과, CNN은 영상인식, 음성인식, 언어번역 등 다양한 머신러닝(machine learning) 분야에서 각광받고 있다.

발명의 내용

해결하려는 과제

[0004] 본 출원은 소형화 가능하면서도 빠른 연산 동작을 수행할 수 있는 인공 신경망 장치를 제공하는데 목적이 있다.

과제의 해결 수단

[0005] 본 출원의 실시 예에 따른 인공 신경망 장치는 입력 특징맵에 대한 인공 신경망 연산 동작을 수행하는 연산부; 및 상기 연산부의 인공 신경망 연산 동작에 기초하여, 상기 입력 특징맵에 대한 분류 동작을 수행하는 분류부를 포함하며, 상기 연산부는 상기 입력 특징맵과 필터에 대한 XNOR 연산 동작을 수행하는 XNOR 연산 회로; 및 상기 XNOR 연산 회로의 XNOR 연산 결과에 기초하여, 이진화 동작을 수행하는 이진화 회로를 포함한다.

[0006] 실시 예에 있어서, 상기 XNOR 연산 회로는 복수의 매치라인들에 연결되며, 상기 입력 특징맵의 데이터를 저장하는 복수의 메모리 셀들을 포함하는 메모리 셀 어레이; 및 상기 복수의 매치라인들 중 적어도 하나의 매치라인을 프리차지하는 프리차지 회로를 포함한다.

[0007] 실시 예에 있어서, 상기 메모리 셀 어레이는 복수의 소스라인들에 연결되며, 상기 복수의 소스라인들을 통하여 상기 필터의 웨이트 데이터가 제공된다.

[0008] 실시 예에 있어서, 상기 복수의 메모리 셀들 중 적어도 하나의 메모리 셀은, 상기 복수의 매치라인들 중 선택된 매치라인에 연결되며, 저장된 데이터의 상보 값에 응답하여 턴-온 또는 턴-오프 되는 제1 트랜지스터; 상기 선

택된 매치라인에 연결되며, 상기 저장된 데이터에 응답하여 턴-온 또는 턴-오프 되는 제2 트랜지스터; 상기 제1 트랜지스터에 연결되며, 상기 복수의 소스라인들 중 선택된 소스라인에 제공되는 데이터에 응답하여 턴-온 또는 턴-오프 되는 제3 트랜지스터; 및 상기 제2 트랜지스터에 연결되며, 상기 선택된 소스라인에 제공되는 데이터의 상보 값에 응답하여 턴-온 또는 턴-오프 되는 제4 트랜지스터를 포함한다.

[0009] 실시 예에 있어서, 상기 저장된 데이터와 상기 선택된 소스라인을 통하여 제공되는 데이터가 서로 동일한 경우에, 상기 적어도 하나의 메모리 셀은 상기 선택된 매치라인에 대한 디스차지 경로를 차단하고, 상기 저장된 데이터와 상기 선택된 소스라인을 통하여 제공되는 데이터가 서로 다른 경우에, 상기 적어도 하나의 메모리 셀은 상기 선택된 매치라인에 대한 디스차지 경로를 생성한다.

[0010] 실시 예에 있어서, 상기 메모리 셀 어레이는, 기준 매치라인에 연결되며, 상기 기준 매치라인에 대한 디스차지 경로를 생성하거나 차단하는 복수의 기준 셀들을 더 포함한다.

[0011] 실시 예에 있어서, 상기 복수의 기준 셀들 중 적어도 하나의 기준 셀은, 상기 기준 매치라인에 연결되며, 디스차지 경로를 차단하는 제1 트랜지스터; 상기 기준 매치라인에 연결되며, 상기 필터의 웨이트 값에 응답하여 턴-온 또는 턴-오프 되는 제2 트랜지스터; 상기 제1 트랜지스터에 연결되며, 상기 복수의 소스라인들 중 선택된 소스라인에 제공되는 데이터에 응답하여 턴-온 또는 턴-오프 되는 제3 트랜지스터; 및 상기 제2 트랜지스터에 연결되며, 상기 선택된 소스라인에 제공되는 데이터의 상보 값에 응답하여 턴-온 또는 턴-오프 되는 제4 트랜지스터를 포함한다.

[0012] 실시 예에 있어서, 인공 신경망 연산 동작이 수행될 때에, 상기 복수의 기준 셀들 중 일부는 상기 기준 매치라인에 대한 디스차지 경로를 생성하고, 상기 복수의 기준 셀들 중 다른 일부는 상기 기준 매치라인에 대한 디스차지 경로를 차단한다.

[0013] 실시 예에 있어서, 상기 복수의 기준 셀들은 데이터를 저장하지 않는다.

[0014] 실시 예에 있어서, 상기 이진화 회로는 복수의 TDA들; 및 상기 복수의 매치라인들과 상기 복수의 TDA들 사이에 배치된 딜레이 버퍼를 포함하며, 상기 복수의 TDA들은 상기 복수의 매치라인들 중 대응하는 매치라인에 각각 연결되고, 상기 기준 매치라인에는 공통적으로 연결된다.

[0015] 실시 예에 있어서, 상기 복수의 TDA들 중 적어도 하나의 TDA는, 상기 복수의 매치라인들 중 대응하는 매치라인의 디스차지 속도와 상기 기준 매치라인의 디스차지 속도의 비교 결과에 기초하여, 출력 특징맵의 데이터를 출력한다.

[0016] 실시 예에 있어서, 상기 복수의 메모리 셀들은 상기 복수의 매치라인들 중 제1 매치라인에 대응하는 제1 그룹의 메모리 셀들; 및 상기 복수의 매치라인들 중 상기 제1 매치라인과 다른 제2 매치라인에 대응하는 제2 그룹의 메모리 셀들을 포함하며, 상기 입력 특징맵의 데이터는 상기 제1 그룹의 메모리 셀들과 상기 제2 그룹의 메모리 셀들에 중첩하여 저장된다.

[0017] 실시 예에 있어서, 상기 제1 그룹에 저장된 데이터에 대한 인공 신경망 연산 동작과 상기 제2 그룹에 저장된 데이터에 대한 인공 신경망 연산 동작은 동시에 수행된다.

[0018] 실시 예에 있어서, 상기 복수의 메모리 셀들은 상기 복수의 매치라인들 중 제1 매치라인에 대응하는 제1 그룹의 메모리 셀들; 및 상기 복수의 매치라인들 중 상기 제1 매치라인과 다른 제2 매치라인에 대응하는 제2 그룹의 메모리 셀들을 포함하며, 상기 입력 특징맵의 데이터는 상기 제1 그룹의 메모리 셀들과 상기 제2 그룹의 메모리 셀들에 중첩되지 않고 저장된다.

[0019] 실시 예에 있어서, 상기 메모리 셀 어레이는, 상기 복수의 메모리 셀들을 저장하는 복수의 뱅크들; 및 상기 복수의 뱅크들 사이에 배치된 매치라인 스위치 회로를 포함한다.

[0020] 실시 예에 있어서, 상기 매치라인 스위치 회로는, 상기 필터의 슬라이딩 동작에 기초하여, 상기 복수의 뱅크들 중 제1 뱅크의 매치라인을 선택하고, 상기 제1 뱅크와 인접한 제2 뱅크의 매치라인을 선택하며, 상기 제1 뱅크의 선택된 매치라인과 상기 제2 뱅크의 선택된 매치라인을 전기적으로 서로 연결시킨다.

[0021] 본 출원의 실시 예에 따른 인공 신경망 장치의 동작 방법은 메모리 셀 어레이에 입력 특징맵의 데이터를 프로그램하는 단계; 상기 메모리 셀 어레이에 연결된 복수의 매치라인들 중 적어도 하나의 매치라인을 프리차지하는 단계; 상기 메모리 셀 어레이에 연결된 복수의 소스라인들 중 적어도 하나의 소스라인을 통하여 필터의 웨이트 데이터를 제공하는 단계; 및 상기 메모리 셀 어레이에 저장된 데이터와 상기 필터의 웨이트 데이터에 기초하여,

상기 적어도 하나의 매치라인에 대한 이벨루에이션 동작이 수행되는 단계를 포함한다.

[0022] 실시 예에 있어서, 상기 적어도 하나의 매치라인에 대한 디스차지 속도와 기준 매치라인에 대한 디스차지 속도의 비교 결과에 기초하여, 출력 특징맵의 데이터를 출력하는 단계를 더 포함한다.

[0023] 실시 예에 있어서, 상기 복수의 매치라인들 중 제1 매치라인 및 상기 제1 매치라인과 다른 제2 매치라인에 대한 이벨루에이션 동작은 동시에 수행된다.

[0024] 실시 예에 있어서, 상기 메모리 셀 어레이에는 상기 입력 특징맵의 데이터가 중첩하여 저장된다.

발명의 효과

[0025] 본 출원의 실시 예에 따른 인공 신경망 장치는 소형화 가능하면서도 빠른 속도로 연산 동작을 수행할 수 있다.

도면의 간단한 설명

[0026] 도 1은 본 출원의 일 실시 예에 따른 CNN 장치의 일 예를 보여주는 블록도이다.

도 2는 도 1의 CNN 장치의 CNN 연산 동작의 일 예를 개념적으로 보여주는 도면이다.

도 3은 본 출원의 일 실시 예에 따른 BNN 장치의 일 예를 보여주는 블록도이다.

도 4는 도 3의 BNN 장치의 BNN 연산 동작의 일 예를 개념적으로 보여주는 도면이다.

도 5는 도 3의 BNN 장치를 하드웨어적으로 구현하는 일 예를 좀 더 자세히 설명하기 위한 블록도이다.

도 6은 오른쪽으로 슬라이딩되면서 BNN 연산이 수행되는 일 예를 보여주는 도면이다.

도 7은 BNN 연산 결과의 일 예를 보여주는 도면이다.

도 8은 도 6의 연산부를 좀 더 자세히 보여주는 블록도이다.

도 9는 도 8의 XNOR 연산 회로 및 이진화 회로를 좀 더 자세히 보여주는 회로도이다.

도 10은 도 9의 메모리 셀의 일 예를 보여주는 회로도이다.

도 11은 도 9의 기준 메모리 셀의 일 예를 보여주는 회로도이다.

도 12는 도 9의 제1 TDA의 일 예를 보여주는 회로도이다.

도 13a 내지 도 13d는 도 8 및 도 9의 XNOR 연산 회로의 XNOR 연산 동작을 좀 더 자세히 설명하기 위한 도면들이다.

도 14는 입력 특징맵의 데이터와 필터 웨이트의 데이터의 매칭 또는 미스매칭된 경우의 예를 보여주는 도면이다.

도 15는 본 출원의 일 실시 예에 따른 BNN 장치의 BNN 연산 동작의 일 예를 보여주는 도면이다.

도 16은 본 출원의 일 실시 예에 따른 BNN 장치의 BNN 연산 동작의 다른 예를 보여주는 도면이다.

도 17은 본 출원의 실시 예에 따른 BNN 장치의 동작을 설명하기 위한 순서도이다.

도 18은 본 출원의 다른 실시 예에 따른 메모리 셀 어레이를 보여주는 도면이다.

도 19는 도 18의 매치라인 스위치 회로에 포함된 복수의 매치라인 스위치들 중 제1 매치라인 스위치를 예시적으로 보여주는 도면이다.

도 20 및 도 21은 도 18 및 도 19의 매치라인 스위치 회로의 동작을 설명하기 위한 도면들이다.

발명을 실시하기 위한 구체적인 내용

[0027] 이하, 본 출원의 기술적 사상을 본 출원의 기술분야에서 통상의 지식을 가진자가 용이하게 실시할 수 있을 정도로 상세히 설명하기 위하여, 본 출원의 실시 예를 첨부된 도면을 참조하여 설명하기로 한다.

[0028] 도 1은 본 출원의 일 실시 예에 따른 CNN 장치(10)의 일 예를 보여주는 블록도이다. 도 1에서는 인공 신경망 장치의 일 예로 CNN 장치(10)가 도시되어 있다. 도 1을 참조하면, CNN 장치(10)는 컨볼루션층 레이어(11) 및 분류

함수(12)를 포함한다.

- [0029] 컨볼루션 레이어(11)는 입력된 데이터로부터 특징(feature)을 추출하는 동작을 수행한다. 예를 들어, 컨볼루션 레이어(11)는 외부로부터 입력 특징맵(input feature map, 이하 "iFMAP")을 수신하고, 이로부터 특징을 추출하여 출력 특징맵(output feature map, 이하 "oFMAP")을 출력할 수 있다. 이 경우, 입력 특징맵(iFMAP)은 이미지(image)일 수 있으나, 이에 한정되는 것은 아니다. 예를 들어, 입력 특징맵(iFMAP)은 영상, 음성, 언어 등 다양한 형태의 데이터일 수도 있다.
- [0030] 컨볼루션 레이어(11)는 특징을 추출하는 필터(filter)와, 필터의 값을 비선형 값으로 바꾸어주는 액티베이션 함수(activation function)를 포함할 수 있다. 또한, 컨볼루션 레이어(11)는 풀링(pooling), 맥스 풀링(max pooling) 등의 잘 알려진 동작을 수행하는 구성을 추가적으로 포함할 수도 있다.
- [0031] 분류 함수(12)는 컨볼루션 레이어(11)로부터 출력 특징맵(oFMAP)을 수신하고, 이를 분류하는 동작을 수행한다. 분류 함수(12)는, 예를 들어, 풀리 커넥티드 레이어(fully connected layer), 소프트맥스(softmax) 함수, 드롭아웃 레이어(dropout layer) 등을 포함할 수 있다.
- [0032] 도 2는 도 1의 CNN 장치(10)의 컨볼루션 레이어(11)의 CNN 연산 동작의 일 예를 개념적으로 보여주는 도면이다. 설명의 편의상, 도 2에서는 출력 특징맵(oFMAP)의 데이터들 중 하나의 데이터(4)를 연산하는 과정이 예시적으로 도시되어 있다.
- [0033] 컨볼루션 레이어(11)의 CNN 연산 동작은 필터(filter)와 입력 특징맵(iFMAP)의 곱셈 및 덧셈으로 정의될 수 있다. 다시 말하면, 컨볼루션 레이어(11)의 CNN 연산 동작은 입력 특징맵(iFMAP)과 필터의 곱으로 정의되는 부분합(pSUM)과, 상기 부분합들(pSUMs)에 대한 덧셈으로 정의되는 출력 특징맵(oFMAP)으로 정의될 수 있다.
- [0034] 예를 들어, 도 2에 도시된 바와 같이, 필터의 채널(channel)의 깊이(depth)가 "6"이라고 가정하자, 이 경우, 제1 부분합(3_1)은 필터의 제1 부분(2_1) 및 이에 대응하는 입력 특징맵의 부분(1_1)의 곱으로 정의될 수 있다. 또한, 제2 부분합(3_2)은 필터의 제2 부분(2_2) 및 이에 대응하는 입력 특징맵의 부분(1_2)의 곱으로 정의될 수 있다. 마찬가지로, 제6 부분합(3_6)은 필터의 제6 부분(2_6) 및 이에 대응하는 입력 특징맵의 부분(1_6)으로 정의될 수 있다. 그리고, 출력 특징맵(4)은 제1 내지 제6 부분합(3_1~3_6)의 덧셈으로 정의될 수 있다.
- [0035] 이와 같이, 출력 특징맵의 데이터들 중 하나의 데이터(4)를 연산하기 위해서는 복수의 부분합들을 연산하는 과정 및 해당 복수의 부분합들을 누적하여 연산하는 과정이 수행되어야 한다. 따라서, 채널(channel)이 증가할수록 연산 과정이 늘어나게 된다. 더욱이, 이와 같은 CNN 연산 동작을 수행하기 위해서는, 입력 특징맵에 대응하는 데이터를 저장하기 위한 메모리 및 필터에 대응하는 데이터를 저장하기 위한 메모리가 필요할 뿐만 아니라, 부분합들에 대응하는 데이터를 저장하기 위한 메모리가 추가로 필요하다. 따라서, 많은 용량의 메모리가 필요하여 소형화 및 집적화에 불리하다.
- [0036] 도 3은 본 출원의 일 실시 예에 따른 BNN(Binarized Neural Network) 장치(20)의 일 예를 보여주는 블록도이다. 도 3에서는 인공 신경망 장치의 일 예로 BNN 장치(20)가 도시되어 있다. 도 3을 참조하면, BNN 장치(20)는 연산부(21) 및 분류부(22)를 포함한다.
- [0037] 연산부(21)는 입력된 데이터로부터 특징(feature)을 추출하는 동작을 수행한다. 예를 들어, 연산부(21)는 외부로부터 입력 특징맵(iFMAP)을 수신하고, 이로부터 특징을 추출하여 출력 특징맵(oFMAP)을 출력할 수 있다. 이 경우, 입력 특징맵(iFMAP)은 이미지(image)일 수 있으나, 이에 한정되는 것은 아니다. 예를 들어, 입력 특징맵(iFMAP)은 영상, 음성, 언어 등 다양한 형태의 데이터일 수도 있다.
- [0038] 분류부(22)는 연산부(21)로부터 출력 특징맵(oFMAP)을 수신하고, 수신된 출력 특징맵에 기초하여 분류 동작을 수행한다. 예를 들어, 입력 특징맵(iFMAP)이 이미지인 경우, 분류부(22)는 수신된 출력 특징맵에 기초하여, 상기 이미지가 속하는 분야를 결정하는 분류 동작을 수행할 것이다. 분류부(22)는, 예를 들어, 풀리 커넥티드 레이어(fully connected layer), 소프트맥스(softmax) 함수, 드롭아웃 레이어(dropout layer) 등을 포함할 수 있다.
- [0039] 본 출원의 기술적 사상에 따른 실시 예에 있어서, 연산부(21)는 BNN 연산 동작을 지원한다. 여기서, BNN 연산이란 입력 특징맵과 필터를 직접(directly) 연산하는 것을 의미한다. 본 출원의 실시 예에 따른 연산부(21)는, 도 1의 컨볼루션 레이어(11)와 달리, 곱셈 연산을 수행한 후에 덧셈 연산을 수행할 필요가 없으며, 이에 따라 도 1의 컨볼루션 레이어(11)에 비하여 연산량이 줄어드는 효과가 있다.
- [0040] 도 4는 도 3의 BNN 장치(20)의 연산부(21)의 BNN 연산 동작의 일 예를 개념적으로 보여주는 도면이다. 설명의

편의상, 도 4에서는 출력 특징맵(oFMAP)의 데이터들 중 하나의 데이터(4')를 연산하는 과정이 예시적으로 도시되어 있다. 또한, 도 2와 같이, 채널(channel)의 깊이는 "6"이라고 가정된다.

- [0041] 도 2의 CNN 연산 동작과 달리, 도 4의 BNN 연산 동작은 입력 특징맵과 필터를 직접 연산한다. 다시 말하면, 도 2의 CNN 연산 동작이 채널 별로 부분합(pSUM)을 연산한 후에, 도출된 복수의 부분합들(pSUMs)에 대한 덧셈 동작을 수행하는 것에 반하여, 도 4의 CNN 연산 동작은 입력 특징맵과 필터를 직접 연산한다.
- [0042] 따라서, 본 출원의 실시 예에 따른 BNN 연산은 CNN 연산에 비하여 연산량이 줄어들뿐만 아니라, 부분합들(pSUMs)을 저장하기 위한 메모리가 필요치 않아 소형화에 유리하다.
- [0043] 도 5는 도 3의 BNN 장치(20)를 하드웨어적으로 구현하는 일 예를 좀 더 자세히 설명하기 위한 블록도이다. 도 5를 참조하면, BNN 장치(100)는 연산부(110) 및 분류부(120)를 포함하며, 연산부(110)는 배타적 부정 논리합 연산 회로(이하, "XNOR 연산 회로", 111) 및 이진화 회로(112)를 포함한다.
- [0044] XNOR 연산 회로(111)는 외부로부터 입력 특징맵을 수신한다. XNOR 연산 회로(111)는 수신된 입력 특징맵에 대한 XNOR 연산을 수행한다. 여기서, XNOR 연산이란 입력 특징맵의 데이터와 이에 대응하는 필터의 데이터를 비교하여 동일 여부를 판단하는 것을 의미한다.
- [0045] BNN 연산을 수행하는데 있어서, 한 번의 BNN 연산에는 다수의 XNOR 연산이 수반될 수 있다. 예를 들어, 도 6 및 도 7에 도시된 바와 같이, 필터의 크기가 $2 \times 2 \times 2$ 이고, 입력 특징맵의 크기가 $4 \times 4 \times 2$ 라고 가정하면, BNN 연산을 한 번 수행하기 위해서는 8차례의 XNOR 연산이 각각 수행되어야 한다.
- [0046] 본 출원의 기술적 사상에 따른 실시 예에 있어서, XNOR 연산 회로(111)는 한번의 BNN 연산에 수반되는 복수의 XNOR 연산이 동시에 수행되도록 구현될 수 있다. 즉, 본 출원의 실시 예에 따른 XNOR 연산 회로(111)는 병렬 XNOR 연산(parallel XNOR operation)을 지원하도록 하드웨어적으로 구현될 수 있다. 병렬 XNOR 연산을 지원하는 XNOR 회로(111)의 구조 및 동작은 이하에서 좀 더 자세히 설명될 것이다.
- [0047] 이진화 회로(112)는 XNOR 연산 회로(111)로부터 XNOR 연산 결과를 수신한다. 이진화 회로(112)는 XNOR 연산 결과에 기초하여 이진화 동작을 수행하고, 이를 출력 특징맵으로 출력한다. 여기서 이진화 동작이란 복수의 XNOR 연산 결과들에 기초하여, 입력 특징맵 데이터와 필터의 웨이트 데이터가 서로 매칭(matching)된 횟수가 많은지 또는 미스매칭(mismatching)된 횟수가 많은지를 판단하고, 이를 "1" 또는 "0"의 이진수로 출력하는 것을 의미한다.
- [0048] 매칭된 횟수가 많은지 또는 미스매칭된 횟수가 많은지를 판단함에 있어서, 본 출원의 실시 예에 따른 이진화 회로(112)는 복수의 XNOR 연산 결과를 일일이 카운팅하지 않는다. 예를 들어, 도 7a에 도시된 바와 같이, XNOR 연산 결과가 [1,1,0,1,1,1,0,1]이라고 가정할 때, 이진화 회로(112)는 "1" 또는 "0"의 개수를 일일이 카운팅하지 않는다.
- [0049] "1" 또는 "0"의 개수를 일일이 카운팅하는 대신에, 본 출원의 실시 예에 따른 이진화 회로(112)는 XNOR 연산이 수행된 매치라인과 기준 매치라인의 디스차지 속도를 비교함으로써, 매칭된 횟수가 많은지 또는 미스매칭된 횟수가 많은지에 대한 판단 동작을 한 번에 수행한다. 본 출원의 실시 예에 따른 이진화 회로(112)의 구조 및 동작은 이하에서 좀 더 자세히 설명될 것이다.
- [0050] 도 6 및 도 7은 도 5의 연산부(110)의 동작을 좀 더 자세히 설명하기 위한 도면들이다. 구체적으로, 도 6은 오른쪽으로 슬라이딩되면서 BNN 연산이 수행되는 일 예를 보여주는 도면이다. 도 7a는 XNOR 연산 결과 "매칭"이 좀 더 많은 경우의 BNN 연산 결과를 보여주는 일 예이며, 도 7b는 XNOR 연산 결과 "미스매칭"이 좀 더 많은 경우의 BNN 연산 결과를 보여주는 일 예이다.
- [0051] 설명의 편의상, 도 6에 도시된 바와 같이, 입력 특징맵(iFMAP)이 $4 \times 4 \times 2$ 의 크기를 갖는다고 가정된다. 그리고 필터(filter)는 2×2 의 크기를 갖는다고 가정된다. 또한, 필터가 오른쪽으로 슬라이딩되면서, 입력 특징맵 중 제1 필터 윈도우(FW1) 내지 제3 필터 윈도우(FW3)에 대한 BNN 연산이 수행된다고 가정된다.
- [0052] 도 6을 참조하면, 먼저 입력 특징맵(iFMAP) 중 제1 필터 윈도우(FW1)에 대한 BNN 연산이 수행된다. 이 경우, 제1 필터 윈도우(FW1)의 데이터와 이에 대응하는 필터 웨이트의 데이터가 서로 동일하면, XNOR 연산 결과로 "1"이 출력될 것이다. 반대로, 제1 필터 윈도우(FW1)의 데이터와 이에 대응하는 필터 웨이트의 데이터가 서로 다르면, XNOR 연산 결과로 "0"이 출력될 것이다.
- [0053] 제1 필터 윈도우(FW1)에 대한 BNN 연산이 완료된 후에, 필터는 오른쪽으로 슬라이딩되며 제2 및 제3 필터 윈도우

우(FW2, FW3)에 대한 BNN 연산이 순차적으로 수행될 것이다.

- [0054] 도 7a를 참조하면, 제1 필터 윈도우(FW1)에 대한 BNN 연산이 수행될 때에, 제1 필터 윈도우(FW1)는 [1,0,0,1,0,0,1,1]의 데이터열을 갖고, 필터는 [1,0,1,1,0,0,0,1]의 필터 웨이트 데이터열을 갖는다고 가정된다. 이 경우, XNOR 연산 회로(111, 도 5 참조)는 XNOR 연산 동작을 수행하며, XNOR 연산 동작 결과로 [1,1,0,1,1,1,0,1]의 데이터열을 출력할 것이다. 이후, 이진화 회로(112, 도 5 참조)는 XNOR 연산 결과 "1"의 데이터가 많으므로(즉, "매칭"이 많으므로), 출력 특징맵의 데이터로 "1"을 출력할 것이다.
- [0055] 도 7b를 참조하면, 제3 필터 윈도우(FW3)는 [0,0,0,0,1,0,1,0]의 데이터열을 갖고, 필터는 [1,0,1,1,0,0,0,1]의 필터 웨이트 데이터열을 갖는다고 가정된다. 이 경우, XNOR 연산 회로는 XNOR 연산 결과로 [0,1,0,0,0,1,0,0]의 데이터열을 출력할 것이다. 이후, 이진화 회로(112)는 XNOR 연산 결과 "0"의 데이터가 많으므로(즉, "미스매칭"이 많으므로), 출력 특징맵의 데이터로 "0"을 출력할 것이다.
- [0056] 도 8 내지 도 12는 도 6의 연산부(110)를 좀 더 자세히 설명하기 위한 도면들이다. 구체적으로, 도 8은 도 6의 연산부(110)를 좀 더 자세히 보여주는 블록도이며, 도 9는 도 8의 XNOR 연산 회로(111) 및 이진화 회로(112)를 좀 더 자세히 보여주는 회로도이다. 도 10은 도 9의 메모리 셀(X1)의 일 예를 보여주는 회로도이고, 도 11은 도 9의 기준 메모리 셀(W1)의 일 예를 보여주는 회로도이며, 도 12는 도 9의 제1 TDA(TDA1)의 일 예를 보여주는 회로도이다.
- [0057] 먼저, 도 8을 참조하면, 연산부(110)는 XNOR 연산 회로(111), 이진화 회로(112) 및 주변 회로(113)를 포함한다. 또한, XNOR 연산 회로(111)는 프리차지 회로(111_1) 및 메모리 셀 어레이(111_2)를 포함하며, 이진화 회로(112)는 딜레이 버퍼(112_1) 및 TDA(112_2)를 포함한다.
- [0058] 프리차지 회로(111_1)는 매치라인(matchline, 이하 "ML")을 프리차지(precharge)하는 동작을 수행한다. 예를 들어, 도 9에 도시된 바와 같이, 프리차지 회로(111_1)는 PMOS 트랜지스터를 이용하여 구현될 수 있으며, 각각의 PMOS 트랜지스터는 제1 내지 제4 매치라인(ML1~ML4) 및 기준 매치라인(MLref)에 연결될 수 있다. XNOR 연산 동작이 수행될 때에, 프리차지 회로(111_1)는 PCHb 신호에 응답하여 제1 내지 제4 매치라인(ML1~ML4) 및 기준 매치라인(MLref)을 Vdd 전압 레벨로 프리차지할 수 있다.
- [0059] 메모리 셀 어레이(111_2)는 복수의 메모리 셀들을 포함한다. 예를 들어, 도 9에 도시된 바와 같이, 메모리 셀 어레이(111_2)는 제1 내지 제32 메모리 셀들(X1~X32)을 포함하도록 구현될 수 있다. 다만, 이는 예시적인 것이며, 메모리 셀들의 개수는 설계자에 따라 다양하게 변경될 수 있다.
- [0060] 각각의 메모리 셀은 데이터를 저장하도록 구현된다. 예를 들어, 도 10에 도시된 바와 같이, 제1 메모리 셀(X1)은 4개의 NMOS 트랜지스터들(M1~M4)과 2개의 인버터를 사용하여 구현될 수 있다. 예를 들어, 제1 및 제3 NMOS 트랜지스터들(M1, M3)은 서로 직렬 연결되도록 구성되고, 제2 및 제4 NMOS 트랜지스터들(M2, M4)은 서로 직렬 연결되도록 구성될 수 있다.
- [0061] 또한, 제1 소스라인(SL1)은 제3 NMOS 트랜지스터(M3)의 게이트에 제공되고, 제1 상보 소스라인(SLB1)은 제4 NMOS 트랜지스터(M4)의 게이트로 제공되도록 구성될 수 있다. 그리고 제1 매치라인(ML1)은 제1 및 제2 NMOS 트랜지스터들(M1, M2)의 드레인에 제공되도록 구성될 수 있다.
- [0062] 하나의 매치라인에는 복수의 메모리 셀들이 대응하도록 배치된다. 예를 들어, 도 9에 도시된 바와 같이, 하나의 매치라인에는 8개의 메모리 셀들이 대응하도록 배치될 수 있다. 다만, 이는 예시적인 것이며, 하나의 매치라인에 대응하는 메모리 셀들의 개수는 설계자에 따라 다양하게 변경될 수 있다.
- [0063] 또한, 메모리 셀 어레이(111_2)는 복수의 기준 셀들을 더 포함한다. 예를 들어, 도 9에 도시된 바와 같이, 메모리 셀 어레이(111_2)는 기준 매치라인(MLref)에 연결된 8개의 기준 셀들(W1~W4, W5~W8)을 포함할 수 있다. 다만, 이는 예시적인 것이며, 기준 매치라인(MLref)에 연결된 기준 셀들의 개수는 설계자에 따라 다양하게 변경될 수 있다.
- [0064] 각각의 기준 셀들은, 기준 매치라인(MLref)을 디스차지(discharge)하도록 구현된다. 예를 들어, 도 11에 도시된 바와 같이, 제1 기준 셀은 4개의 NMOS 트랜지스터들(M5~M8)을 이용하여 구현될 수 있다. 이 경우, 제5 및 제7 NMOS 트랜지스터(M5, M7)는 서로 직렬 연결되고, 제6 및 제8 NMOS 트랜지스터(M6, M8)는 서로 직렬 연결되도록 구현될 수 있다.
- [0065] 또한, 제1 소스라인(SL1)은 제7 NMOS 트랜지스터(M7)의 게이트에 연결되고, 제1 상보 소스라인(SLB1)은 제8 NMOS 트랜지스터(M8)의 게이트에 연결되도록 구현될 수 있다. 그리고 기준 매치라인(MLref)은 제5 및 제6 NMOS

트랜지스터(M5, M6)의 드레인에 연결되도록 구현될 수 있다.

- [0066] BNN 연산이 수행되는 경우, 본 출원의 기술적 사상의 실시 예에 따른 복수의 기준 셀들 중 일부는 기준 매치라인(MLref)을 디스차지하도록 제어되고, 일부는 기준 매치라인(MLref)을 디스차지하지 않도록 제어될 수 있다. 특히, 본 출원의 기술적 사상의 실시 예에 있어서, 기준 매치라인(MLref)의 디스차지 속도는 매칭이 많은 매치라인의 디스차지 속도보다 빠르고, 미스매칭이 많은 매치라인의 디스차지 속도보다 느리도록 제어될 수 있다.
- [0067] 예를 들어, 각각의 기준 셀의 제6 NMOS 트랜지스터(M6)의 게이트에는 필터의 웨이트 데이터 또는 그 상보 값이 선택적으로 제공될 수 있다. 이 경우, 도 9에 도시된 바와 같이, 제1 내지 제4 기준 셀들(W1~W4)에는 각각 제1 및 제4 웨이트 데이터의 상보 값(W1~W4)이 제공되며, 제5 내지 제8 기준 셀들(W5~W8)에는 제5 내지 제8 웨이트 데이터(W5~W8)가 제공될 수 있다.
- [0068] 이와 같이, 복수의 기준 셀들의 제6 NMOS 트랜지스터(M6)의 게이트에 제공되는 데이터 값을 달리함으로써, 복수의 기준 셀들에 연결된 기준 매치라인(MLref)의 디스차지 속도를 조절할 수 있다. 예를 들어, 기준 매치라인(MLref)의 디스차지 속도는 "매칭"이 많은 매치라인과 "미스매칭"이 많은 매치라인의 디스차지 속도의 중간이 되도록 제어될 수 있다.
- [0069] 다시 도 8 및 도 9를 참조하면, 이진화 회로(112)의 딜레이 버퍼(112_1)는 매치라인(ML) 및 기준 매치라인(MLref)에 연결되며, 매치라인(ML) 및 기준 매치라인(MLref)으로부터 수신된 신호를 TDA(112_2)에 전달하는 동작을 수행한다.
- [0070] TDA(112_2)는 딜레이 버퍼(112_1)를 통하여 매치라인(ML)의 신호들 및 기준 매치라인(MLref)의 신호를 수신한다. 예를 들어, 도 9에 도시된 바와 같이, TDA(112_2)는 제1 내지 제4 TDA(TDA1~TDA4)를 포함하며, 제1 내지 제4 TDA(TDA1~TDA4)는 각각 제1 내지 제4 매치라인(ML1~ML4)에 연결될 수 있다. 또한, 제1 내지 제4 TDA(TDA1~TDA4)는 각각 기준 매치라인(MLref)에 연결될 수 있다.
- [0071] TDA(112_2)는 매치라인(ML) 신호 및 기준 매치라인(MLref) 신호를 비교하고, 비교 결과를 출력 특징맵(ofMAP)의 데이터로 출력하도록 구현될 수 있다.
- [0072] 예를 들어, 도 12에 도시된 바와 같이, 제1 TDA(TDA1)는 제1 감지 노드(S01)를 통하여 제1 매치라인(ML1)의 신호를 수신하고, 기준 감지 노드(S0ref)를 통하여 기준 매치라인(MLref)의 신호를 수신한다. 제1 TDA(TDA1)는 매치라인(ML1) 및 기준 매치라인(MLref)의 디스차지 속도의 비교 결과에 기초하여, "1" 또는 "0"의 데이터 중 어느 하나를 출력하도록 구현될 수 있다. 다만, 도 12에 도시된 TDA의 구성은 예시적인 것이며, 설계자에 따라 TDA는 다양하게 변경될 수 있다.
- [0073] 도 13a 내지 도 13d는 도 8 및 도 9의 XNOR 연산 회로(111)의 XNOR 연산 동작을 좀 더 자세히 설명하기 위한 도면들이다.
- [0074] 설명의 편의상, 제1 매치라인(ML1)에 연결된 제1 및 제n 메모리 셀(X1, Xn)에는 모두 데이터 "0"이 저장되어 있다고 가정된다. 다시 말하면, BNN 연산이 수행될 입력 특징맵(iFMAP)의 데이터 중 일부가 제1 및 제n 메모리 셀(X1, Xn)에 로드(load)되되, 해당 입력 특징맵의 데이터는 "0"인 것으로 가정된다.
- [0075] 먼저, 도 13a를 참조하면, 제1 매치라인(ML1)이 프리차지된다. 즉, 프리차지 회로(111_1, 도 8 및 도 9 참조)에 의하여 제1 매치라인(ML1)이 Vdd 전압 레벨로 프리차지된다.
- [0076] 도 13b를 참조하면, 필터의 웨이트 데이터가 소스라인 및 상보 소스라인에 제공된다. 예를 들어, 제1 내지 제n 소스라인(SL1~SLn)에 제공되는 필터의 웨이트 데이터가 모두 "1"이라고 가정하자. 이 경우, 도 13b에 도시된 바와 같이, 제1 소스라인(SL1) 내지 제n 소스라인(SLn)에는 데이터 "1"에 대응하는 논리 하이(H)의 전압이 제공될 수 있다. 그리고 제1 상보 소스라인(SLB1) 내지 제n 상보 소스라인(SLBn)에는 데이터 "0"에 대응하는 논리 로우(L)의 전압이 제공될 수 있다.
- [0077] 도 13c를 참조하면, 제1 매치라인(ML1)에 대한 이벨류에이션(evaluation) 동작이 수행된다. 즉, 제1 매치라인(ML1)에 연결된 메모리 셀들이 턴-온 되거나 턴-오프 됨으로써, 제1 매치라인(ML1)에 대한 디스차지 경로(discharge path)가 생성되거나 차단되는 동작이 수행된다.
- [0078] 예를 들어, 도 13c에 도시된 바와 같이, 제1 내지 제n 소스라인(SL1~SLn)에 제공되는 필터의 웨이트 데이터가 모두 "1"인 경우, 제1 메모리 셀(X1)의 제3 NMOS 트랜지스터(M3)는 턴-온 된다. 또한, 제1 메모리 셀(X1)에 저장된 데이터가 "1"이기 때문에, 제1 NMOS 트랜지스터(M1) 역시 턴-온 된다. 이에 따라, 제1 메모리 셀(X1)의 제

1 및 제3 NMOS 트랜지스터(M1, M3)에 의하여 디스차지 경로가 생성되며, 제1 매치라인(ML1)이 디스차지된다.

- [0079] 이와 유사하게, 제n 메모리 셀(Xn)에 저장된 데이터가 "1"이기 때문에, 제n 메모리 셀(Xn)의 제1 및 제3 NMOS 트랜지스터(M1, M3)에 의하여 디스차지 경로가 생성될 수 있다. 이와 같이, 메모리 셀들에 저장된 데이터와 필터 웨이트의 데이터가 동일하지 않은 경우(즉, "미스매칭"인 경우)에는 디스차지 경로가 생성되기 때문에, 제1 매치라인(ML1)은 빠르게 디스차지 될 수 있다.
- [0080] 이와 반대로, 소스 라인(SL)을 통하여 제공되는 데이터가 "0"인 경우, 디스차지 경로가 차단될 수 있다. 예를 들어, 도 13d에 도시된 바와 같이, 제1 내지 제n 소스라인(SL1~SLn)에 제공되는 데이터가 모두 "0"이라고 가정하자. 이 경우, 제1 메모리 셀(X1)의 제3 NMOS 트랜지스터(M3)는 턴-오프 된다. 또한, 제1 메모리 셀(X1)에 데이터 "0"이 저장되어 있기 때문에, 제2 NMOS 트랜지스터(M2) 역시 턴-오프 된다. 따라서, 제1 매치라인(ML1)에 대한 디스차지 경로가 차단된다.
- [0081] 마찬가지로, 제n 메모리 셀(Xn)에 저장된 데이터가 "0"이기 때문에, 제n 메모리 셀(Xn)에 의하여도 디스차지 경로는 차단된다. 이와 같이, 메모리 셀들에 저장된 데이터와 필터 웨이트의 데이터가 동일한 경우(즉, "매칭"인 경우)에는 디스차지 경로가 차단되기 때문에, 제1 매치라인(ML1)은 디스차지되지 않거나 천천히 디스차지될 수 있다.
- [0082] 도 14a 내지 도 14c는 도 8 및 도 9의 이진화 회로(112)의 이진화 동작을 설명하기 위한 도면들이다. 구체적으로, 도 14a는 입력 특징맵의 데이터와 필터 웨이트의 데이터가 서로 "매칭"된 횟수가 많은 경우의 일 예를 보여주는 도면이다. 도 14b는 입력 특징맵의 데이터와 필터 웨이트의 데이터가 서로 "미스매칭"된 횟수가 많은 경우의 일 예를 보여주는 도면이다. 도 14c는 "매칭"된 횟수에 따른 디스차지 속도 차이 및 이진화 동작을 간략히 보여주는 도면이다.
- [0083] 먼저, 도 14a를 참조하면, 입력 특징맵(iFMAP)의 데이터열이 [1,0,0,1,0,0,1,1]이고, 필터 웨이트(Filter Weight)의 데이터열이 [1,0,1,1,0,0,0,1]이라고 가정된다. 또한, 제1 시간(t1)에 제1 매치라인(ML1)이 Vdd로 프리차지되고, 제2 시간(t2)에 소스라인(SL)을 통하여 필터 웨이트의 데이터열이 제1 매치라인(ML1)에 연결된 메모리 셀들에 제공된다고 가정하자.
- [0084] 이 경우, 입력 특징맵과 이에 대응하는 필터 웨이트의 데이터가 서로 동일한 횟수(즉, "매칭"된 횟수)가 서로 동일하지 않은 횟수(즉, "미스매칭"된 횟수)보다 많다. 이에 따라, 도 13d에서 설명된 바와 같이, 동일한 매치라인에 연결된 복수의 메모리 셀들 중 디스차지 경로를 차단하는 메모리 셀들의 개수가 디스차지 경로를 생성하는 메모리 셀들의 개수보다 많아지게 된다.
- [0085] 따라서, "매칭"이 많은 경우에 해당 매치라인은 기준 매치라인에 비하여 천천히 디스차지된다. 이 경우, 제1 TDA(TDA1)로 제공되는 기준 매치라인(MLref)이 제1 매치라인(ML1)에 비하여 보다 빨리 논리 로우(L)로 천이되며, 이에 따라 제1 TDA(TDA1)는 출력 특징맵의 데이터로 "1"을 출력할 것이다.
- [0086] 도 14b를 참조하면, 입력 특징맵(iFMAP)의 데이터열이 [0,0,0,0,1,0,1,0]이고, 필터 웨이트(Filter Weight)의 데이터열이 [1,0,1,1,0,0,0,1]이라고 가정된다. 또한, 제1 시간(t1)에 제1 매치라인(ML1)이 Vdd로 프리차지되고, 제2 시간(t2)에 소스라인(SL)을 통하여 필터 웨이트의 데이터열이 제1 매치라인(ML1)에 연결된 메모리 셀들에 제공된다고 가정된다.
- [0087] 이 경우, 입력 특징맵과 이에 대응하는 필터 웨이트의 데이터가 서로 "미스매칭"되는 횟수가 서로 "매칭"되는 횟수보다 많다. 이에 따라, 디스차지 경로가 생성하는 메모리 셀들이 디스차지 경로를 차단하는 메모리 셀들보다 많아지게 된다. 따라서, "미스매칭"이 많은 경우에 해당 매치라인은 기준 매치라인에 비하여 빨리 디스차지된다. 이 경우, 제1 TDA(TDA1)로 제공되는 제1 매치라인(ML1)이 기준 매치라인(MLref)에 비하여 보다 빨리 논리 로우(L)로 천이되며, 이에 따라 제1 TDA(TDA1)는 출력 특징맵의 데이터로 "0"을 출력할 것이다.
- [0088] 도 14c를 참조하면, "미스매칭"된 횟수에 따른 이진화 동작이 간략하게 도시되어 있다. 앞서 설명된 바와 같이, "미스매칭"의 경우에 디스차지 경로가 생성된다. 따라서, 입력 특징맵의 데이터열과 필터 웨이트의 데이터열이 모두 "매칭"된다면, 디스차지 경로가 생성되지 않으며, 이 경우 TDA(112_2)는 출력 특징맵의 데이터로 "1"을 출력한다.
- [0089] 또한, "매칭"이 "미스매칭"보다 많다면, 해당 매치라인은 기준 매치라인보다 천천히 디스차지되며, 이 경우에 TDA(112_2)는 출력 특징맵의 데이터로 역시 "1"을 출력한다.
- [0090] 그러나, "미스매칭"이 "매칭"보다 많다면, 해당 매치라인은 기준 매치라인보다 빨리 디스차지되면, 이 경우에

TDA(112_2)는 출력 특징맵의 데이터로 "0"을 출력한다.

- [0091] 도 15는 본 출원의 일 실시 예에 따른 BNN 장치(100)의 BNN 연산 동작의 일 예를 보여주는 도면이다. 예를 들어, 도 15에서는, 필터가 오른쪽으로 슬라이딩 되면서 BNN 연산 동작이 수행되는 경우가 도시되어 있다.
- [0092] 본 출원의 기술적 사상에 따른 실시 예에 있어서, 서로 다른 필터 웨이트들에 대한 BNN 연산 동작들이 병렬적으로 수행될 수 있다.
- [0093] 도 15를 참조하여 좀 더 자세히 설명하면, 입력 특징맵(iFMAP) 중 제1 필터 윈도우(FW1)의 데이터열이 제1 매치라인(ML1)의 메모리 셀들(X1~X8)에 저장될 수 있다. 입력 특징맵(iFMAP) 중 제2 필터 윈도우(FW2)의 데이터열이 제2 매치라인(ML2)의 메모리 셀들(X9~X16)에 저장될 수 있다. 이와 유사하게, 제3 필터 윈도우(FW3)의 데이터열 및 제4 필터 윈도우(FW4)의 데이터열은 각각 제3 및 제4 매치라인(ML3, ML4)의 메모리 셀들에 저장될 수 있다.
- [0094] 이후, 제1 내지 제4 매치라인(ML1~ML4)이 모두 프라치지 되고, 소스라인(SL)을 통하여 필터 웨이트의 데이터열이 제공될 수 있다.
- [0095] 이 경우, 제1 내지 제4 매치라인(ML1~ML4)에 대한 디스차지 동작이 동시에 수행될 수 있으며, 제1 내지 제4 매치라인(ML1~ML4)과 기준 매치라인의 디스차지 속도를 비교하는 동작 역시 동시에 수행될 수 있다.
- [0096] 결국, 본 출원의 실시 예에 따른 BNN 장치(100)는 제1 내지 제4 필터 웨이트(FW1~FW4)에 대한 BNN 연산 동작을 병렬적으로 수행할 수 있으며, 이에 따라 BNN 연산 동작이 더욱 빠르게 수행될 수 있다.
- [0097] 한편, 상술한 설명은 예시적인 것이며, 본 출원의 기술적 사상은 이에 한정되지 않음이 이해될 것이다. 예를 들어, 도 15에서는, 제1 필터 윈도우(FW1)와 제2 필터 윈도우(FW2)의 일부 영역은 서로 중첩되며, 이에 따라 필터 윈도우의 데이터가 메모리 셀 어레이에 중첩되어 저장되는 것으로 설명되었다. 그러나 본 출원의 기술적 사상의 실시 예에 따른 BNN 장치는 중첩되지 않는 필터 윈도우의 데이터가 메모리 셀 어레이에 저장되도록 구현될 수도 있다.
- [0098] 도 16은 본 출원의 일 실시 예에 따른 BNN 장치(100)의 BNN 연산 동작의 다른 예를 보여주는 도면이다. 도 16에서는, 서로 중첩되지 않는 필터 윈도우가 메모리 셀 어레이에 저장되는 예가 도시되어 있다.
- [0099] 도 16을 참조하면, 입력 특징맵(iFMAP) 중 제1 내지 제4 필터 윈도우들(FW1~FW4)의 데이터열이 각각 제1 내지 제4 매치라인(ML1~ML4)에 대응하는 메모리 셀들에 저장될 수 있다. 이 경우, 도 15와 달리, 제1 내지 제4 필터 윈도우(FW1~FW4)는 서로 중첩되지 않도록 설정될 수 있다.
- [0100] 필터 윈도우들이 서로 중첩되지 않도록 설정되는 경우, 메모리 셀 어레이에 저장되는 데이터들 역시 서로 중첩되지 않는다. 따라서, 메모리 셀 어레이에는 좀 더 용량의 입력 특징맵이 저장될 수 있다. 또한, 도 15와 유사하게, 제1 내지 제4 필터 윈도우들(FW1~FW4)에 대한 BNN 연산 동작들이 병렬적으로 수행 가능하기 때문에, BNN 연산 동작이 빠르게 수행될 수 있다.
- [0101] 도 17은 본 출원의 실시 예에 따른 BNN 장치의 동작을 설명하기 위한 순서도이다.
- [0102] S110 단계에서, 입력 특징맵(iFMAP)의 데이터가 메모리 셀 어레이에 프로그램 된다. 예를 들어, 도 15에서 설명된 바와 같이, 필터 윈도우들 중 일부가 서로 중첩될 수 있으며, 이에 따라 일부 데이터가 중첩된 채로, 메모리 셀 어레이에 저장될 수 있다. 다른 예로, 도 16에서 설명된 바와 같이, 필터 윈도우들이 서로 중첩되지 않을 수 있으며, 이에 따라 데이터의 중첩 없이, 메모리 셀 어레이에 필터 윈도우들에 대응하는 데이터들이 저장될 수 있다.
- [0103] S120 단계에서, 매치라인(ML)이 프리차지 된다. 예를 들어, 도 13a에서 설명된 바와 같이, 프리차지 회로에 의하여 매치라인은 Vdd 전압으로 프리차지될 것이다.
- [0104] S130 단계에서, 필터 웨이트에 기초하여 소스라인(SL)이 활성화된다. 예를 들어, 도 13b에서 설명된 바와 같이, 필터 웨이트의 데이터열이 소스라인(SL)에 제공될 것이다.
- [0105] S140 단계에서, 매치라인(ML)에 대한 이벨류에이션 동작이 수행된다. 예를 들어, 도 13c에서 설명된 바와 같이, 소스라인(SL)을 통하여 제공되는 데이터 및 메모리 셀에 저장된 데이터에 따라, 매치라인(ML)에 대한 디스차지 경로가 생성될 것이다.
- [0106] 예를 들어, 소스라인(SL)의 데이터와 메모리 셀에 저장된 데이터가 서로 다른 경우, 매치라인(ML)에 대한 디스차지 경로가 생성될 것이다. 다른 예로, 소스라인(SL)의 데이터와 메모리 셀에 저장된 데이터가 서로 동일한 경

우, 매치라인(ML)에 대한 디스차지 경로는 차단될 것이다.

- [0107] S150 단계에서, 이진화 동작이 수행될 것이다. 예를 들어, 도 14a 및 도 14b에서 설명된 바와 같이, 매치라인(ML)의 디스차지 속도 및 기준 매치라인(MLref)의 디스차지 속도를 서로 비교함으로써, 출력 특징맵의 데이터를 출력하는 이진화 동작이 수행될 수 있다.
- [0108] 예를 들어, 매치라인(ML)이 기준 매치라인(MLref)에 비하여 느리게 디스차지되는 경우, 입력 특징맵의 필터 윈도우와 필터 웨이트는 서로 "매칭"되는 데이터가 많다고 판단되며, 출력 특징맵의 데이터로 "1"이 출력될 것이다. 다른 예로, 매치라인(ML)이 기준 매치라인(MLref)보다 빨리 디스차지 되는 경우, 입력 특징맵의 필터 윈도우와 필터 웨이트는 서로 "미스매칭"되는 데이터가 많다고 판단되며, 출력 특징맵의 데이터로 "0"이 출력될 것이다.
- [0109] 도 8 내지 도 17을 참조하여 설명된 바와 같이, 본 출원의 기술적 사상의 실시 예에 따른 BNN 장치(100)는 XNOR 연산 및 이진화 동작을 수행하는 회로를 제공한다. 특히, 본 출원의 기술적 사상의 실시 예에 따른 BNN 장치(100)는 매치라인(ML)에 연결된 메모리 셀들에 입력 특징맵의 데이터열을 저장하고, 소스라인(SL)을 통하여 필터 웨이트의 데이터열을 제공함으로써, 복수의 XNOR 연산을 동시에 수행하도록 구현될 수 있다.
- [0110] 또한, 본 출원의 실시 예에 따른 BNN 장치(100)는 매치라인과 기준 매치라인의 디스차지 속도를 비교하여 출력 특징맵의 데이터를 비교한다. 따라서, XNOR 연산 결과를 일일이 카운트할 필요가 없이, "매칭"이 많은지 또는 "미스매칭"이 많은지에 대한 판단 및 출력 특징맵의 데이터 출력 동작을 한번에 수행하도록 구현될 수 있다.
- [0111] 한편, 상술한 설명은 예시적인 것이며, 본원의 기술적 사상은 이에 한정되지 않음이 이해될 것이다. 또한, 본 출원의 기술적 사상은 다양하게 적용 및 응용될 수 있음이 이해될 것이다. 이하에서는, 본 출원의 기술적 사상의 다양한 적용 예 및 응용 예들이 좀 더 자세히 설명될 것이다.
- [0112] 도 18 및 도 19는 본 출원의 다른 실시 예에 따른 BNN 장치(200)를 보여주는 도면들이다. 구체적으로, 도 18은 본 출원의 다른 실시 예에 따른 메모리 셀 어레이(211_2)를 보여주는 도면이며, 도 19는 도 18의 매치라인 스위치 회로(211_3)에 포함된 복수의 매치라인 스위치들 중 제1 매치라인 스위치(ML Switch_1)를 예시적으로 보여주는 도면이다.
- [0113] 도 18 및 도 19의 BNN 장치(200)는 도 8 내지 도 17의 BNN 장치(100)와 유사하다. 따라서, 동일하거나 유사한 구성은 동일하거나 유사한 참조번호를 사용하여 표기될 것이며, 중복되는 설명은 간략한 설명을 위하여 이하 생략될 것이다.
- [0114] 도 18을 참조하면, 메모리 셀 어레이(211_2)는 복수의 뱅크들을 포함하며, 뱅크와 뱅크 사이에는 매치라인 스위치 회로(211_3)가 위치하도록 배치될 수 있다. 매치라인 스위치 회로(211_3)는 복수의 매치라인 스위치들을 포함하며, 각각의 매치라인 스위치는 레프트 매치라인(left ML)과 라이트 매치라인(right ML)을 서로 전기적으로 연결시키는 동작을 수행한다.
- [0115] 예를 들어, 도 19에 도시된 바와 같이, 제1 매치라인 스위치(ML Switch 1)는 하나의 레프트 매치라인(ML_l2)과 두 개의 라이트 매치라인(ML_r1, ML_r2) 사이에 배치될 수 있으며, 외부의 제어 신호에 응답하여 제2 레프트 매치라인(ML_l2)을 제1 및 제2 라이트 매치라인들(ML_r1, ML_r2) 중 어느 하나에 연결시킬 수 있다.
- [0116] 다만, 이는 예시적인 것이며, 매치라인 스위치(ML Switch 1)에 대응하는 레프트 매치라인의 개수와 라이트 매치라인의 개수는 설계자에 따라 다양하게 변경될 수 있다.
- [0117] 본 출원의 기술적 사상에 따른 실시 예에 있어서, 매치라인 스위치 회로(211_3)는 필터의 라이트 슬라이딩(right sliding)에 기초하여 스위칭 동작을 수행할 수 있다. 특히, 중첩되지 않은 복수의 필터 윈도우들에 대한 BNN 연산 동작이 동시에 수행되는 경우에, 본 출원의 기술적 사상에 따른 매치라인 스위치 회로(211_3)가 구비될 수 있다. 이는 이하의 도 20에서 좀 더 자세히 설명될 것이다.
- [0118] 도 20 및 도 21은 도 18 및 도 19의 매치라인 스위치 회로(211_3)의 동작을 설명하기 위한 도면들이다.
- [0119] 설명의 편의상, 입력 특징맵(iFMAP)은 14 X 14의 크기를 가지며, 필터는 5 X 5의 크기를 갖는다고 가정된다. 또한, 필터에 대응하는 필터 윈도우 역시 5 X 5의 크기를 가지며, 제1 내지 제4 필터 윈도우(FW1~FW4)에 대한 BNN 연산 동작이 병렬적으로 수행된다고 가정된다. 또한, 필터는 오른쪽으로 슬라이딩되며, 스트라이드(stride)는 "1"이라고 가정된다. 그리고 메모리 셀 어레이는 5개의 뱅크들을 포함하며, 뱅크와 뱅크 사이에 제1 내지 제4 매치라인 스위치 회로들(211_31~211_34)이 배치된다고 가정된다. 입력 특징맵의 좌표가 (row, column)으로 표시

되어 있다.

- [0120] 먼저, 도 20a를 참조하면, 입력 특징맵(iFMAP)의 데이터가 메모리 셀 어레이에 저장될 것이다. 예를 들어, 제1 필터 윈도우(FW1)에 대응하는 데이터는 제1 내지 제5 뱅크(BANK 1~BANK 5)의 제1 로우에 저장되고, 제2 필터 윈도우(FW2)에 대응하는 데이터는 제1 내지 제5 뱅크(BANK 1~BANK 5)의 제2 로우에 저장될 것이다. 또한, 제3 필터 윈도우(FW3)에 대응하는 데이터는 제1 내지 제5 뱅크(BANK 1~BANK 5)의 제4 로우에 저장되고, 제4 필터 윈도우(FW4)에 대응하는 데이터는 제1 내지 제5 뱅크(BANK 1~BANK 5)의 제5 로우에 저장될 것이다.
- [0121] 제1 필터 윈도우(FW1)를 예로 들어 좀 더 자세히 설명하면, 제1 필터 윈도우(FW1)의 첫 번째 컬럼((1,1)~(5,1))에 대응하는 데이터는 제1 뱅크(BANK 1)의 제1 로우에 저장되고, 두 번째 컬럼((1,2)~(5,2))에 대응하는 데이터는 제2 뱅크(BANK 2)의 제1 로우에 저장된다. 이와 같은 방법으로, 제1 필터 윈도우(FW1)의 데이터는 제1 내지 제5 뱅크들(BANK 1~BANK 5)의 제1 로우에 저장된다.
- [0122] 계속해서 도 20a를 참조하면, 제1 내지 제4 필터 윈도우(FW1~FW4)에 대한 BNN 연산 동작이 수행될 것이다. 이 경우, 도 16에서 설명된 바와 같이, 제1 내지 제4 필터 윈도우(FW1~FW4)에 대한 BNN 연산 동작은 병렬적으로 동시에 수행될 수 있다. 또한, 제1 내지 제4 필터 윈도우들(FW1~FW4)의 데이터들이 제1 내지 제5 뱅크들(BANK 1~BANK 5)의 동일한 로우에 위치하기 때문에, 제1 내지 제4 매치라인 스위치들(211_31~211_34)은 각각 동일한 로우에 위치하는 레프트 매치라인과 라이트 매치라인을 서로 연결할 것이다.
- [0123] 도 20b를 참조하면, 필터가 오른쪽으로 "1"만큼 슬라이딩 된다. 이 경우, 제1 내지 제4 필터 윈도우들(FW1~FW4) 역시 오른쪽으로 "1"만큼 슬라이딩 되며, 이에 따라 제1 내지 제4 필터 윈도우들(FW1~FW4)에 대응하는 데이터들 중 일부 데이터가 변경된다.
- [0124] 좀 더 자세히 설명하면, 필터가 오른쪽으로 "1"만큼 슬라이딩 됨에 따라, 이전에 제2 필터 윈도우(FW2)에 속하던 (1,6) 내지 (5,6)의 데이터가 제1 필터 윈도우(FW1)의 데이터로 변경된다. 따라서, 제1 매치라인 스위치 회로(211_31)는 제1 뱅크(BANK 1)의 제2 레프트 매치라인(ML_12)과 제2 뱅크(BANK 2)의 제1 라이트 매치라인(ML_r1)이 서로 연결되도록 스위칭 동작을 수행한다.
- [0125] 이와 유사하게, 제1 매치라인 스위치 회로(211_31)는 제1 뱅크(BANK 1)의 제3 레프트 매치라인(ML_13)과 제2 뱅크(BANK 2)의 제2 라이트 매치라인(ML_r2)을 서로 연결하고, 제1 뱅크(BANK 1)의 제5 레프트 매치라인(ML_15)과 제2 뱅크(BANK 2)의 제4 라이트 매치라인(ML_r4)을 서로 연결하며, 제1 뱅크(BANK 1)의 제6 레프트 매치라인(ML_16)과 제2 뱅크(BANK 2)의 제5 라이트 매치라인(ML_r5)을 서로 연결할 것이다.
- [0126] 도 20c를 참조하면, 필터가 오른쪽으로 "1"만큼 더 슬라이딩 된다. 이 경우, 도 20b와 비교하여, 제1 내지 제4 필터 윈도우들(FW1~FW4) 역시 오른쪽으로 "1"만큼 더 슬라이딩 되며, 이에 따라 제1 내지 제4 필터 윈도우들(FW1~FW4)에 대응하는 데이터들 중 일부 데이터가 변경된다.
- [0127] 좀 더 자세히 설명하면, 필터가 오른쪽으로 "1"만큼 슬라이딩 됨에 따라, 이전에 제2 필터 윈도우(FW2)에 속하던 (1,7) 내지 (5,7)의 데이터가 제1 필터 윈도우(FW1)의 데이터로 변경된다. 따라서, 제2 매치라인 스위치 회로(211_32)는 제2 뱅크(BANK 2)의 제2 레프트 매치라인(ML_12)과 제3 뱅크(BANK 3)의 제1 라이트 매치라인(ML_r1)이 서로 연결되도록 스위칭 동작을 수행한다.
- [0128] 이와 유사하게, 제2 매치라인 스위치 회로(211_32)는 제2 뱅크(BANK 2)의 제3 레프트 매치라인(ML_13)과 제3 뱅크(BANK 3)의 제2 라이트 매치라인(ML_r2)을 서로 연결하고, 제2 뱅크(BANK 2)의 제5 레프트 매치라인(ML_15)과 제3 뱅크(BANK 3)의 제4 라이트 매치라인(ML_r4)을 서로 연결하며, 제2 뱅크(BANK 2)의 제6 레프트 매치라인(ML_16)과 제3 뱅크(BANK 3)의 제5 라이트 매치라인(ML_r5)을 서로 연결할 것이다.
- [0129] 상술한 바와 같이, 본 출원의 실시 예에 따른 BNN 장치(200)는 입력 특징맵의 데이터를 중첩 없이 메모리 셀 어레이에 저장된 상태에서 복수의 필터 윈도우들에 대한 BNN 연산 동작을 병렬적으로 수행할 수 있다. 또한, 필터가 오른쪽으로 슬라이딩 되는 경우, 매치라인 스위치(211_3)를 스위칭시키는 동작만으로 BNN 연산 동작을 용이하게 수행할 수 있다.
- [0130] 한편, 필터가 아래 방향으로 슬라이딩 되는 경우, 메모리 셀 어레이에 저장된 데이터 중 일부를 재-프로그램하는 동작이 수행될 수 있다. 예를 들어, 도 21에 도시된 바와 같이, 필터가 아래 방향으로 "1"만큼 슬라이딩 된다고 가정하자. 이 경우, 제1 필터 윈도우(FW1)의 제1 컬럼에는 (2,1) 내지 (6,1)의 데이터가 대응되며, 이에 따라 제1 뱅크(BANK 1)의 제1 로우에는 (2,1) 내지 (6,1)의 데이터가 저장되어야 한다. 이 때, (2,1) 내지 (5,1)은 이미 제1 뱅크(BANK 1)에 저장되어 있으므로, (7,1)에 대응하는 데이터가 제1 뱅크(BANK 1)의 제1 로우

에 재-프로그램 될 것이다.

[0131] 이와 유사하게, 제1 내지 제5 뱅크들(BANK 1~BANK 5)에 저장된 데이터들 중 일부 데이터가 필터의 아래 방향으로의 슬라이딩에 따라 재-프로그램될 수 있다.

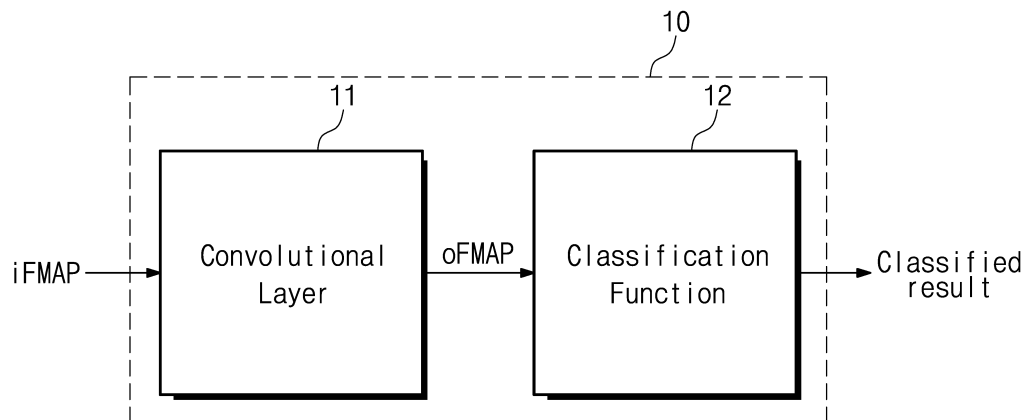
[0132] 이상과 같이 본 발명에 대해서 예시한 도면을 참조로 하여 설명하였으나, 본 명세서에 개시된 실시 예와 도면에 의해 본 발명이 한정되는 것은 아니며, 본 발명의 기술사상의 범위 내에서 통상의 기술자에 의해 다양한 변형이 이루어질 수 있음은 자명하다. 아울러 앞서 본 발명의 실시 예를 설명하면서 본 발명의 구성에 따른 작용 효과를 명시적으로 기재하여 설명하지 않았을 지라도, 해당 구성에 의해 예측 가능한 효과 또한 인정되어야 함은 당연하다.

부호의 설명

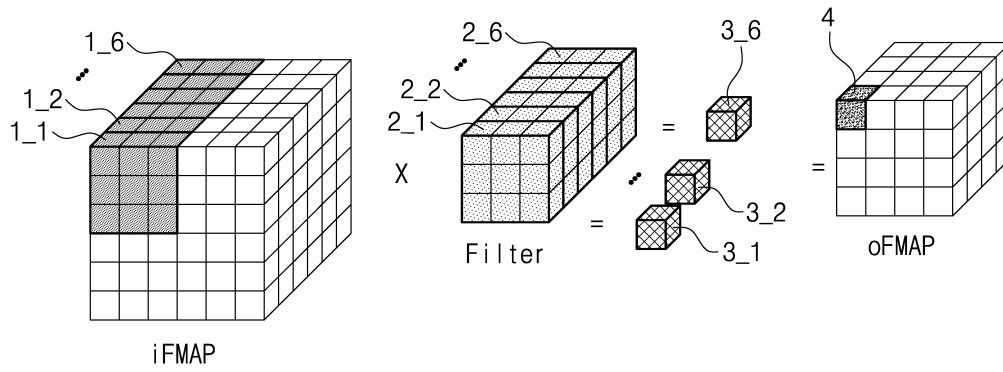
[0133] 10, 20, 100: 인공 신경망 장치
11: 컨볼루션층 레이어
12: 분류 함수
21, 110: 연산부
22, 120: 분류부
111: XNOR 연산 회로
111_1: 프리차지 회로
111_2: 메모리 셀 어레이
112: 이진화 회로
112_1: 딜레이 버퍼
112_2: TDA

도면

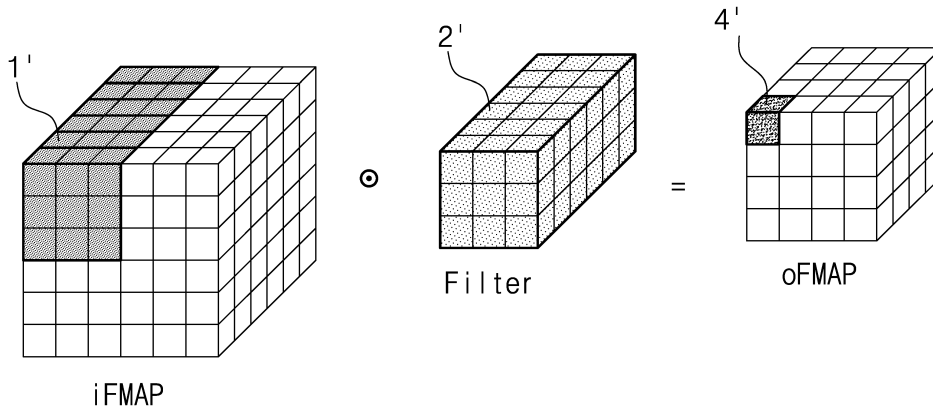
도면1



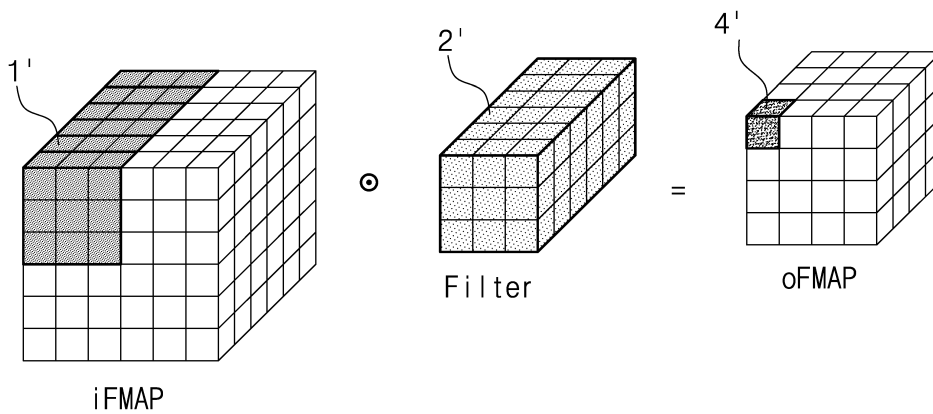
도면2



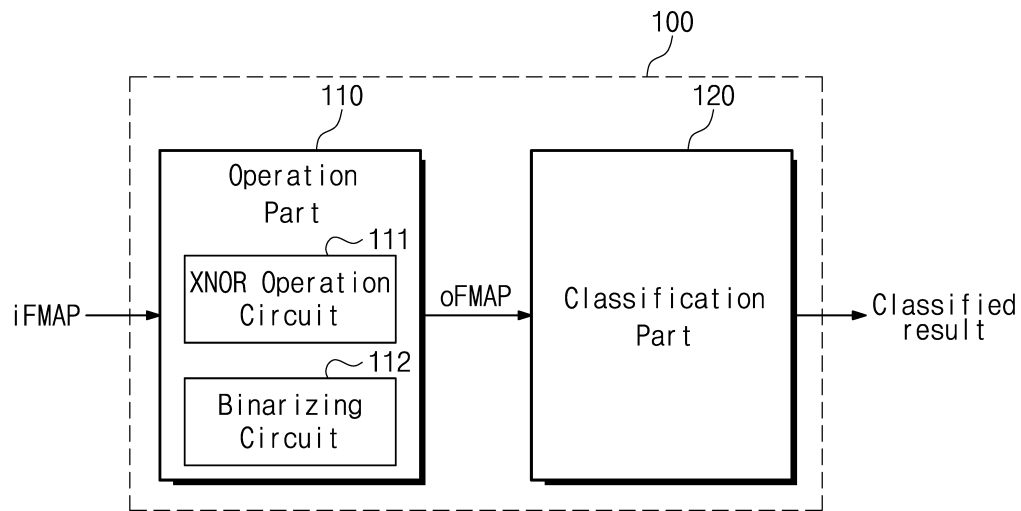
도면3



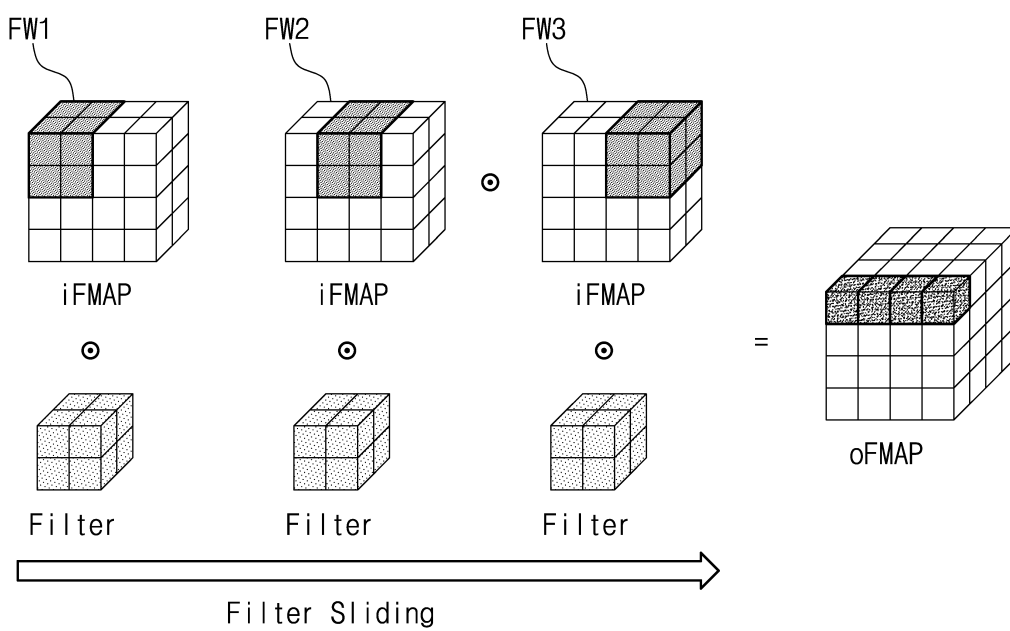
도면4



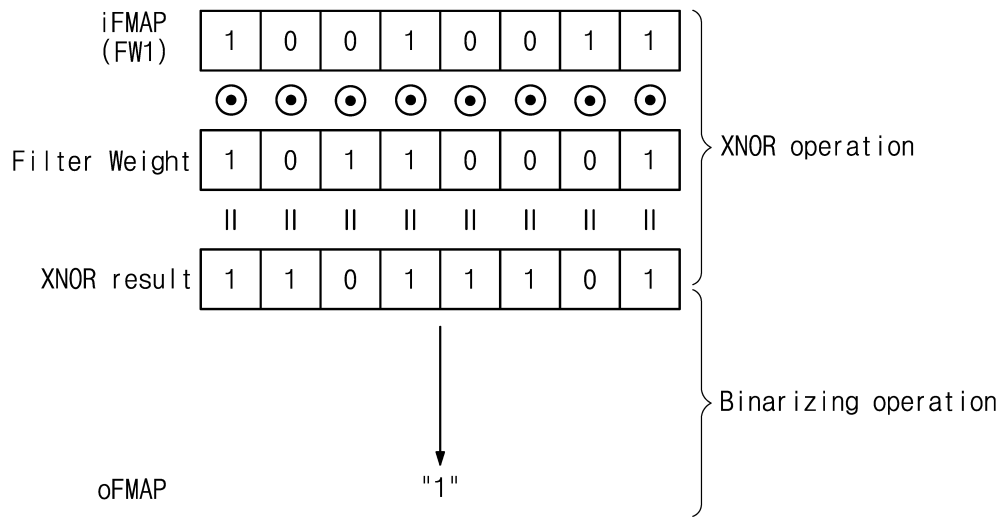
도면5



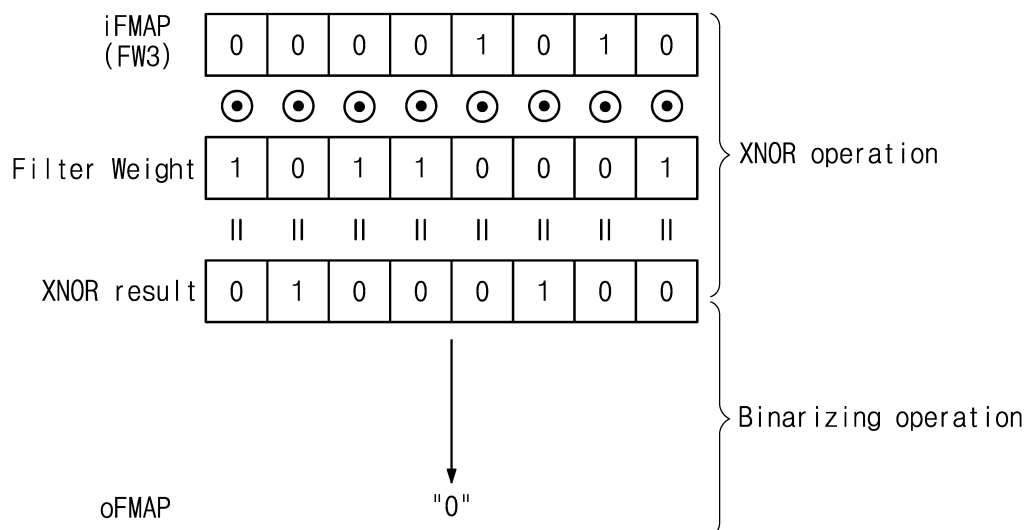
도면6



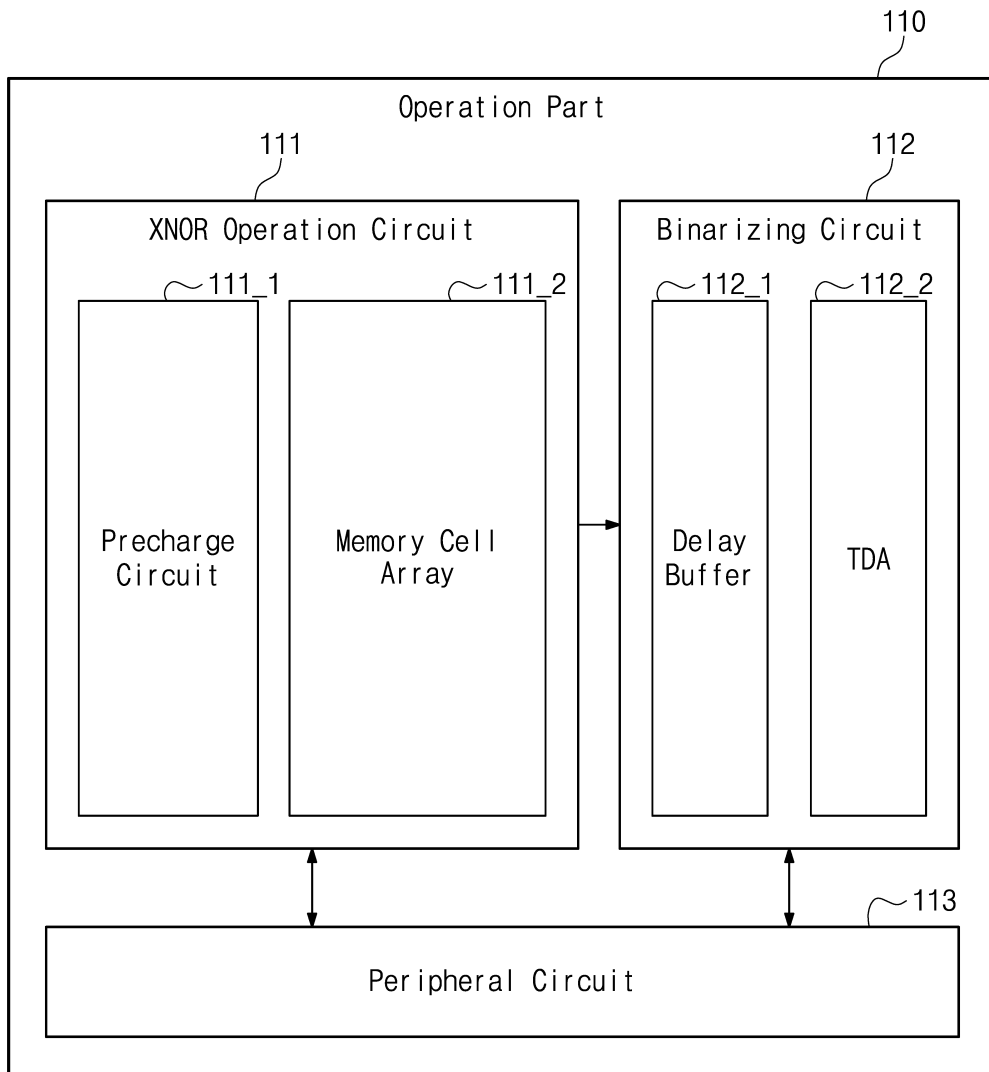
도면7a



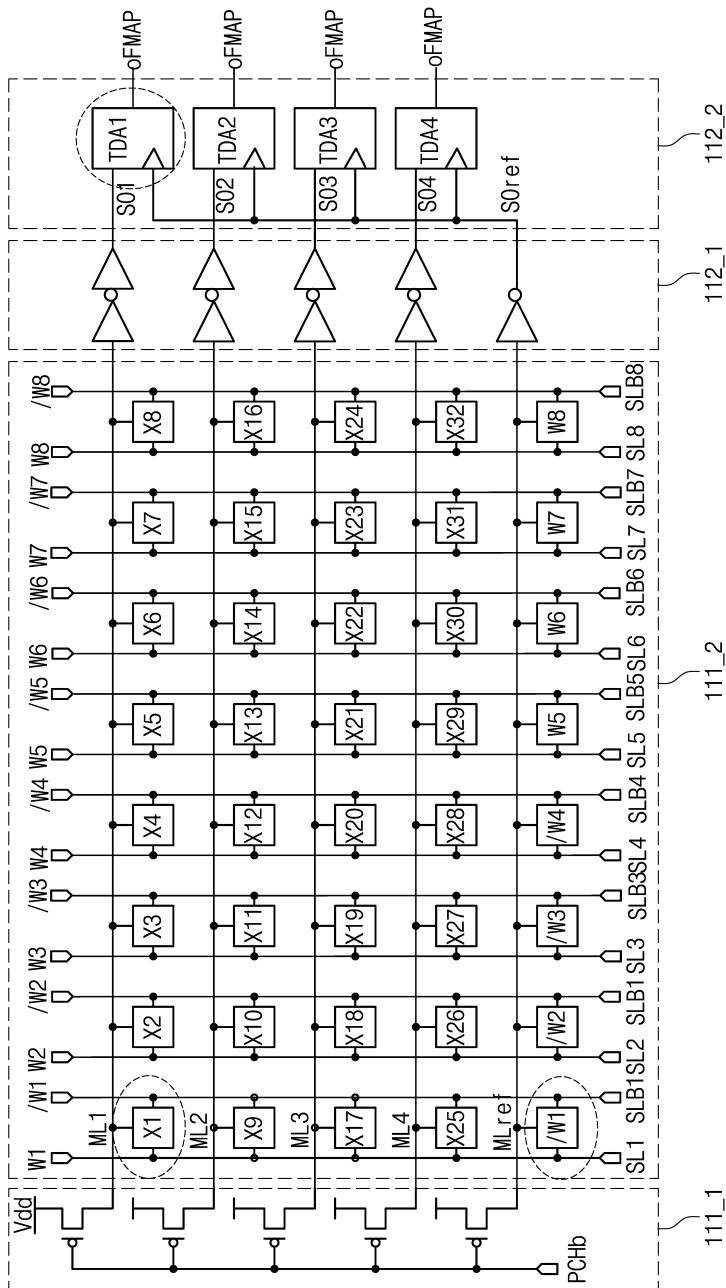
도면7b



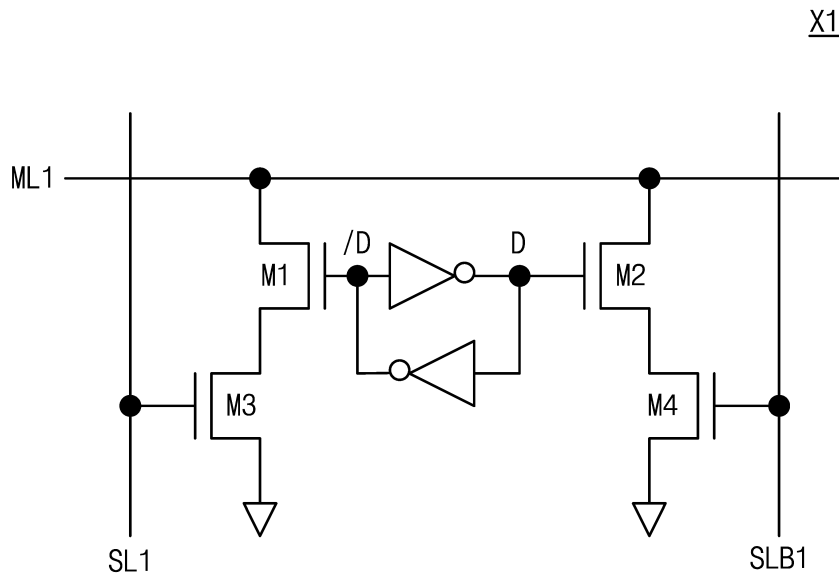
도면8



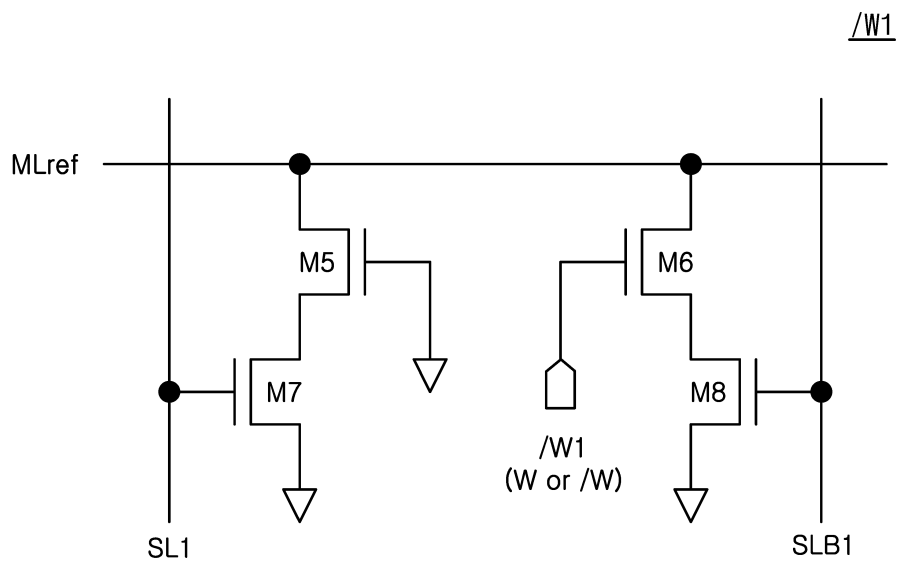
도면9



도면10

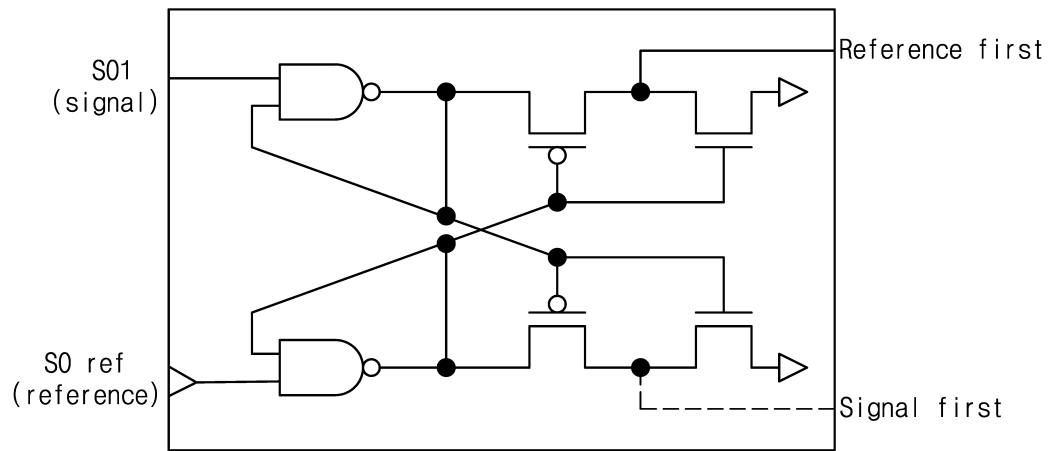


도면11

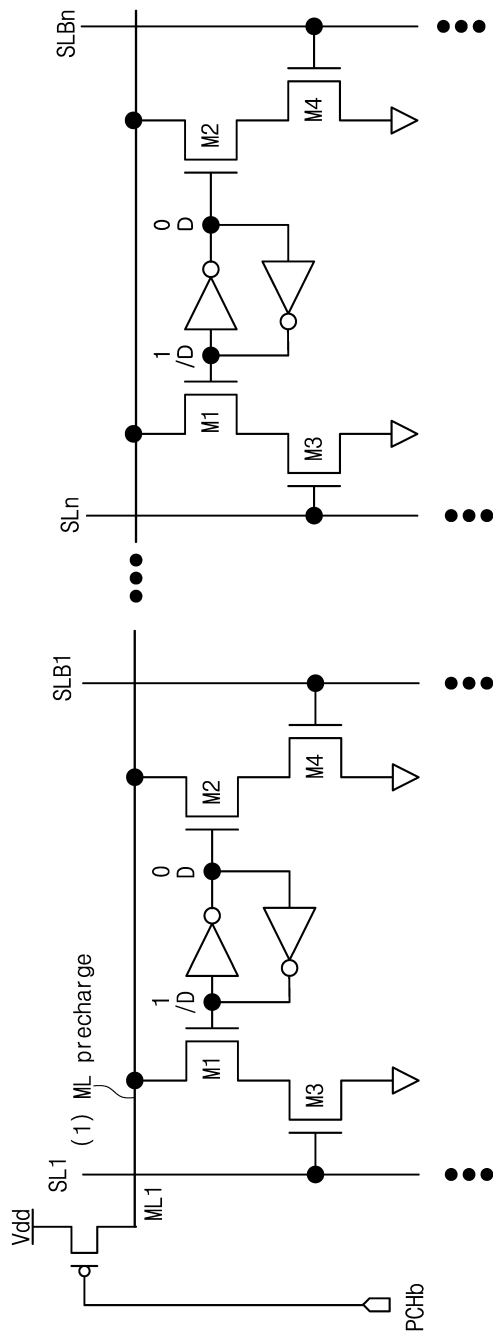


도면12

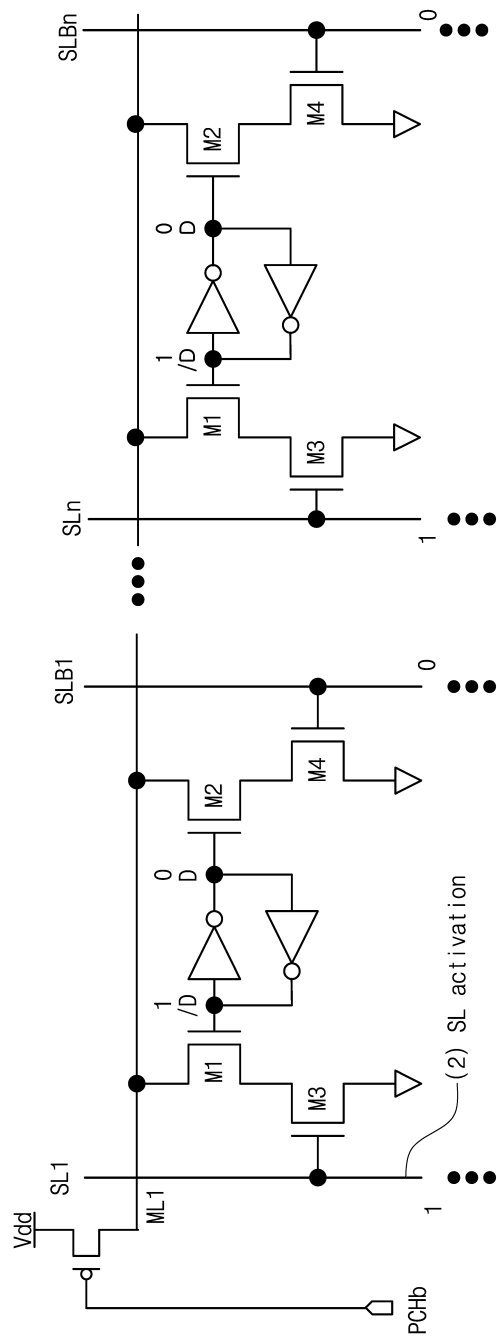
TDA1



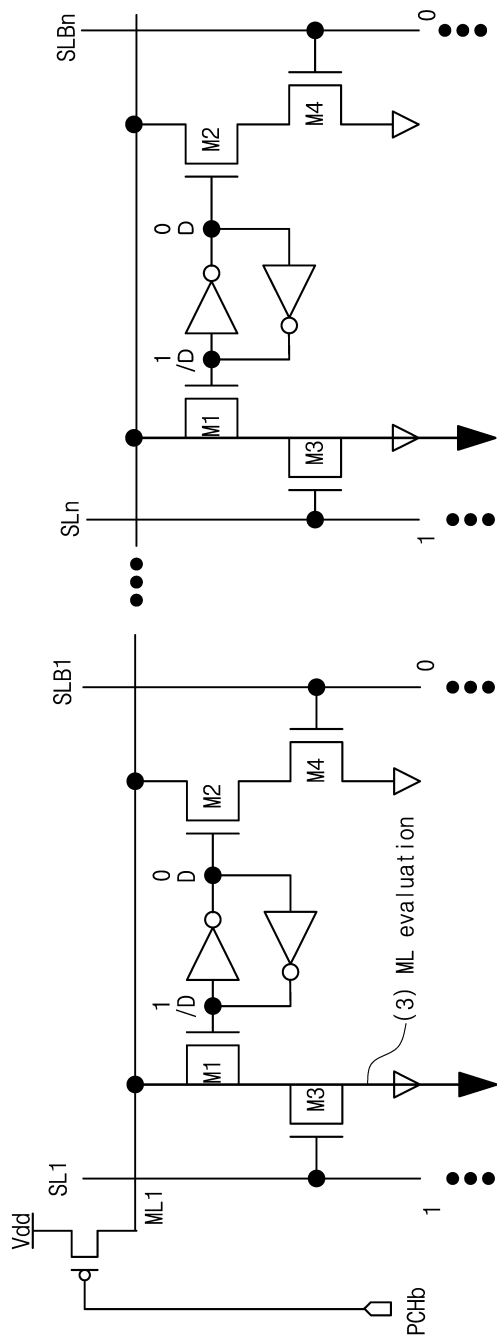
도면13a



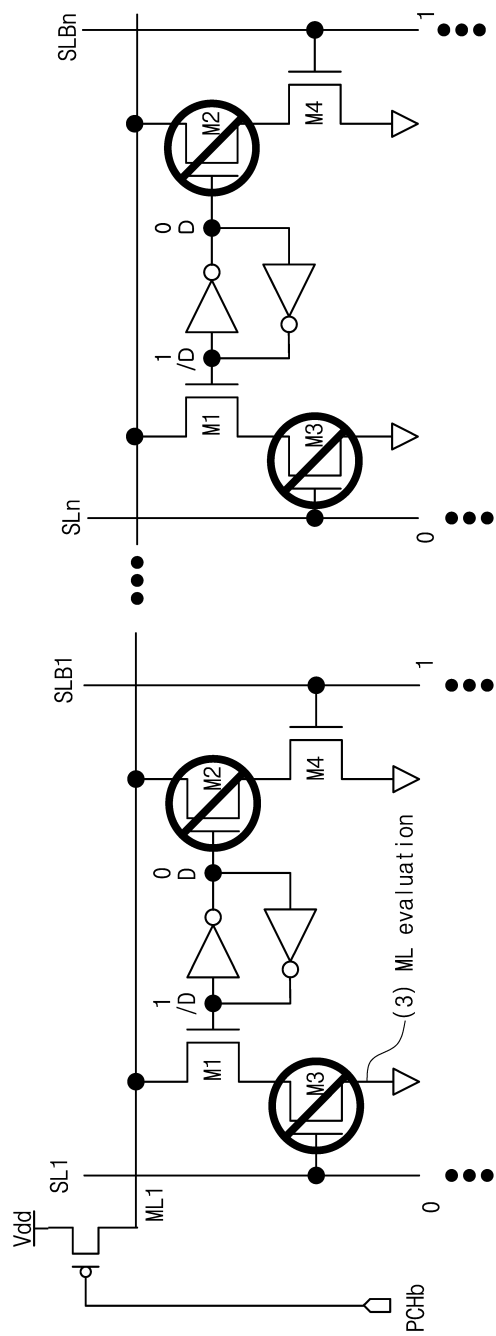
도면13b



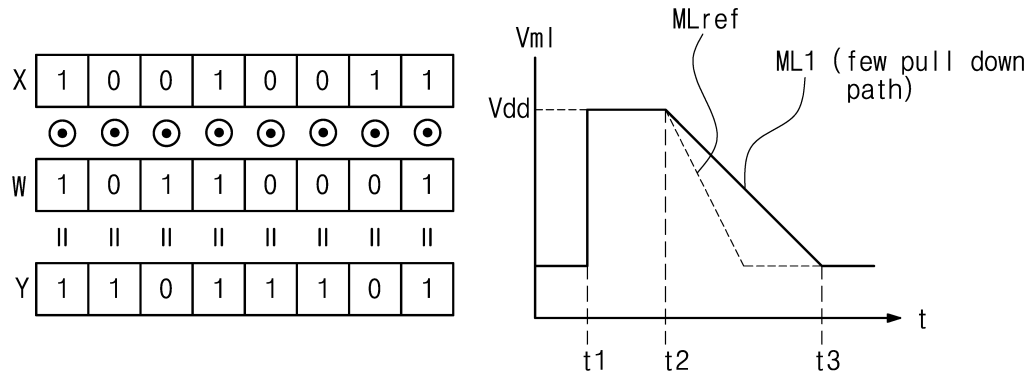
도면13c



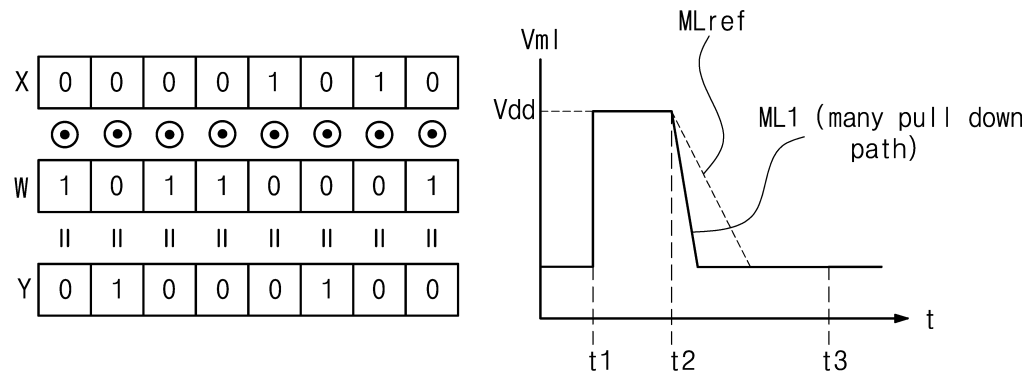
도면13d



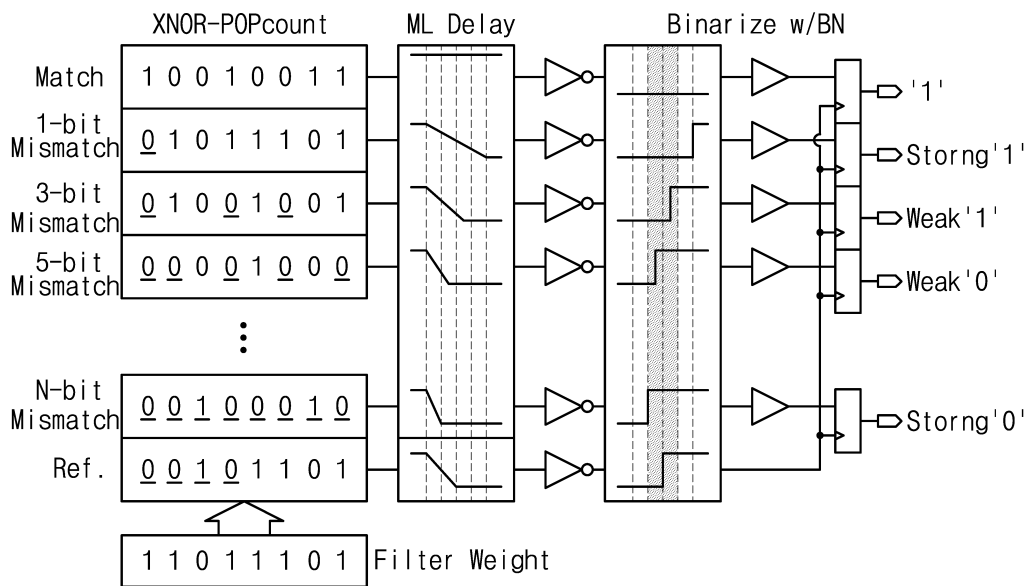
도면14a



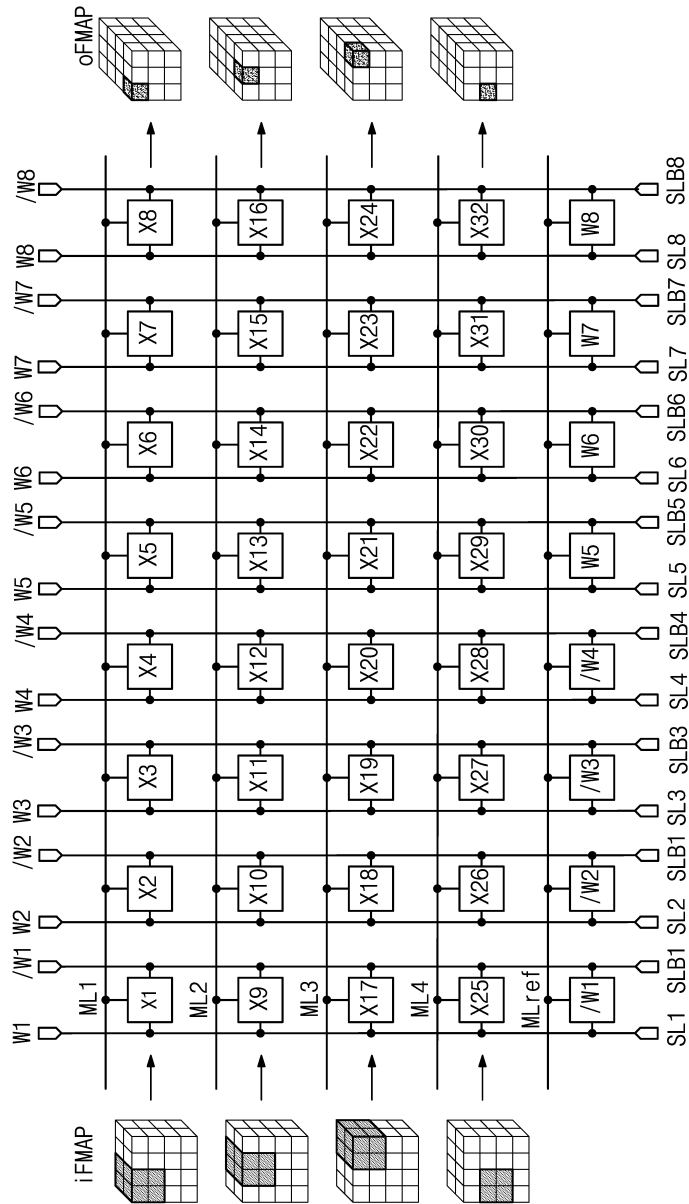
도면14b



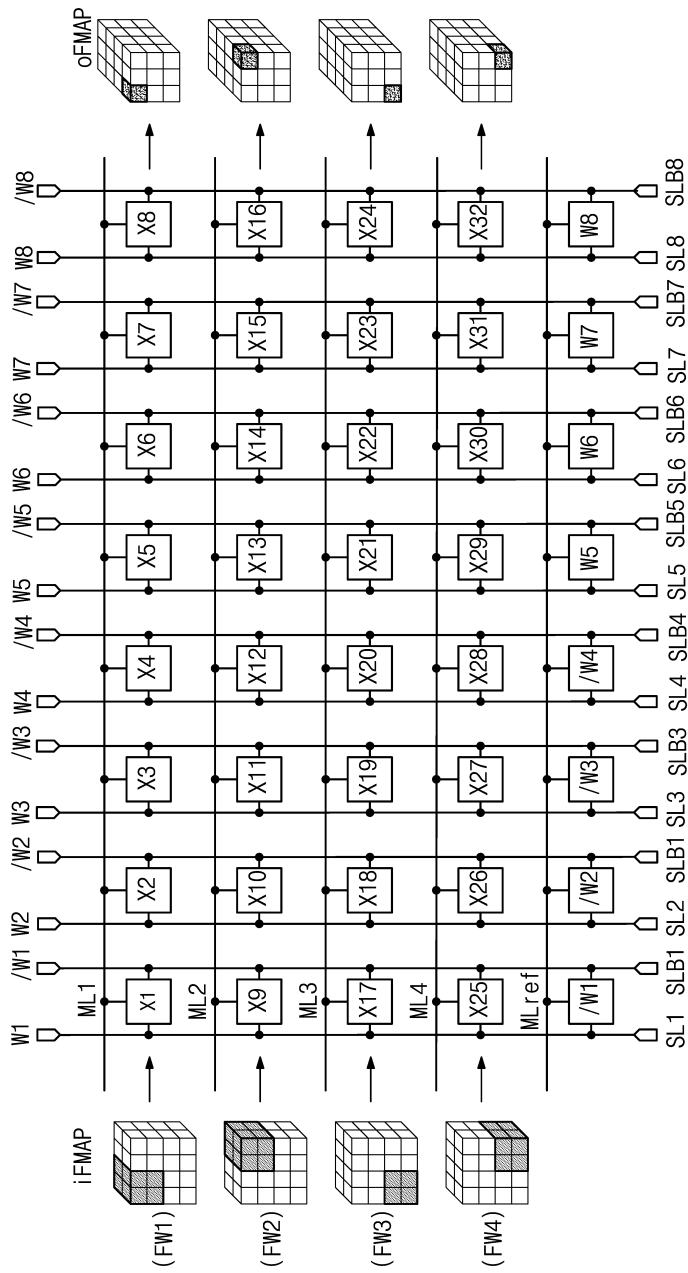
도면14c



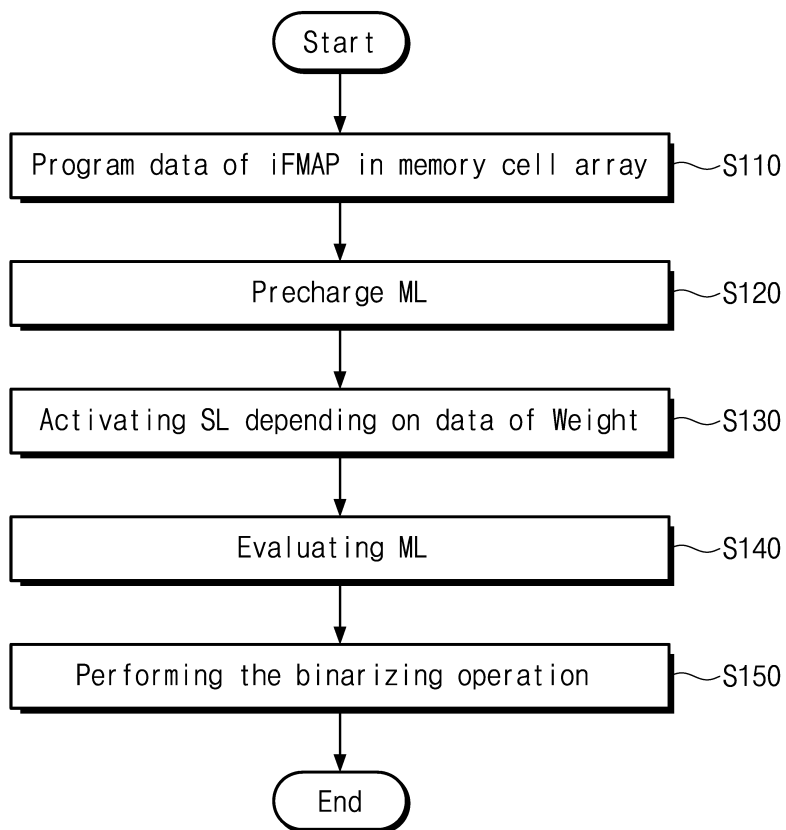
도면15



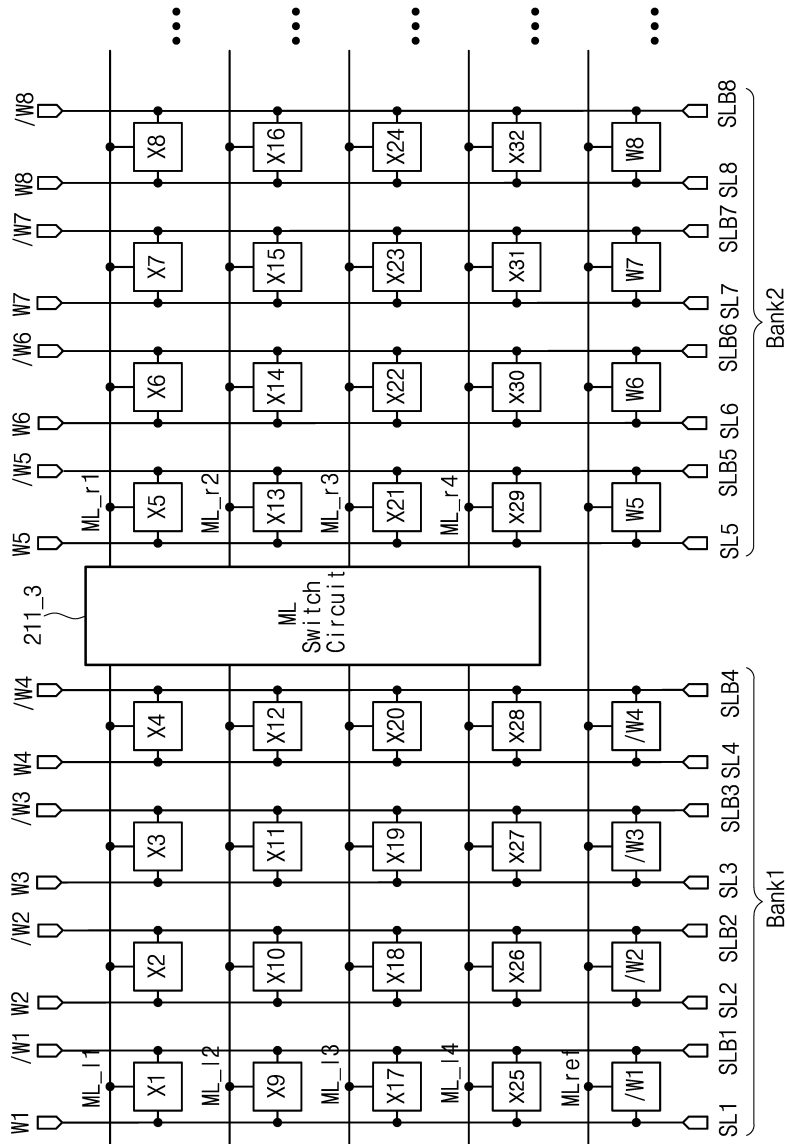
도면16



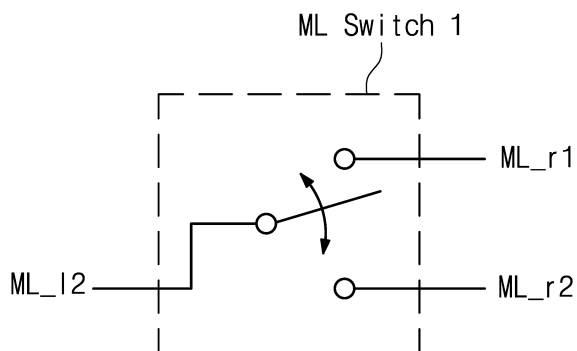
도면17



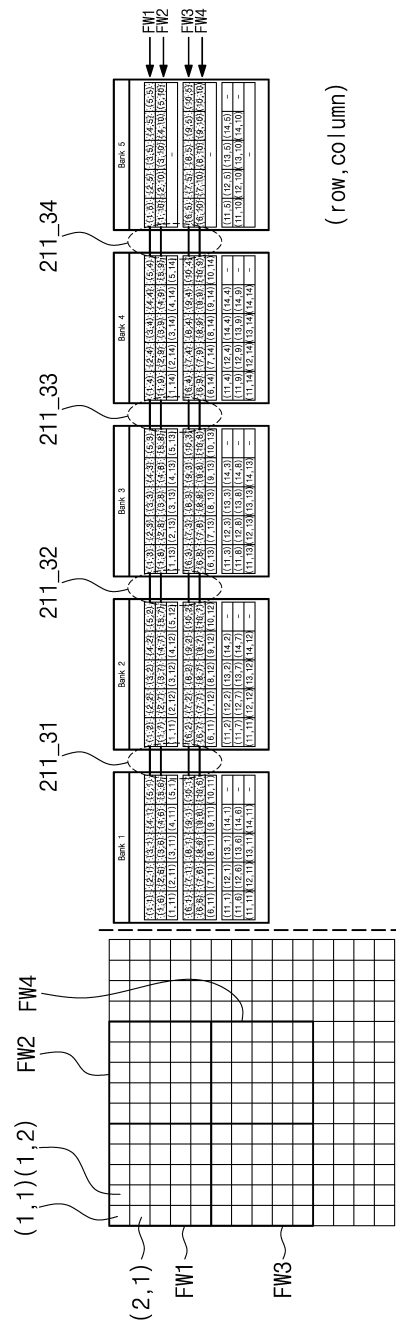
도면18



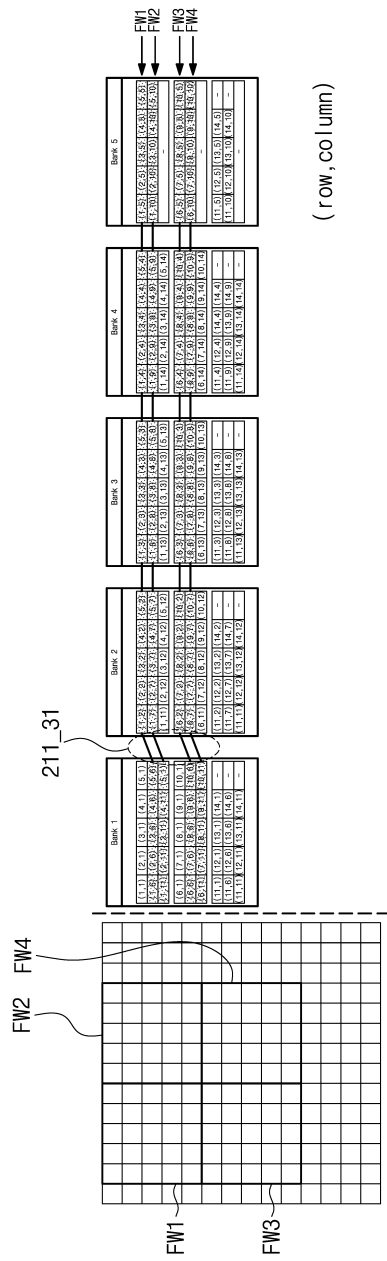
도면19



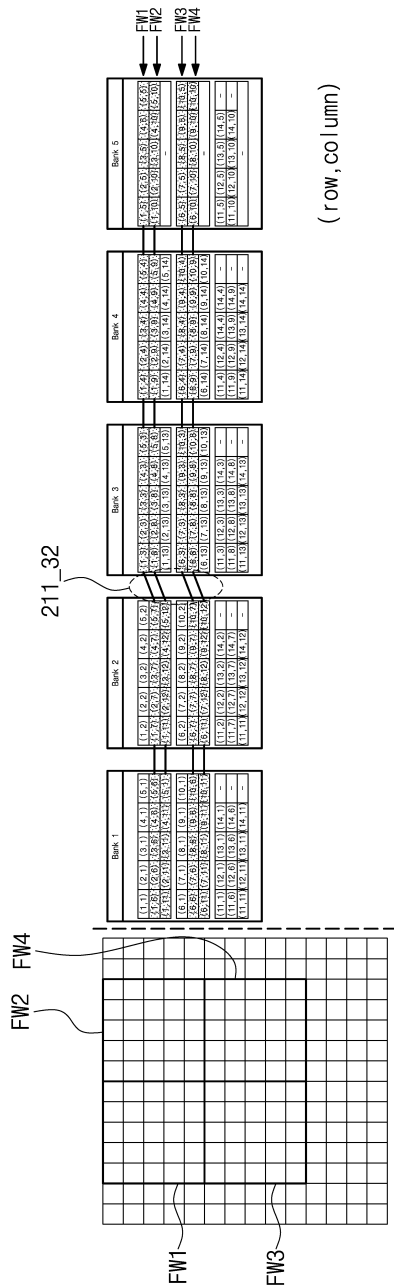
도면20a



도면20b



도면20c



도면21

