



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0078024
(43) 공개일자 2023년06월02일

(51) 국제특허분류(Int. Cl.)
H10B 63/00 (2023.01) G06N 3/063 (2023.01)
H10N 70/00 (2023.01)
(52) CPC특허분류
H10B 63/30 (2023.02)
G06N 3/063 (2013.01)
(21) 출원번호 10-2021-0165347
(22) 출원일자 2021년11월26일
심사청구일자 2021년11월26일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
이승현
경기도 성남시 분당구 서판교로 29, 912동 403호 (판교동, 판교원마을)
바틸백 알림카놀리
경기도 수원시 영통구 매영로 412, 511호 (영통동)
현기환
경기도 수원시 영통구 반달로 60-2, 103호 (영통동)
(74) 대리인
김연권

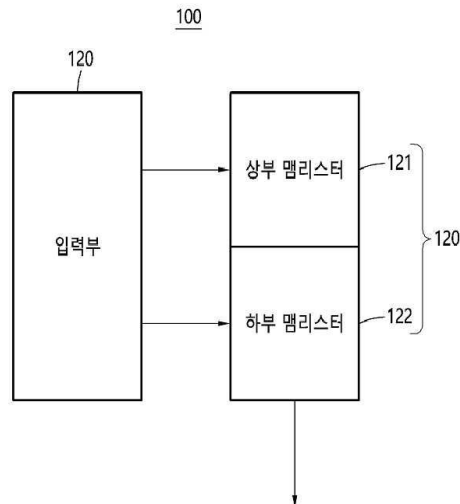
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 삼항 정밀도 XOR 논리 연산을 기반으로 하는 반도체 소자 및 이를 포함하는 뉴로모픽 컴퓨팅 시스템

(57) 요약

본 발명은 삼항 정밀도 XOR 논리 연산을 기반으로 하는 반도체 소자 및 이를 포함하는 뉴로모픽 컴퓨팅 시스템에 관한 것으로서, 일실시예에 따른 반도체 소자는 상부 입력신호와 하부 입력신호를 수신하는 입력부 및 하부 메모리스터와 상기 하부 메모리스터에 적층된 상부 메모리스터를 포함하고, 상기 상부 메모리스터로 입력되는 상기 상부 입력신호와 상기 하부 메모리스터로 입력되는 상기 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행하는 논리 연산부를 포함한다.

대 표 도 - 도1



(52) CPC특허분류

H10N 70/20 (2023.02)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711128470
과제번호	2020M3F3A2A01085755
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	차세대지능형반도체기술개발(R&D)
연구과제명	ReRAM 기반 near-memory 비트벡터 연산기를 위한 ReRAM 셀 어레이 개발
기 여 율	1/2
과제수행기관명	경희대학교 산학협력단
연구기간	2020.07.01 ~ 2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711133798
과제번호	10085646
부처명	과학기술정보통신부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	미래소자 원천기술개발사업
연구과제명	3D 메모리스터 어레이 결점 극복을 위한 뉴로모픽 시스템
기 여 율	1/2
과제수행기관명	경희대학교 산학협력단
연구기간	2017.12.01 ~ 2021.12.31

명세서

청구범위

청구항 1

상부 입력신호와 하부 입력신호를 수신하는 입력부 및

하부 맴리스터와 상기 하부 맴리스터에 적층된 상부 맴리스터를 포함하고, 상기 상부 맴리스터로 입력되는 상기 상부 입력신호와 상기 하부 맴리스터로 입력되는 상기 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행하는 논리 연산부

를 포함하는 반도체 소자.

청구항 2

제1항에 있어서,

상기 논리 연산부는,

상기 상부 맴리스터와 상기 하부 맴리스터 각각의 저항 상태에 따라 가중치 값이 결정되며, 상기 상부 입력신호, 상기 하부 입력신호 및 상기 가중치 값에 기초하여 상기 XNOR 연산을 수행하는

반도체 소자.

청구항 3

제2항에 있어서,

상기 논리 연산부는,

상기 상부 맴리스터가 저저항 상태(LRS)이고 상기 하부 맴리스터가 고저항 상태(HRS)이면 상기 가중치 값이 '1'로 결정되고,

상기 상부 맴리스터가 고저항 상태이고 상기 하부 맴리스터가 저저항 상태이면 상기 가중치 값이 '-1'로 결정되며,

상기 상부 맴리스터가 고저항 상태이고 상기 하부 맴리스터가 고저항 상태이면 상기 가중치 값이 '0'으로 결정되는

반도체 소자.

청구항 4

제2항에 있어서,

상기 논리 연산부는,

상기 상부 입력 신호의 값이 '1'이고 상기 하부 입력 신호의 값이 '-1'이며 상기 가중치의 값이 '1'이면, 상기 XNOR 연산의 수행 결과로 '1'을 출력하고,

상기 상부 입력 신호의 값이 '1'이고 상기 하부 입력 신호의 값이 '-1'이며 상기 가중치의 값이 '-1'이면, 상기 XNOR 연산의 수행 결과로 '-1'을 출력하는

반도체 소자.

청구항 5

제2항에 있어서,

상기 논리 연산부는,

상기 상부 입력 신호의 값이 '-1'이고 상기 하부 입력 신호의 값이 '1'이며 상기 가중치의 값이 '1'이면, 상기

XNOR 연산의 수행 결과로 '-1'을 출력하고,

상기 상부 입력 신호의 값이 '-1'이고 상기 하부 입력 신호의 값이 '1'이며 상기 가중치의 값이 '-1'이면, 상기 XNOR 연산의 수행 결과로 '1'을 출력하는

반도체 소자.

청구항 6

제2항에 있어서,

상기 논리 연산부는,

상기 상부 입력 신호와 상기 하부 입력 신호의 차이 값이 '0'이거나, 상기 가중치 값이 '0'이면, 상기 XNOR 연산의 수행 결과로 '0'을 출력하는

반도체 소자.

청구항 7

제1항에 있어서,

출력 트랜지스터를 포함하고, 상기 출력 트랜지스터의 온/오프 동작에 기초하여 상기 XNOR 연산의 수행 결과를 출력하는 출력부

를 더 포함하는 반도체 소자.

청구항 8

적어도 하나의 상부 입력신호와 적어도 하나의 하부 입력신호를 수신하는 입력부;

상기 적어도 하나의 상부 입력신호와 상기 적어도 하나의 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행하는 복수의 시냅스를 구비하는 시냅스 어레이 및

상기 XNOR 연산의 수행 결과를 출력하는 출력부

를 포함하고,

상기 복수의 시냅스 각각은,

하부 맴리스터 및 상기 하부 맴리스터에 적층된 상부 맴리스터를 포함하고,

상기 상부 맴리스터를 통해 상기 적어도 하나의 상부 입력신호 중 대응되는 어느 하나의 상부 입력신호를 수신하고, 상기 하부 맴리스터를 통해 상기 적어도 하나의 하부 입력신호 중 대응되는 어느 하나의 하부 입력신호를 수신하는

뉴로모픽 컴퓨팅 시스템.

청구항 9

제8항에 있어서,

상기 복수의 시냅스 각각은,

상기 상부 맴리스터와 상기 하부 맴리스터 각각의 저항 상태에 따라 가중치 값이 결정되며, 상기 어느 하나의 상부 입력신호, 상기 어느 하나의 하부 입력신호 및 상기 가중치 값에 기초하여 상기 XNOR 연산을 수행하는

뉴로모픽 컴퓨팅 시스템.

청구항 10

제8항에 있어서,

상기 출력부는,

출력 트랜지스터를 포함하고, 상기 출력 트랜지스터의 온/오프 동작에 기초하여 상기 XNOR 연산의 수행 결과를

출력하는

뉴로모픽 컴퓨팅 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 반도체 소자 및 이를 포함하는 뉴로모픽 컴퓨팅 시스템에 관한 것으로, 보다 상세하게는 반도체 소자를 이용하여 삼항 정밀도 XOR 논리 연산을 수행하는 기술적 사상에 관한 것이다.

배경 기술

[0002] 기존 폰-노이만 구조는 메모리와 연산부가 분리되어 있어 심층 신경망 네트워크(DNN)를 집적회로에 적용하는 것이 상당히 제한되며, 이에 따라 뉴로모픽 컴퓨팅 구조는 비휘발성 메모리(eNVM)가 최적의 메모리 소자로서 활용되는 메모리 중심 아키텍처로 전환될 것으로 전망되고 있다.

[0003] 특히, 비휘발성 메모리 중 저항 메모리(RRAM)는 높은 집적도, 빠른 동작 속도, 낮은 소비 전력, 높은 CMOS와의 호환성, 뛰어난 반복성/내구성을 갖추어 메모리 중심 아키텍처에 가장 적합한 메모리로 평가되고 있다.

[0004] 뉴로모픽 컴퓨팅과 관련한 최근 연구에 따르면, 32 비트 부동 소수점 방식에서 1 비트 이진 정밀도 (-1, +1)로 양자화 함으로써 모델의 사이즈의 감소와 학습 정확도의 악화 사이의 최적의 타협점에 도달하였다.

[0005] 따라서, 벡터-행렬 곱 정밀도는 연산 XNOR과 비트 수 카운터 동작으로 더욱 향상된 덧셈/뺄셈 연산으로 단순화될 수 있고, 추론 프로세스가 진행되는 동안 에너지 효율, 연산 속도를 향상시키고 연산 시 소비하는 에너지를 줄일 수 있게 되었으며, 이러한 심층 신경망 모델은 특정 가중치 소거(weight-pruning)를 활용한 최적화를 통해 삼항 정밀도 (-1, 0, +1)를 구현할 수 있다.

[0006] 한편, 기존 크로스-포인트 아키텍처에서는 적층 수에 비례하여 비트 당 소비 에너지 효율과 메모리 어레이의 적층 집적도를 제한하는 적층 수가 빠르게 증가하고 있으며, 이와 달리 3차원 수직형 저항 메모리(3D VRRAM)는 최적의 비트 당 소비 에너지 효율성을 갖춘 고집적 어레이로 구현될 수 있어 그 활용성이 주목을 받고 있다. 구체적으로, VRRAM은 높은 리드/라이트(Read/Write) 마진, 낮은 에너지 소비, 병렬적 프로그래밍 등의 이점을 보이고 있다.

[0007] 그러나, 기존의 크로스-포인트 어레이가 접합에서의 키르히호프 법칙으로 인해 가중 합 연산을 쉽게 구현할 수 있기 때문에 뇌 신경 모사 컴퓨팅의 주요 동작 중 하나인 벡터-행렬 곱 연산(VMM)에 유리한 반면, 3D VRRAM은 워드 평면(word-plane, WP)의 모놀리식(monolithic) 구조로 인해 입력 벡터의 수가 제한되므로 VMM 연산을 방해한다는 문제가 있다.

[0008] 이를 해결하기 위해, VRRAM 어레이에서 가중 합 연산을 수행하기 위한 방법으로써, 기동 전극을 입력 뉴런으로 WP를 출력 뉴런으로 1T1R 구조를 형성하는 방법이 제시되었으나, 이러한 구조에서는 출력 뉴런의 개수가 적층된 WP의 개수에 크게 의존하며, 결국 예칭 시의 종횡비에 따라 달라진다는 한계가 존재한다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 한국등록특허 제10-1973678호, "멤리스터로 구현된 순차적 메모리 회로 및 그 구동방법"

(특허문헌 0002) 한국공개특허 제2019-0046603호, "뉴로모픽 컴퓨팅에 사용 가능한 변화 저항성 자기 접합형 XNOR 셀을 제공하는 방법 및 시스템"

발명의 내용

해결하려는 과제

[0010] 본 발명은 적층된 복수의 멤리스터를 이용하여 XNOR 연산을 수행함으로써, 높은 메모리 집적도, 낮은 소비 전력 및 저 정밀도의 심층 신경망을 구현할 수 있는 메모리 소자 및 이를 포함하는 뉴로모픽 컴퓨팅 시스템을 제공하

고자 한다.

[0011] 또한, 본 발명은 적층된 복수의 맴리스터를 이용하는 XNOR 아키텍처를 VRRAM에 적용하여 개별 메모리 셀과 전체적인 어레이의 성능을 향상시킬 수 있는 메모리 소자 및 이를 포함하는 뉴로모픽 컴퓨팅 시스템을 제공하고자 한다.

과제의 해결 수단

[0012] 본 발명의 일실시예에 따른 반도체 소자는 상부 입력신호와 하부 입력신호를 수신하는 입력부 및 하부 맴리스터와 하부 맴리스터에 적층된 상부 맴리스터를 포함하고, 상부 맴리스터로 입력되는 상부 입력신호와 하부 맴리스터로 입력되는 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행하는 논리 연산부를 포함할 수 있다.

[0013] 일측에 따르면, 논리 연산부는 상부 맴리스터와 하부 맴리스터 각각의 저항 상태에 따라 가중치 값이 결정되며, 상부 입력신호, 하부 입력신호 및 가중치 값에 기초하여 XNOR 연산을 수행할 수 있다.

[0014] 일측에 따르면, 논리 연산부는 상부 맴리스터가 저저항 상태(LRS)이고 하부 맴리스터가 고저항 상태(HRS)이면 가중치 값이 '1'로 결정되고, 상부 맴리스터가 고저항 상태이고 하부 맴리스터가 저저항 상태이면 가중치 값이 '-1'로 결정되며, 상부 맴리스터가 고저항 상태이고 하부 맴리스터가 고저항 상태이면 가중치 값이 '0'으로 결정될 수 있다.

[0015] 일측에 따르면, 논리 연산부는 상부 입력 신호의 값이 '1'이고 하부 입력 신호의 값이 '-1'이며 가중치의 값이 '1'이면, XNOR 연산의 수행 결과로 '1'을 출력하고, 상부 입력 신호의 값이 '1'이고 하부 입력 신호의 값이 '-1'이며 가중치의 값이 '-1'이면, XNOR 연산의 수행 결과로 '-1'을 출력할 수 있다.

[0016] 일측에 따르면, 논리 연산부는 상부 입력 신호의 값이 '-1'이고 하부 입력 신호의 값이 '1'이며 가중치의 값이 '1'이면, XNOR 연산의 수행 결과로 '-1'을 출력하고, 상부 입력 신호의 값이 '-1'이고 하부 입력 신호의 값이 '1'이며 가중치의 값이 '-1'이면, XNOR 연산의 수행 결과로 '1'을 출력할 수 있다.

[0017] 일측에 따르면, 논리 연산부는 상부 입력 신호와 하부 입력 신호의 차이 값이 '0'이거나, 가중치 값이 '0'이면, XNOR 연산의 수행 결과로 '0'을 출력할 수 있다.

[0018] 일측에 따르면, 반도체 소자는 출력 트랜지스터를 구비하고 출력 트랜지스터의 온/오프 동작에 기초하여 XNOR 연산의 수행 결과를 출력하는 출력부를 더 포함할 수 있다.

[0019] 본 발명의 일실시예에 따른 뉴로모픽 컴퓨팅 시스템은 적어도 하나의 상부 입력신호와 적어도 하나의 하부 입력신호를 수신하는 입력부, 적어도 하나의 상부 입력신호와 적어도 하나의 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행하는 복수의 시냅스를 구비하는 시냅스 어레이 및 XNOR 연산의 수행 결과를 출력하는 출력부를 포함할 수 있다.

[0020] 여기서, 복수의 시냅스 각각은 하부 맴리스터 및 하부 맴리스터에 적층된 상부 맴리스터를 포함하고 상부 맴리스터를 통해 적어도 하나의 상부 입력신호 중 대응되는 어느 하나의 상부 입력신호를 수신하고, 하부 맴리스터를 통해 적어도 하나의 하부 입력신호 중 대응되는 어느 하나의 하부 입력 신호를 수신할 수 있다.

[0021] 일측에 따르면, 복수의 시냅스 각각은 상부 맴리스터와 하부 맴리스터 각각의 저항 상태에 따라 가중치 값이 결정되며, 어느 하나의 상부 입력신호, 어느 하나의 하부 입력신호 및 가중치 값에 기초하여 XNOR 연산을 수행할 수 있다.

[0022] 일측에 따르면, 출력부는 출력 트랜지스터를 포함하고, 출력 트랜지스터의 온/오프 동작에 기초하여 XNOR 연산의 수행 결과를 출력할 수 있다.

발명의 효과

[0023] 일실시예에 따르면, 본 발명은 적층된 복수의 맴리스터를 이용하여 XNOR 연산을 수행함으로써, 높은 메모리 집적도, 낮은 소비 전력 및 저 정밀도의 심층 신경망을 구현할 수 있다.

[0024] 일실시예에 따르면, 본 발명은 적층된 복수의 맴리스터를 이용하는 XNOR 아키텍처를 VRRAM에 적용하여 개별 메모리 셀과 전체적인 어레이의 성능을 향상시킬 수 있다.

도면의 간단한 설명

- [0025] 도 1은 일실시예에 따른 반도체 소자를 설명하는 도면이다.
- 도 2는 일실시예에 따른 뉴로모픽 컴퓨팅 시스템을 설명하는 도면이다.
- 도 3은 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 구현예를 설명하는 도면이다.
- 도 4는 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 구동 과정을 설명하는 도면이다.
- 도 5a 내지 도 5c는 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 동작 특성을 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.
- [0027] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0028] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 용어들에 의해 한정되어서는 안 된다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만, 예를 들면 본 발명의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0029] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의 관계를 설명하는 표현들, 예를 들면 "~사이에"와 "바로~사이에" 또는 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.
- [0030] 본 명세서에서 사용한 용어는 단지 특정한 실시예들을 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0031] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0033] 이하, 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.
- [0035] 도 1은 일실시예에 따른 반도체 소자를 설명하는 도면이다.
- [0036] 도 1을 참조하면, 일실시예에 따른 반도체 소자(100)는 적층된 복수의 맴리스터를 이용하여 XNOR 연산을 수행함으로써, 높은 메모리 집적도, 낮은 소비 전력 및 저 정밀도의 심층 신경망을 구현할 수 있다.
- [0037] 또한, 반도체 소자(100)는 적층된 복수의 맴리스터를 이용하는 XNOR 아키텍처를 VRRAM에 적용하여 개별 메모리

셀과 전체적인 어레이의 성능을 향상시킬 수 있다.

- [0038] 이를 위해, 반도체 소자(100)는 입력부(110) 및 논리 연산부(120)를 포함할 수 있다.
- [0039] 일실시예에 따른 입력부(110)는 상부 입력신호와 하부 입력신호를 수신할 수 있다.
- [0040] 예를 들면, 상부 입력신호 및 하부 입력신호는 입력값 '1'에 대응되는 기설정된 크기의 양의 리드(read) 펄스, 입력값 '-1'에 대응되는 기설정된 크기의 음의 리드 펄스 중 적어도 하나일 수 있다.
- [0041] 또한, 입력부(110)를 통해 펄스가 인가되지 않거나 기설정된 임계 크기 이하로 매우 작은 크기의 펄스가 인가되는 경우는 입력값 '0'에 대응될 수 있다. 일례로, 반도체 소자(100)에 대응되는 수직 트랜지스터를 오프(off)하면 입력값 '0'을 구현할 수 있다.
- [0042] 일측에 따르면, 반도체 소자(100)는 저항 메모리 시스템에 구비되는 메모리 소자일 수 있으며, 입력부(110)는 상부 입력신호 및 하부 입력신호를 메모리 컨트롤러로부터 수신할 수 있다.
- [0043] 또한, 반도체 소자(100)는 뉴로모픽 컴퓨팅 시스템에 구비되는 시냅스일 수 있으며, 입력부(110)는 상부 입력신호 및 하부 입력신호를 시냅스의 전단에 연결된 적어도 하나의 뉴런으로부터 수신할 수도 있다.
- [0044] 일실시예에 따른 논리 연산부(120)는 하부 맴리스터(122)와, 하부 맴리스터(122)에 적층된 상부 맴리스터(121)를 포함할 수 있다.
- [0045] 또한, 논리 연산부(120)는 상부 맴리스터(121)로 입력되는 상부 입력신호와 하부 맴리스터(122)로 입력되는 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행할 수 있다.
- [0046] 일측에 따르면, 논리 연산부(120)는 상부 맴리스터(121)와 하부 맴리스터(122) 각각의 저항 상태에 따라 가중치 값이 결정되며, 상부 입력신호, 하부 입력신호 및 가중치 값에 기초하여 XNOR 연산을 수행할 수 있다.
- [0047] 다시 말해, 반도체 소자(100)의 시냅스 가중치 행렬은 상부 맴리스터(121)와 하부 맴리스터(122), 즉 두 개의 수직 레이어를 따라 구현되며, 반도체 소자(100)가 연결된 비트 라인(BL)에 흐르는 출력 전류는 상부 입력신호, 하부 입력신호 및 가중치 값에 따라 가변할 수 있다.
- [0048] 예를 들면, 논리 연산부(120)는 상부 맴리스터(121)가 저저항 상태(LRS)이고 하부 맴리스터(122)가 고저항 상태(HRS)이면 가중치 값이 '1'로 결정될 수 있다.
- [0049] 또한, 논리 연산부(120)는 상부 맴리스터(121)가 고저항 상태이고 하부 맴리스터(122)가 저저항 상태이면 가중치 값이 '-1'로 결정될 수 있다.
- [0050] 또한, 논리 연산부(120)는 상부 맴리스터(121)가 고저항 상태이고 하부 맴리스터(122)가 고저항 상태이면 가중치 값이 '0'으로 결정될 수 있다.
- [0051] 일측에 따르면, 논리 연산부(120)는 하기 표1에 기초하는 XNOR 연산을 수행할 수 있다.

표 1

[0052]

입력값(Input)	가중치 값(Weight)	출력 값(Output)
1	1	1
1	-1	-1
-1	1	-1
-1	-1	1
X	0	0
0	X	0

- [0053] 표1에서 입력값은 상부 입력신호와 하부 입력신호의 조합값을 의미하는 것으로, 입력값 '1'은 상부 입력 신호의 값이 '1'이고 하부 입력 신호의 값이 '-1'인 상태를 의미하고, 입력값 '-1'은 상부 입력 신호의 값이 '-1'이고 하부 입력 신호의 값이 '1'인 상태를 의미하며, 입력값 '0'은 상부 입력 신호와 하부 입력 신호의 차이 값이 '0'이거나, 기설정된 임계 크기 이하로 매우 작은 크기의 값인 상태를 의미할 수 있다.
- [0054] 또한, 표1에서 'X'는 'don't care' 상태, 즉 값이 '1', '0', '-1' 중 어느 하나인 상태를 의미할 수 있다.
- [0055] 즉, 논리 연산부(120)는 상부 입력 신호의 값이 '1'이고 하부 입력 신호의 값이 '-1'이며 가중치의 값이 '1'이

면 XNOR 연산의 수행 결과로 '1'을 출력하고, 상부 입력 신호의 값이 '1'이고 하부 입력 신호의 값이 '-1'이며 가중치의 값이 '-1'이면 XNOR 연산의 수행 결과로 '-1'을 출력할 수 있다.

- [0056] 또한, 논리 연산부(120)는 상부 입력 신호의 값이 '-1'이고 하부 입력 신호의 값이 '1'이며 가중치의 값이 '1'이면 XNOR 연산의 수행 결과로 '-1'을 출력하고, 상부 입력 신호의 값이 '-1'이고 하부 입력 신호의 값이 '1'이며 가중치의 값이 '-1'이면 XNOR 연산의 수행 결과로 '1'을 출력할 수 있다.
- [0057] 또한, 논리 연산부(120)는 상부 입력 신호와 하부 입력 신호의 차이 값이 '0'이거나 가중치 값이 '0'이면, XNOR 연산의 수행 결과로 '0'을 출력할 수 있다.
- [0058] 다시 말해, 반도체 소자(100)는 입력값이 '1'이고 가중치 값이 '1', 즉 상부 메모리스터(121)가 저저항 상태이고 하부 메모리스터(122)가 고저항 상태이면, 반도체 소자(100)의 기동 전극에 $I_{LRS}-I_{HRS}$ 만큼의 전류가 흐르고, 이는 XNOR 논리에 따라 '1'로 표현될 수 있다.
- [0059] 반면, 가중치 값이 '0', 즉 상부 메모리스터(121)와 하부 메모리스터(122)가 모두 고저항 상태로 프로그래밍되면 자연스러운 시냅스 소거 동작이 이루어지고, 이로 인해 기동 전극에 극히 소량의 전류가 흐르게 되며, 이는 XNOR 논리에 따라 '0'으로 표현될 수 있다.
- [0060] 한편, 반도체 소자(100)는 출력 트랜지스터를 포함하고, 출력 트랜지스터의 온/오프(On/Off) 동작에 기초하여 XNOR 연산의 수행 결과를 출력하는 출력부를 더 포함할 수 있으며, 일례로 출력부는 반도체 소자(100)의 다음단에 연결된 뉴런으로 XNOR 연산의 수행 결과를 출력할 수 있다.
- [0061] 즉, 본 발명은 저 비트 정밀도 심층 신경망을 효과적으로 활용할 수 있는 아키텍처로, 1 비트 삼항 정밀도를 위한 XNOR 계산을 통해, 벡터-행렬 곱 수행과 뇌 모사 컴퓨팅을 위한 VRRAM 어레이 구현이 가능하며, 이러한 구조에서는 특정 가중치 소거 동작과 비트-카운터 동작을 통해 연산 시의 에너지 소비를 효과적으로 줄일 수 있다.
- [0062]
- [0063] 도 2는 일실시예에 따른 뉴로모픽 컴퓨팅 시스템을 설명하는 도면이다.
- [0064] 도 2는 도 1을 통해 설명한 일실시예에 따른 반도체 소자의 적용 예시를 설명하는 도면으로, 이하에서 도 2를 통해 설명하는 내용 중 도 1을 통해 설명한 내용과 중복되는 설명은 생략하기로 한다.
- [0065] 도 2를 참조하면, 일실시예에 따른 뉴로모픽 컴퓨팅 시스템(200)은 입력부(210), 시냅스 어레이(220) 및 출력부(230)를 포함할 수 있다.
- [0066] 일실시예에 따른 입력부(210)는 적어도 하나의 상부 입력신호와 적어도 하나의 하부 입력신호를 수신할 수 있다.
- [0067] 일실시예에 따른 시냅스 어레이(220)는 적어도 하나의 상부 입력신호와 적어도 하나의 하부 입력신호에 기초하여 XNOR(Exclusive-NOR) 연산을 수행하는 복수의 시냅스를 구비할 수 있다.
- [0068] 여기서, 복수의 시냅스 각각은 하부 메모리스터(M_d) 및 하부 메모리스터(M_d)에 적층된 상부 메모리스터(M_u)를 포함할 수 있다.
- [0069] 또한, 복수의 시냅스 각각은 상부 메모리스터(M_u)를 통해 적어도 하나의 상부 입력신호 중 대응되는 어느 하나의 상부 입력신호(V_u)를 수신하며, 하부 메모리스터(M_d)를 통해 적어도 하나의 하부 입력신호 중 대응되는 어느 하나의 하부 입력 신호(V_d)를 수신할 수 있다.
- [0070] 예를 들면, 시냅스 어레이(220)는 동일한 로우(row) 라인(입력신호가 인가되는 방향)에 연결된 시냅스들에게 동일한 상부 입력신호와 하부 입력신호가 인가될 수 있다.
- [0071] 또한, 시냅스는 도 1의 반도체 소자일 수 있다. 즉, 시냅스 어레이(220)는 3D RRAM의 메모리 어레이일 수 있다.
- [0072] 일측에 따르면, 복수의 시냅스 각각은 상부 메모리스터(M_u)와 하부 메모리스터(M_d) 각각의 저항 상태에 따라 가중치 값이 결정되며, 어느 하나의 상부 입력신호(V_u), 어느 하나의 하부 입력신호(V_d) 및 가중치 값에 기초하여 XNOR 연산을 수행할 수 있다.
- [0073] 일측에 따르면, 복수의 시냅스 각각은 상부 메모리스터(M_u)가 저저항 상태이고 하부 메모리스터(M_d)가 고저항 상태이

면 가중치 값이 '1'로 결정될 수 있다.

- [0074] 또한, 복수의 시냅스 각각은 상부 맴리스터(M_u)가 고저항 상태이고 하부 맴리스터(M_d)가 저저항 상태이면 가중치 값이 '-1'로 결정될 수 있다.
- [0075] 또한, 복수의 시냅스 각각은 상부 맴리스터(M_u)가 고저항 상태이고 하부 맴리스터(M_d)가 고저항 상태이면 가중치 값이 '0'으로 결정될 수 있다.
- [0076] 일측에 따르면, 복수의 시냅스 각각은 상부 입력신호(V_u)의 값이 '1'이고 하부 입력신호(V_d)의 값이 '-1'이며 가중치의 값이 '1'이면 XNOR 연산의 수행 결과로 '1'을 출력하고, 상부 입력신호(V_u)의 값이 '1'이고 하부 입력신호(V_d)의 값이 '-1'이며 가중치의 값이 '-1'이면 XNOR 연산의 수행 결과로 '-1'을 출력할 수 있다.
- [0077] 또한, 복수의 시냅스 각각은 상부 입력신호(V_u)의 값이 '-1'이고 하부 입력신호(V_d)의 값이 '1'이며 가중치의 값이 '1'이면 XNOR 연산의 수행 결과로 '-1'을 출력하고, 상부 입력신호(V_u)의 값이 '-1'이고 하부 입력신호(V_d)의 값이 '1'이며 가중치의 값이 '-1'이면 XNOR 연산의 수행 결과로 '1'을 출력할 수 있다.
- [0078] 또한, 복수의 시냅스 각각은 상부 입력신호(V_u)와 하부 입력신호(V_d)의 차이 값이 '0'이거나 가중치 값이 '0'이면, XNOR 연산의 수행 결과로 '0'을 출력할 수 있다.
- [0079] 한편, 출력부(230)는 출력 트랜지스터(231)를 포함하고, 출력 트랜지스터(231)의 온/오프(On/Off) 동작에 기초하여 XNOR 연산의 수행 결과를 출력할 수 있다. 일례로, 출력부(230)는 시냅스 어레이(220)의 다음 단계에 연결된 뉴런으로 XNOR 연산의 수행 결과를 출력할 수 있으며, 출력 트랜지스터(231)의 온/오프 동작은 메모리 컨트롤러에 의해 제어될 수 있다.
- [0081] 도 3은 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 구현예를 설명하는 도면이다.
- [0082] 다시 말해, 도 3는 도 2를 통해 설명한 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 구현 예시를 설명하는 도면으로, 이하에서 도 3을 통해 설명하는 내용 중 도 2를 통해 설명한 내용과 중복되는 설명은 생략하기로 한다.
- [0083] 도 3을 참조하면, 도면부호 300은 모놀리식 워드 평면(단일 그래핀층)을 활용한 시냅스 어레이(3D VRRAM 어레이)를 도시하며, 여기서 비트 라인은 출력 뉴런으로 활용될 수 있다.
- [0084] 구체적으로, 일실시예에 따른 뉴로모픽 컴퓨팅 시스템은 상부 전극(Pillar electrode)이 TiN 전극으로 구성되고, 워드 라인이 그래핀 워드 평면(Graphene Word Plane)으로 구현될 수 있다.
- [0085] 그래핀 워드 평면은 상부 전극의 하부에 배치되도록 구성되고, 수직 구조 메모리 셀(상부 맴리스터 및 하부 맴리스터)은 상부 전극과 그래핀 워드 평면 사이에 메모리 물질로 형성될 수 있다. 일례로, 메모리 물질은 HfO_2 , TiO_2 , Al_2O_3 및 Ta_2O_5 중 적어도 하나를 포함할 수 있다.
- [0086] 그래핀 워드 평면은 적어도 적어도 둘 이상의 그래핀 층을 포함하고, 그래핀 층 각각은 금속 콘택(metal contact)과 연결될 수 있다. 또한, 그래핀 층은 소정의 두께(t_{di})로 이루어진 SiO_2 보호층(passivation layer)를 더 포함할 수 있다.
- [0087] 한편, 수직 구조 메모리 셀은 셀렉터 수직 트랜지스터(Selector vertical transistor)와 각각 연결되고, 셀렉터 수직 트랜지스터는 온/오프 동작의 제어를 위해 대응되는 셀렉터 라인(Selector lines)과 연결될 수 있다.
- [0089] 도 4는 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 구동 과정을 설명하는 도면이다.
- [0090] 도 4를 참조하면, 도 4의 (a)는 뉴로모픽 컴퓨팅 시스템(3D VRRAM)의 라이트 동작을 설명하는 도면이고, 도 4의 (b)는 뉴로모픽 컴퓨팅 시스템(3D VRRAM)의 리드 동작을 설명하는 도면이며, 도 4의 (c)는 비트 라인을 출력 뉴런으로 활용하는 평면 어레이 기반 가중 합 연산 동작을 설명하는 도면이다.
- [0091] 구체적으로, 도 4의 (a) 및 (b)에 따르면, 1/2 바이어스의 라이트 동작의 경우 선택된 WP 및 BL 각각은 완전한 라이트 전압(V_w)이 인가되고 접지가 수행되는 반면, 선택되지 않은 다른 모든 셀들은 절반의 라이트 전압($V_w/2$)

이 인가되거나 수직 트랜지스터에 의해 선택되지 않으며, 이러한 라이트 동작을 통해 상부 맴리스터 및 하부 맴리스터 각각의 저항 상태가 결정될 수 있으며, 리드 동작의 경우, 읽기 전압(V_R)이 인가되는 선택된 WP를 제외한 나머지 WP와 BL이 접지될 수 있다.

[0092] 도 4의 (c)에 따르면, 일실시예에 따른 뉴로모픽 컴퓨팅 시스템은 출력 뉴런의 수가 적층된 WP의 수에 크게 영향을 받으며, 이는 에칭 중횡비에 따라 달라질 수 있다.

[0093] 또한, 뉴로모픽 컴퓨팅 시스템은 WP와 SL(selector line)을 결합하여 벡터-행렬 곱 연산(VMM)을 수행하며, 여기서 선택된 WP와 선택되지 않은 WP는 각각 리드 전압(V_R)과 0V의 전압이 인가되고, 선택된 BL은 접지될 수 있다. 따라서 각 BL에서 판독된 출력 전류는 평면 어레이의 y축(즉, 메모리 셀의 적층 방향)을 따라 위치한 메모리 셀의 입력 전압 및 컨덕턴스의 WS(weighted-sum)을 의미할 수 있다.

[0095] 도 5a 내지 도 5c는 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 동작 특성을 설명하는 도면이다.

[0096] 도 5a 내지 도 5c를 참조하면, 도면부호 510은 이진(흑/백) 입력 벡터를 통한 이중층 MLP(2-layer perceptron) 신경망 네트워크를 도시하고, 도면부호 520은 XNOR 연산을 수행하는 1비트 삼항(-1, 0, 1) 시냅스의 개략도를 도시한다.

[0097] 또한, 도면부호 530은 1비트 심진 정밀도 체계로 가중치를 양자화하는 과정을 설명하는 과정으로 여기서 Sign bit과 MSB(즉, condition bits)는 삼항 수준을 결정할 수 있다.

[0098] 구체적으로, 일실시예에 따른 뉴로모픽 컴퓨팅 시스템의 시냅스 어레이, 즉 VRRAM 어레이의 인식 성능은 MNIST(Modified National Institute of Standards and Technology) 데이터베이스의 필기 숫자를 이용하여 평가할 수 있다.

[0099] 일실시예에 따른 뉴로모픽 컴퓨팅 시스템은 그래핀 기반 3D VRRAM로 구현될 수 있으며, 이는 도면부호 510과 같이 400 개의 입력 레이어, 200개의 은닉 레이어 및 10개의 출력 레이어로 구성된 이중층 MLP 신경망 네트워크를 통해 생산성(productivity)이 추정될 수 있다.

[0100] 이 과정에서 도면부호 520과 같이, WP 평면 크기를 줄이고 삼항 XNOR 연산을 구현하기 위해 원본 이미지를 자르고 이진화시킬 수 있다.

[0101] 트레이닝 과정을 위해 필요한 가중치 업데이트 동작은 하기 수식1에 기초하여 수행될 수 있다.

[0102] [수식1]

$$\Delta W_{ij} = \alpha \cdot \delta_j^b(l) V_i^b(l) \text{ for } b = 1, \dots, B$$

[0103] 여기서 α 는 학습률(learning rate), B는 순차적 가중치 업데이트를 위해 샘플이 계산되는 배치 크기이며, V_i , δ_j 는 각각 $m \times n$ 시냅스 크기의 l번째 레이어에 대한 특정 뉴런 입력 값 및 출력 오차를 의미한다.

[0105] 일실시예에 따른 뉴로모픽 컴퓨팅 시스템은 적층 특성으로 인해 트레이닝을 6비트 이상의 가중치 정밀도로 수행할 수 있으며 더 나아가 이를 최적화할 수 있다.

[0106] 또한, 삼항 가중치에 대한 XNOR 연산을 이용하면(도면부호 520), 비트 카운트 연산과 자연적 가중치 프루닝에 의해 연산 및 에너지 사용에 따른 자원을 줄일 수 있다.

[0108] 결국, 본 발명을 이용하면, 적층된 복수의 맴리스터를 이용하여 XNOR 연산을 수행함으로써, 높은 메모리 집적도, 낮은 소비 전력 및 저 정밀도의 심층 신경망을 구현할 수 있다.

[0109] 또한, 본 발명을 이용하면, 적층된 복수의 맴리스터를 이용하는 XNOR 아키텍처를 VRRAM에 적용하여 개별 메모리 셀과 전체적인 어레이의 성능을 향상시킬 수 있다.

[0111] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면

상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들면, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 장치, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

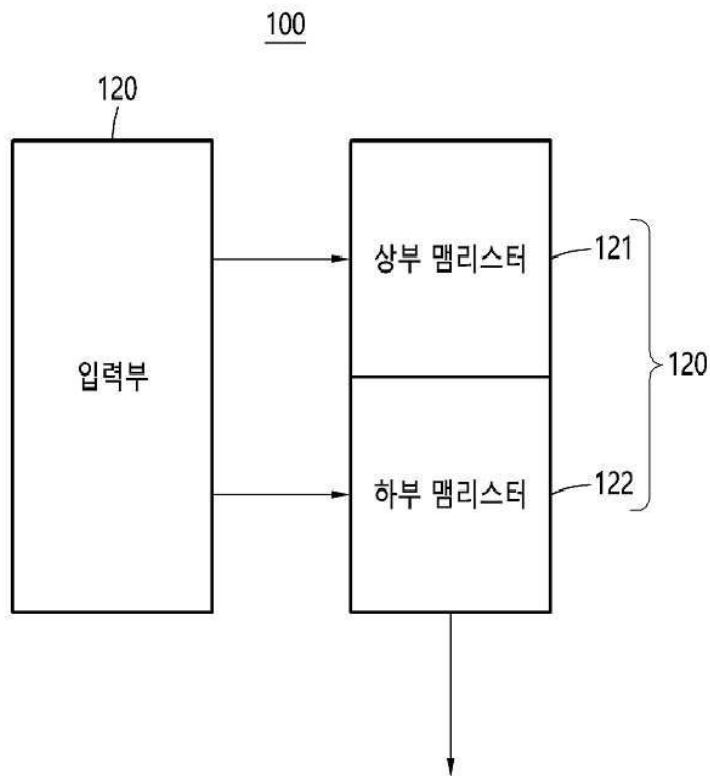
[0112] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

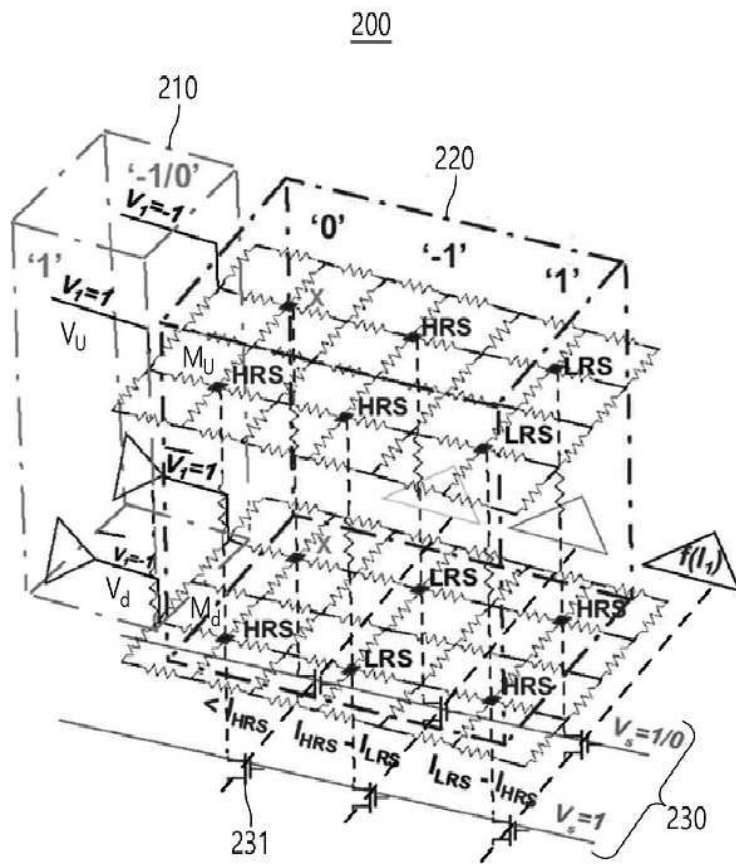
[0114] 100: 반도체 소자 110: 입력부
120: 논리 연산부 121: 상부 매퍼스터
122: 하부 매퍼스터

도면

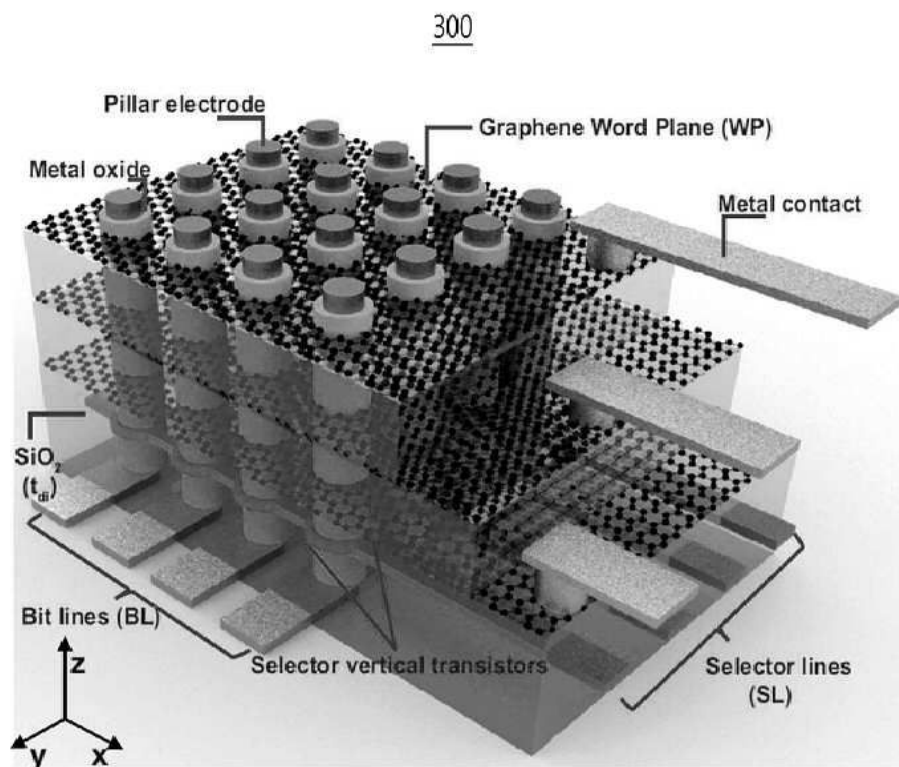
도면1



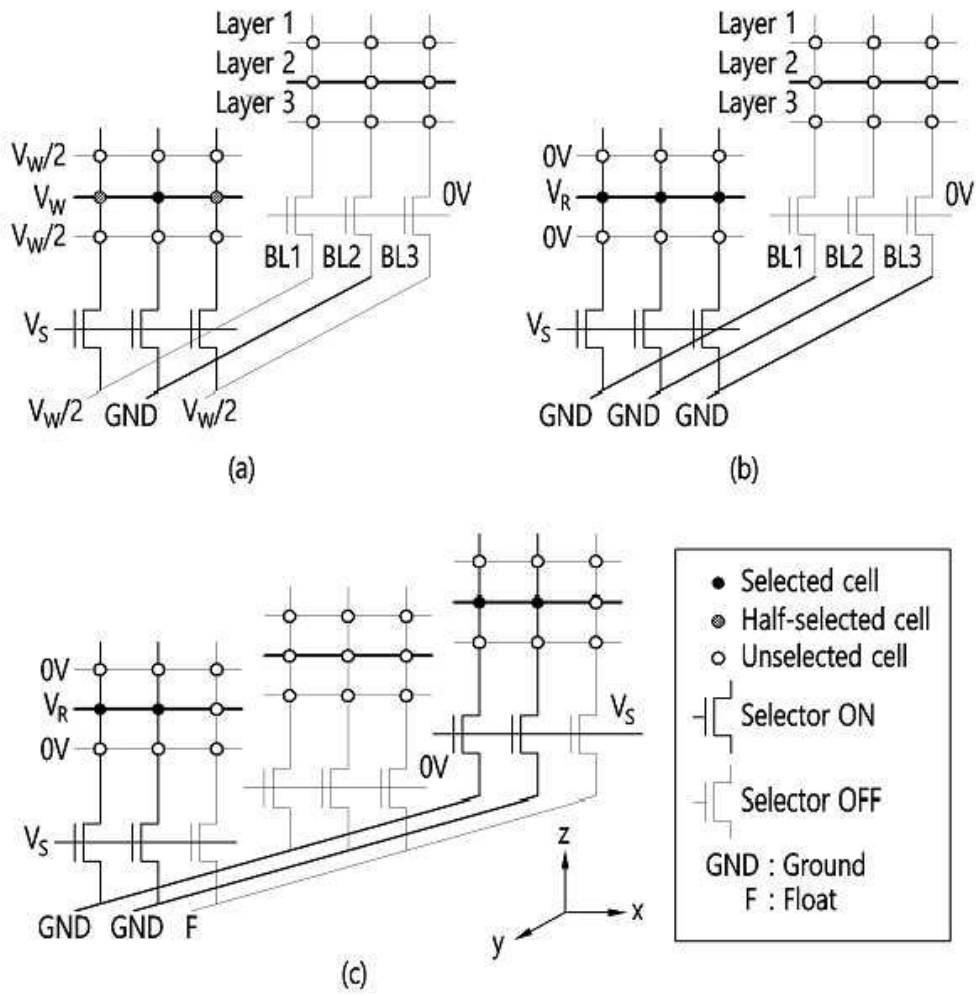
도면2



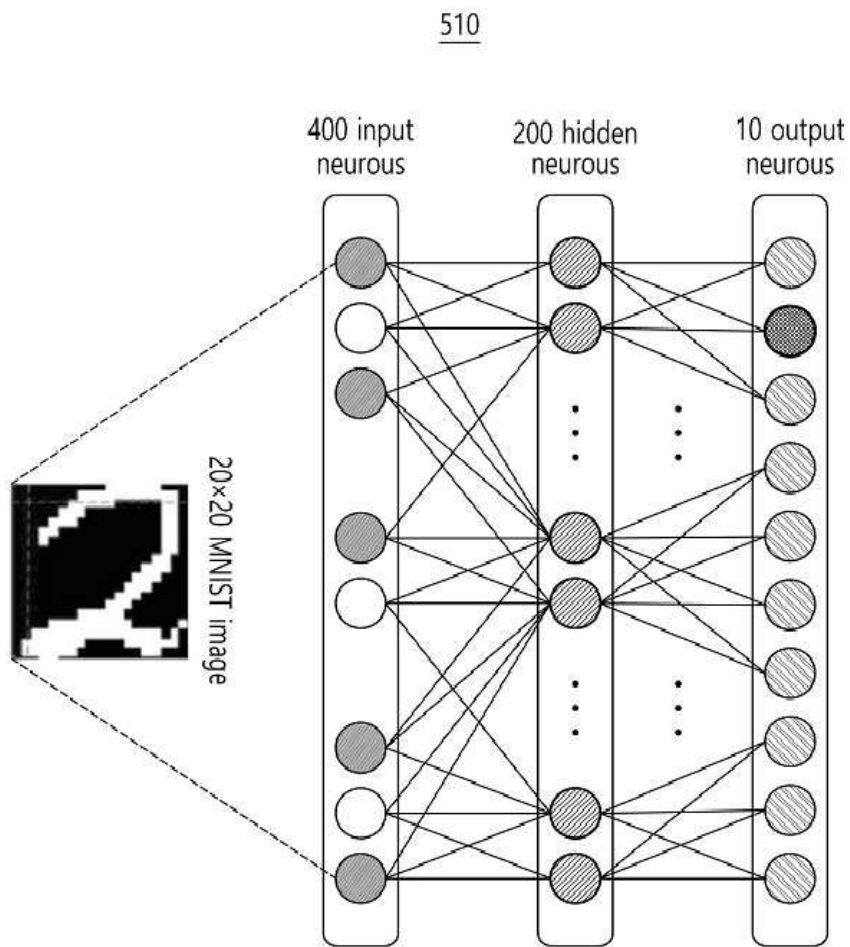
도면3



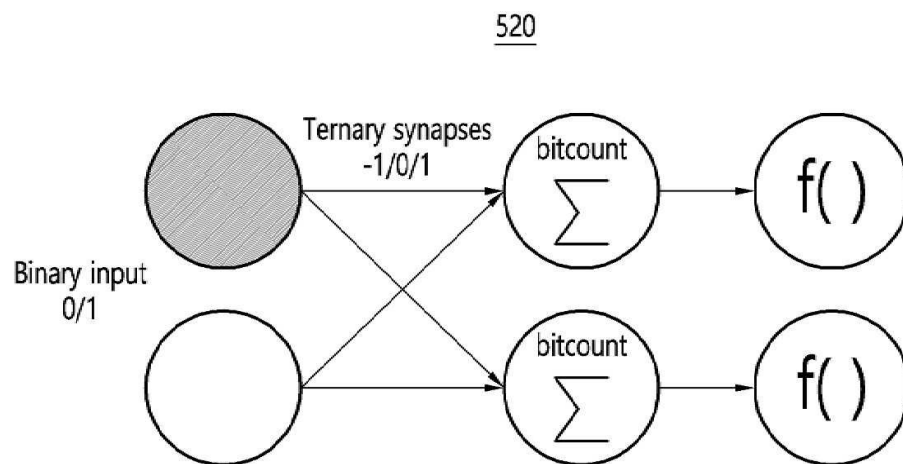
도면4



도면5a



도면5b



도면5c

