



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0124803
(43) 공개일자 2020년11월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) H01L 43/02 (2006.01)
H01L 43/08 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
H01L 43/02 (2013.01)
(21) 출원번호 10-2019-0047992
(22) 출원일자 2019년04월24일
심사청구일자 2019년04월24일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
성균관대학교산학협력단
경기도 수원시 장안구 서부로 2066 (천천동, 성균관대학교내)
(72) 발명자
홍종일
서울특별시 서대문구 연세로 50 신소재공학과 (신촌동, 연세대학교)
권기원
경기도 성남시 분당구 내정로166번길 42, 119동 3001호(수내동, 파크타운삼익아파트)
(74) 대리인
민영준

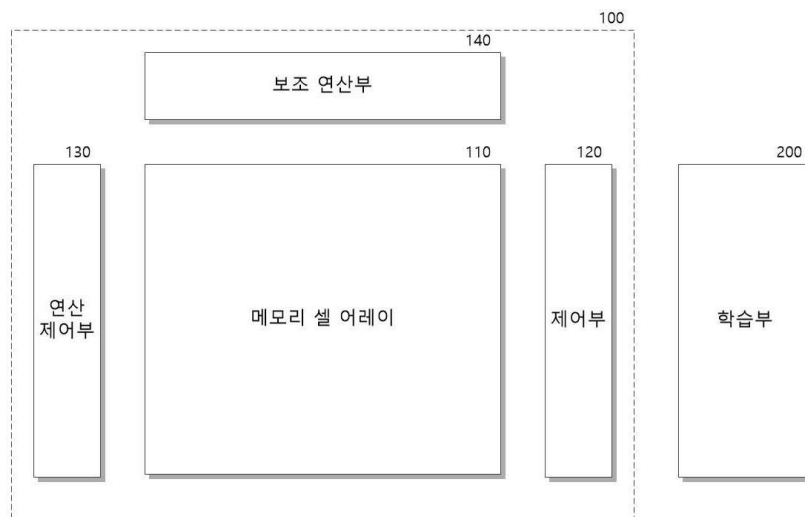
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 자성/비자성 다층 박막 메모리 기반 고속 인공 신경망 가속기 및 이의 운용 방법

(57) 요약

본 발명은 자성/비자성 다층 박막 메모리 소자로 구현되어 데이터를 아날로그적 저항값의 형식으로 저장할 수 있는 다수의 메모리 셀로 구성된 메모리 셀 어레이, 인공 신경망의 다수의 레이어를 구성하는 다수의 가중치를 다수의 메모리 셀에 저장하고, 가중치가 저장된 다수의 메모리 셀에 연산 데이터에 대응하는 연산 전류를 공급하여 가중치와 연산 데이터 사이의 기지정된 연산 결과를 획득하는 연산 제어부 및 가중치와 연산 데이터 사이의 연산 결과에 대해 기지정된 보조 연산을 수행하여 인공 신경망의 패턴 인식 결과를 출력하는 보조 연산부를 포함하는 고속 인공 신경망 및 이의 운용 방법을 제공할 수 있다.

대표도



(52) CPC특허분류

H01L 43/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	2015M3D1A1070465
부처명	미래창조과학부
과제관리(전문)기관명	한국연구재단
연구사업명	미래소재디스커버리사업
연구과제명	설계기반 Spin-Orbitronics 소재 개발
기 여 율	1/1
과제수행기관명	고대학교 산학협력단
연구기간	2015.12.04 ~ 2021.12.03

명세서

청구범위

청구항 1

자성/비자성 다층 박막 메모리 소자로 구현되어 데이터를 아날로그적 저항값의 형식으로 저장할 수 있는 다수의 메모리 셀로 구성된 메모리 셀 어레이;

인공 신경망의 다수의 레이어를 구성하는 다수의 가중치를 다수의 메모리 셀에 저장하고, 가중치가 저장된 다수의 메모리 셀에 연산 데이터에 대응하는 연산 전류를 공급하여 상기 가중치와 상기 연산 데이터 사이의 기지정된 연산 결과를 획득하는 연산 제어부; 및

상기 가중치와 상기 연산 데이터 사이의 연산 결과에 대해 기지정된 보조 연산을 수행하여 상기 인공 신경망의 패턴 인식 결과를 출력하는 보조 연산부;를 포함하는 인공 신경망 가속기.

청구항 2

제1 항에 있어서, 상기 연산 제어부는

상기 다수의 가중치를 대응하는 크기의 쓰기 전류로 변환하여, 변환된 쓰기 전류를 메모리 셀로 공급하여 상기 메모리 셀의 저항값을 가변하여 저장하고,

상기 연산 데이터를 대응하는 크기의 연산 전류로 변환하여 상기 가중치가 저항값의 형식으로 저장된 메모리 셀로 공급하고, 상기 연산 전류가 공급된 메모리 셀에서 출력되는 전압(또는 전류)을 감지하여 상기 가중치와 상기 연산 데이터 사이의 곱셈 연산 결과를 판별하는 인공 신경망 가속기.

청구항 3

제2 항에 있어서, 상기 다수의 메모리 셀 각각은

공급되는 쓰기 전류에 대응하는 스핀 전류를 배치된 평면에 수직 방향으로 생성하는 스핀 전류 생성층; 및

상기 스핀 전류 생성층 상에 배치되어 스핀 전류에 대응하는 터널 자기저항(TMR)을 저항값으로 갖는 자기터널 접합층;을 포함하는 인공 신경망 가속기.

청구항 4

제3 항에 있어서, 상기 스핀 전류 생성층은

기관 상에 배치된 제1 비자성 도전층;

상기 제1 비자성 도전층 상에 정렬되어 배치된 제2 비자성 도전층; 및

상기 제1 비자성 도전층과 상기 제2 비자성 도전층 사이에 배치되고 수직자기 이방성을 갖는 자성층;을 포함하고,

상기 자기터널 접합층은

스핀 전류에 의해 자기 모멘트의 방향이 가변되는 자유 자성층;

상기 자유 자성층 상에 배치되는 터널 절연층; 및

상기 터널 절연층 상에 배치되는 고정 자성층;을 포함하는 인공 신경망 가속기.

청구항 5

제4 항에 있어서, 상기 자유 자성층은

상기 스핀 전류 생성층에서 생성되어 인가되는 상기 스핀 전류에 따라 홀 저항이 선형적으로 가변되는 구간이 포함되고,

상기 쓰기 전류는

상기 자유 자성층의 홀 저항이 선형적으로 가변되는 구간의 스핀 전류가 생성되는 구간에서 상기 다수의 가중치를 대응하는 크기를 갖는 인공 신경망 가속기.

청구항 6

제4 항에 있어서, 상기 메모리 셀 어레이는

다수의 메모리가 3차원 적층 구조로 적층되어 구성되는 인공 신경망 가속기.

청구항 7

제4 항에 있어서, 상기 메모리 셀 어레이는

다수의 메모리 셀 각각이 하단 또는 상단에 적층된 메모리 셀로 연산 결과를 전달하는 인공 신경망 가속기.

청구항 8

자성/비자성 다층 박막 메모리 소자로 구현되어 아날로그적인 저항값의 형식으로 데이터를 저장하는 다수의 메모리 셀로 구성된 메모리 셀 어레이를 포함하는 인공 신경망 가속기의 운용 방법에 있어서,

인공 신경망의 다수의 레이어를 구성하는 다수의 가중치를 획득하는 단계;

상기 가중치에 대응하는 쓰기 전류를 메모리 셀에 공급하여, 상기 가중치를 저항값의 형식으로 메모리 셀에 저장하는 단계;

상기 가중치가 저장된 다수의 메모리 셀에 연산 데이터에 대응하는 전류를 공급하여, 상기 가중치와 상기 연산 데이터 사이의 기지정된 연산 결과를 획득하는 단계; 및

상기 가중치와 상기 연산 데이터 사이의 연산 결과에 대해 기지정된 보조 연산을 수행하여 상기 인공 신경망의 패턴 인식 결과를 출력하는 단계; 를 포함하는 인공 신경망 가속기의 운용 방법.

청구항 9

제8 항에 있어서, 상기 메모리 셀에 저장하는 단계는

상기 메모리 셀 어레이의 다수의 메모리 셀 중 상기 다수의 가중치를 저장할 메모리 셀을 선택하는 단계; 및

상기 다수의 가중치를 대응하는 크기의 쓰기 전류로 변환하는 단계; 및

상기 쓰기 전류를 선택된 메모리 셀로 공급하여 상기 메모리 셀의 저항값을 가변하는 단계; 를 포함하는 인공 신경망 가속기의 운용 방법.

청구항 10

제8 항에 있어서, 상기 저항값을 가변하는 단계는

선택된 상기 메모리 셀의 스핀 전류 생성층에 상기 쓰기 전류가 공급되는 단계;

상기 스핀 전류 생성층이 상기 쓰기 전류에 대응하여 배치 평면에 수직 방향의 스핀 전류를 생성하는 단계; 및

상기 스핀 전류 생성층 상에 배치된 자기터널 접합층이 상기 스핀 전류에 대응하는 터널 자기저항(TMR)을 저항값으로 갖는 단계; 를 포함하는 인공 신경망 가속기의 운용 방법.

청구항 11

제10 항에 있어서, 상기 스핀 전류를 생성하는 단계는

상기 스핀 전류 생성층이 기판 상에 배치된 다수의 비자성 도전층 및 상기 다수의 비자성 도전층 사이에 배치되고 수직자기 이방성을 갖는 적어도 하나의 자성층을 포함하여 상기 쓰기 전류의 크기에 대응하는 상기 스핀 전류를 생성하는 인공 신경망 가속기의 운용 방법.

청구항 12

제11 항에 있어서, 상기 터널 자기저항(TMR)을 저항값으로 갖는 단계는

상기 자기터널 접합층의 자유 자성층이 상기 쓰기 전류에 대응하여 생성된 상기 스핀 전류의 크기에 따라 자기 모멘트의 방향 변경되는 단계; 및

상기 자기터널 접합층에서 상기 자기터널 접합층의 자기 모멘트 방향과 상기 자유 자성층 상에 배치된 터널 절연층 상에 배치되는 고정 자성층의 자기 모멘트의 방향에 의해 터널 자기저항(TMR)이 조절되는 단계; 를 포함하는 인공 신경망 가속기의 운용 방법.

청구항 13

제12 항에 있어서, 상기 쓰기 전류가 공급되는 단계는

상기 자유 자성층의 홀 저항이 선형적으로 가변되는 구간의 스핀 전류가 생성되는 구간에서 상기 다수의 가중치를 대응하는 크기를 갖도록 상기 쓰기 전류를 공급하는 인공 신경망 가속기의 운용 방법.

청구항 14

제8 항에 있어서, 상기 연산 결과를 획득하는 단계는

상기 연산 데이터를 대응하는 크기의 연산 전류로 변환하는 단계;

상기 가중치가 저항값의 형식으로 저장된 메모리 셀로 상기 연산 전류를 공급하는 단계; 및

상기 연산 전류가 공급된 메모리 셀에서 출력되는 전압(또는 전류)을 감지하여 상기 가중치와 상기 연산 데이터 사이의 곱셈 연산 결과를 판별하는 단계; 를 포함하는 인공 신경망 가속기의 운용 방법.

발명의 설명

기술 분야

[0001] 본 발명은 고속 인공 신경망 가속기 및 이의 운용 방법에 관한 것으로, 자성/비자성 다층 박막 메모리 기반 고속 인공 신경망 가속기 및 이의 운용 방법에 관한 것이다.

배경 기술

[0002] 현재 인간의 두뇌가 패턴을 인식하는 방법을 모사하여 두뇌와 비슷한 방식으로 여러 정보를 처리하도록 구성된 인공 신경망(artificial neural network)을 이용한 딥 러닝에 대한 연구가 활발하게 진행되고 있다. 딥 러닝은 일예로 객체 분류, 객체 검출, 음성 인식, 자연어 처리, 자율 주행 등의 다양한 분야에 적용되고 있으며, 적용 분야가 계속 확장되고 있다.

[0003] 이러한 인공 신경망이 요구되는 패턴 인식 성능을 나타내기 위해서는 방대한 학습 데이터를 기반으로 학습이 수행되어야 하며, 이 과정에서 대량의 연산, 특히 곱셈 및 덧셈 연산을 요구한다.

[0004] 기존에 인공 신경망은 대부분 소프트웨어로 구현되며, CPU(Central Processing Unit)나 GPU(Graphics Processing Unit)와 같은 프로세서와 메모리 등의 하드웨어를 이용하여 요구되는 연산을 수행한다. 이때 프로세서와 메모리가 별도로 구비됨에 따라 프로세서는 연산되어야 하는 입력 데이터를 메모리로부터 전달받아야 하며, 연산 결과인 출력값을 다시 메모리로 전달하여 저장해야 한다. 인공 신경망이 학습을 수행하거나, 빅데이터를 처리하기 위해 이용되는 경우와 같이 대량의 연산을 수행하는 경우, 프로세서와 메모리 사이에는 대규모의 데이터 전송이 필요하여 연산 속도를 크게 저하시킬 뿐만 아니라, 대량의 전력 소모를 유발한다. 현재 아키텍처에서 연산 프로세서와 메모리 사이의 데이터 전송은 연산 프로세서의 부동 소수점 연산 대비 100배 이상의 전력 소비가 요구되는 경우도 있다.

[0005] 이러한 프로세서와 메모리가 별도로 구비되어 발생하는 비효율성을 극복하기 위해, 최근에는 메모리 칩 내부에 연산 로직을 이식한 PIM(processing-in-memory)을 이용한 인공 신경망 가속기(artificial neural network accelerator)에 대한 연구가 활발히 진행되고 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 공개 특허 제10-2018-0028966호 (2018.03.19 공개)

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은 자성/비자성 다층 박막 메모리를 이용하여 메모리 칩 내부에 연산 로직이 이식된 고효율 PIM 구조의 인공 신경망 가속기 및 이의 운용 방법을 제공하는데 있다.

[0008] 본 발명의 다른 목적은 자성/비자성 다층 박막 메모리를 메모리 셀로 이용하여 인공 신경망의 가중치를 아날로그 값으로 저장할 수 있는 고속 인공 신경망 가속기 및 이의 운용 방법을 제공하는데 있다.

[0009] 본 발명의 또 다른 목적은 가중치가 저항값의 형태로 저장된 메모리 셀에 연산 데이터를 인가하여 가중치가 저장된 메모리 셀에서 연산 결과를 출력할 수 있는 고속 인공 신경망 가속기 및 이의 운용 방법을 제공하는데 있다.

과제의 해결 수단

[0010] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 고속 인공 신경망 가속기는 자성/비자성 다층 박막 메모리 소자로 구현되어 데이터를 아날로그적 저항값의 형식으로 저장할 수 있는 다수의 메모리 셀로 구성된 메모리 셀 어레이; 인공 신경망의 다수의 레이어를 구성하는 다수의 가중치를 다수의 메모리 셀에 저장하고, 가중치가 저장된 다수의 메모리 셀에 연산 데이터에 대응하는 연산 전류를 공급하여 상기 가중치와 상기 연산 데이터 사이의 기지정된 연산 결과를 획득하는 연산 제어부; 및 상기 가중치와 상기 연산 데이터 사이의 연산 결과에 대해 기지정된 보조 연산을 수행하여 상기 인공 신경망의 패턴 인식 결과를 출력하는 보조 연산부; 를 포함한다.

[0011] 상기 연산 제어부는 상기 다수의 가중치를 대응하는 크기의 쓰기 전류로 변환하여, 변환된 쓰기 전류를 메모리 셀로 공급하여 상기 메모리 셀의 저항값을 가변하여 저장하고, 상기 연산 데이터를 대응하는 크기의 연산 전류로 변환하여 상기 가중치가 저항값의 형식으로 저장된 메모리 셀로 공급하고, 상기 연산 전류가 공급된 메모리 셀에서 출력되는 전압(또는 전류)을 감지하여 상기 가중치와 상기 연산 데이터 사이의 곱셈 연산 결과를 판별할 수 있다.

[0012] 다수의 메모리 셀 각각은 공급되는 쓰기 전류에 대응하는 스핀 전류를 배치된 평면에 수직 방향으로 생성하는 스핀 전류 생성층; 및 상기 스핀 전류 생성층 상에 배치되어 스핀 전류에 대응하는 터널 자기저항(TMR)을 저항값으로 갖는 자기터널 접합층; 을 포함할 수 있다.

[0013] 상기 스핀 전류 생성층은 기판 상에 배치된 제1 비자성 도전층; 상기 제1 비자성 도전층 상에 정렬되어 배치된 제2 비자성 도전층; 및 상기 제1 비자성 도전층과 상기 제2 비자성 도전층 사이에 배치되고 수직자기 이방성을 갖는 자성층; 을 포함하고, 상기 자기터널 접합층은 스핀 전류에 의해 자기 모멘트의 방향이 가변되는 자유 자성층; 상기 자유 자성층 상에 배치되는 터널 절연층; 및 상기 터널 절연층 상에 배치되는 고정 자성층; 을 포함할 수 있다.

[0014] 상기 자유 자성층은 상기 스핀 전류 생성층에서 생성되어 인가되는 상기 스핀 전류에 따라 홀 저항이 선형적으로 가변되는 구간이 포함되고, 상기 쓰기 전류는 상기 자유 자성층의 홀 저항이 선형적으로 가변되는 구간의 스핀 전류가 생성되는 구간에서 상기 다수의 가중치를 대응하는 크기를 가질 수 있다.

[0015] 상기 메모리 셀 어레이는 다수의 메모리가 3차원 적층 구조로 적층되어 구성될 수 있으며, 다수의 메모리 셀 각각은 하단 또는 상단에 적층된 메모리 셀로 연산 결과를 전달할 수 있다.

[0016] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 고속 인공 신경망 가속기의 운용 방법은 자성/비자성 다층 박막 메모리 소자로 구현되어 아날로그적인 저항값의 형식으로 데이터를 저장하는 다수의 메모리 셀로 구성된 메모리 셀 어레이를 포함하는 인공 신경망 가속기의 운용 방법에 있어서, 인공 신경망의 다수의 레이어를 구성하는 다수의 가중치를 획득하는 단계; 상기 가중치에 대응하는 쓰기 전류를 메모리 셀에 공급하여, 상기 가중치를 저항값의 형식으로 메모리 셀에 저장하는 단계; 상기 가중치가 저장된 다수의 메모리 셀에 상기 연산 데이터에 대응하는 전류를 공급하여, 상기 가중치와 상기 연산 데이터 사이의 기지정된 연산 결과를 획득하는 단계; 및 상기 가중치와 상기 연산 데이터 사이의 연산 결과에 대해 기지정된 보조 연산을 수행하여 상기 인공 신

경망의 패턴 인식 결과를 출력하는 단계; 를 포함한다.

발명의 효과

[0017] 따라서, 본 발명의 실시예에 따른 고속 인공 신경망 가속기 및 이의 운용 방법은 인공 신경망의 가중치 및 연산 데이터 중 적어도 하나를 아날로그 값의 형태로 저장할 수 있는 자성/비자성 다층 박막 메모리를 메모리 셀로 이용하여 메모리 칩 내부에 연산 로직이 이식된 고효율 PIM 구조의 인공 신경망 가속기를 구현함으로써, 가중치 또는 연산 데이터를 저장하기 위한 메모리 용량을 크게 줄일 수 있다. 그리고 가중치가 저장된 메모리 셀에 연산 데이터를 전달하여, 메모리 셀 각각에서 가중치와 연산 데이터 사이의 연산 결과가 출력되도록 함으로써, 연산을 위한 데이터 전송을 최소화할 수 있을 뿐만 아니라 대규모 병렬 연산을 수행할 수 있어, 고속 고효율의 연산이 가능하며, 이에 따라 전력 소모를 크게 줄일 수 있다.

[0018] 또한 자성/비자성 다층 박막 메모리의 자기터널 접합층과 스핀 전류 생성층 중 스핀 전류 생성층이 다수의 비자성 도전층과 적어도 하나의 자성층을 교대로 적층하여 구성됨으로써, 자성/비자성 다층 박막 메모리로 구현되는 메모리 셀에 저장되는 터널 자기저항(TMR)을 적은 전력으로 용이하게 변경할 수 있을 뿐만 아니라, 다수의 자성/비자성 다층 박막 메모리를 용이하게 3D 적층할 수 있어 초소형 고용량의 인공 신경망 가속기를 용이하게 구현할 수 있다.

도면의 간단한 설명

[0019] 도1 은 인공 신경망의 개념적 구조를 나타낸다.
 도2 는 본 발명의 일 실시예에 따른 고속 인공 신경망 가속기의 개략적 구조를 나타낸다.
 도3 은 도2 의 메모리 어레이의 메모리 셀인 자성/비자성 다층 박막 메모리 구조의 일예를 나타낸다.
 도4 는 자성/비자성 다층 박막 메모리의 전류에 따른 홀 저항 변화의 아날로그적 특성을 나타낸다.
 도5 는 도2 의 메모리 어레이에서 메모리 셀의 구조의 다른 예를 나타낸다.
 도6 은 자성/비자성 다층 박막 메모리 소자로 구현된 메모리 셀의 3차원 적층 구조의 일예를 나타낸다.
 도7 은 본 발명의 일 실시예에 따른 자성/비자성 다층 박막 메모리를 이용한 고속 인공 신경망 가속기의 운용 방법을 나타낸다.
 도8 은 도7 의 학습 가중치 저장 단계를 상세하게 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0020] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시예에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0021] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0022] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0023] 도1 은 인공 신경망의 개념적 구조를 나타낸다.

[0024] 도1 에 도시된 바와 같이, 일반적으로 인공 신경망(artificial neural network)은 패턴이 인식되어야 할 입력 데이터가 인가되는 입력 레이어(input layer)와 인식된 패턴에 대응하는 출력 데이터를 출력하는 출력 레이어(output layer) 및 입력 레이어(input layer)와 출력 레이어(output layer) 사이에 배치되고 인간의 신경망을 모사한 디지털 뉴런(digital neuron)으로 구성되어 입력 데이터의 패턴을 인식하는 다수의 은닉 레이어(hidden layer)를 포함한다.

[0025] 여기서 입력 데이터는 음성 데이터, 이미지 데이터, 다수 프레임을 갖는 동영상 데이터 등의 다양한 종류의 데

이터일 수 있으며, 행렬(또는 벡터) 형태로 입력될 수 있으나 이에 한정되지 않는다.

- [0026] 다수의 은닉 레이어 각각에는 학습에 의해 획득되는 가중치(가중치 벡터라고도 함)가 할당되며, 이전 레이어로부터 대응하는 연산 데이터를 인가받아 할당된 가중치와 기지정된 연산을 수행한다. 그리고 연산의 수행 결과로서 인가된 데이터의 특징 패턴을 추출한다. 즉 인공 신경망에 인가되는 입력 데이터의 특징 패턴을 추출한다.
- [0027] 여기서 은닉 레이어의 개수와 이전 레이어로부터 인가받는 대응하는 연산 데이터는 인공 신경망에 따라 다양하게 설정될 수 있으며, 각각의 은닉 레이어 또한 행렬(또는 벡터) 형태로 데이터를 출력할 수 있다.
- [0028] 대표적인 인공 신경망으로 영상 인식, 음성 인식, 자연어 처리, 필기체 인식 등에 주로 사용되는 컨볼루션 신경망(Convolution Neural Network: 이하 CNN)의 경우, 다수의 은닉 레이어가 연산 데이터를 할당된 가중치와 컨볼루션 연산을 수행하여 특징 패턴을 추출한다. 즉 다수의 은닉 레이어 각각은 일반적으로 대량의 곱셈 연산과 덧셈 연산을 수행하여 특징 패턴을 추출한다. CNN 이외에도 대부분의 인공 신경망은 곱셈 연산과 덧셈 연산을 반복 수행함으로써, 특징 패턴을 추출한다.
- [0029] 한편 인공 신경망은 학습 시에 학습 데이터를 인공 신경망의 입력 레이어로 입력하고, 출력 레이어에서 출력되는 출력 데이터와 학습 데이터에 대해 미리 지정된 결과 데이터 사이의 오차를 분석하고 분석된 오차를 역전파하여, 이전 할당된 가중치를 업데이트함으로써 학습한다. 즉 인공 신경망은 학습 데이터를 이용하여 가중치를 패턴 인식에 적합한 값으로 다시 할당함으로써 학습이 수행된다.
- [0030] 기존에는 소프트웨어적으로 구현된 인공 신경망을 범용 연산 프로세서와 메모리를 이용하여 실행함에 따라, 프로세서가 다수의 은닉 레이어 각각에 대한 연산 데이터와 가중치를 메모리로부터 인가받아 지정된 연산을 수행하고, 연산 수행 결과를 다시 메모리에 전달하여 저장함에 따라 효율성이 크게 떨어진다. 그에 반해, 인공 신경망을 메모리 칩 내부에 연산 로직을 이식한 PIM(process-in-memory)를 이용하여 실행하는 경우, 메모리 내에 구비된 연산 로직으로 데이터를 전송하므로 대용량 메모리의 장점을 유지하면서 대역폭을 향상시킬 수 있어 효율성을 크게 높일 수 있다. 즉 PIM은 인공 신경망 가속기로 유용하게 이용될 수 있다.
- [0031] 다만 일반적인 PIM 구조에서는 연산을 수행하는 연산 로직과 데이터를 저장하는 메모리 로직이 완전하게 구분되어 대역폭의 증가로 인한 효율성을 향상시킬 수 있으나 연산 로직이 메모리 로직으로부터 데이터 및 가중치를 전달받아 연산을 수행하고 다시 메모리 로직에 저장해야하는 기본 동작에는 변함이 없다.
- [0032] 또한 기존에는 통상적으로 부동소수점 포맷의 입력 데이터와 가중치를 디지털 데이터로 메모리에 저장하고 연산을 수행함에 따라 가중치 또는 입력 데이터 각각에 대해 다수의 메모리 셀이 필요하여 대량의 메모리 용량을 요구하였다.
- [0033] 한편, 최근 PIM은 in-memory라 불리는 멤리스터(Memristor) 소자로 메모리 셀을 구현될 수 있으며, PIM 구조의 인공 신경망 가속기에서 멤리스터 소자는 세미-아날로그 저항(전도도) 상태로 가중치를 저장할 수 있다. 메모리 셀이 아날로그적인 저항값으로 가중치 또는 입력 데이터를 저장하는 경우, 하나의 메모리 셀이 하나의 가중치 또는 입력 데이터를 저장할 수 있으므로, 요구되는 메모리 셀의 개수를 크게 줄일 수 있다. 다만, 메모리 셀이 아날로그적인 저항값으로 가중치 또는 입력 데이터를 안정적으로 저장하기 위해서는 저항값이 가역적이고 선형적이며, 점진적으로 증가 또는 감소할 수 있어야 한다.
- [0034] 도2 는 본 발명의 일 실시예에 따른 고속 인공 신경망 가속기의 개략적 구조를 나타내고, 도3 은 도2 의 메모리 어레이의 메모리 셀인 자성/비자성 다층 박막 메모리 구조의 일예를 나타낸다. 그리고 도4 는 자성/비자성 다층 박막 메모리의 전류에 따른 홀 저항 변화의 특성을 나타낸다.
- [0035] 도2 내지 도4 를 참조하면, 본 실시예에 따른 고속 인공 신경망 가속기는 메모리 셀 어레이(110), 제어부(120), 연산 제어부(130) 및 보조 연산부(140)를 포함할 수 있다.
- [0036] 메모리 셀 어레이(110)는 도3 에 도시된 바와 같이, 다수의 쓰기 라인(WL)과 다수의 비트 라인(BL) 및 다수의 읽기 라인(RL) 사이에 연결되고 멤리스터 소자인 자성/비자성 다층 박막 메모리(Magnetic/Nonmagnetic Multilayer Thin Film Memory)로 구현되는 다수의 메모리 셀(MC)을 포함한다. 여기서 쓰기 라인(WL), 비트 라인(BL) 및 읽기 라인(RL)은 설명의 편의를 위해 임의로 지정된 명칭으로, 다양하게 변경될 수 있다.
- [0037] 다수의 쓰기 라인(WL)과 다수의 읽기 라인(RL) 및 다수의 비트 라인(BL)은 연산 제어부(130)에 의해 활성화 또는 비활성화될 수 있다. 이때 다수의 쓰기 라인(WL)과 다수의 읽기 라인(RL) 및 다수의 비트 라인(BL) 중 적어도 하나는 트랜지스터 등으로 구현되는 스위치(미도시)를 통해 대응하는 메모리 셀(MC)과 연결되고, 연산 제어

부(130)는 각 라인의 스위치를 온/오프하여 메모리 셀(MC)에 대응하는 쓰기 라인(WL)과 읽기 라인(RL) 및 비트 라인(BL)을 활성화 또는 비활성화 할 수 있다.

- [0038] 도3 을 참조하면, 자성/비자성 다층 박막 메모리는 기관(310) 상에 배치되고 공급되는 쓰기 전류(I_w)에 따라 기관(310) 평면과 수직 방향의 스핀 전류를 생성하는 스핀 전류 생성층(320)과 스핀 전류 생성층(320) 상에 배치되어 스핀 전류에 의해 터널 자기저항(Tunnelling MagnetoResistance: TMR)이 가변되는 자기터널 접합층(330)을 포함한다.
- [0039] 스핀 전류 생성층(320)은 기관 상에 배치된 제1 비자성 도전층(323)과 제1 비자성 도전층(323) 상에 정렬되어 배치된 제2 비자성 도전층(329) 및 제1 비자성 도전층(323)과 제2 비자성 도전층(329) 사이에 배치되고 수직자기 이방성(perpendicular magnetic anisotropy)을 갖는 자성층(327)을 포함한다. 본 실시예에서 자성층(327)은 결정 구조를 가지며, 제1 비자성 도전층(323)과 제2 비자성 도전층(329)에 의하여 인장 변형력(tensile strain)을 받는다.
- [0040] 본 실시예에서 자성층(327)의 기준 벌크 격자 상수(reference bulk lattice constant)는 제1 비자성 도전층(323) 및 제2 비자성 도전층(329)의 격자 상수보다 작다. 그리고 자성층(327)의 기준 벌크 격자 상수에 대한 자성층(327)의 박막 격자 상수(thin film lattice constant)의 비로 주어지는 길이 방향의 변형률은 9 퍼센트 이상일 수 있다.
- [0041] 일례로 제1 및 제2 비자성 도전층(323, 329)은 팔라듐(Pd)이고, 자성층(327)은 코발트(Co)일 수 있다. 자성층(327)은 비자성 산화물층(Co_3O_4)에 대한 수소 이온 조사(Hydrogen ion irradiation)에 의해 상자성체에서 금속 상태의 코발트(Co) 박막의 강자성체로 상변화되어 자성층(327)을 형성함과 동시에, 박막 격자 상수를 상수는 제1 비자성 도전층(323) 및 제2 비자성 도전층(329)의 격자 상수와 유사하게 유지할 수 있어 수직자기 이방성을 발현할 수 있다.
- [0042] 비자성 산화물층(Co_3O_4)과 비자성 도전층(Pd)(323, 329) 구조($[\text{Co}_3\text{O}_4/\text{Pd}]_n$)에 대해 낮은 에너지의 수소 이온 또는 양성자를 조사하여 환원된 $[\text{Co}/\text{Pd}]_n$ 초격자(superlattice) 구조를 갖게 되며, 이는 통상적인 금속 $[\text{Co}/\text{Pd}]_n$ 초격자(superlattice) 구조보다 높은 수직자기 이방성을 나타낸다. 이에 환원된 $[\text{Co}/\text{Pd}]_n$ 초격자(superlattice) 구조는 통상적인 금속 $[\text{Co}/\text{Pd}]_n$ 초격자(superlattice) 구조보다 큰 스핀 궤도 토크를 제공할 수 있다. 즉 0.57nm의 격자 상수를 갖는 비자성 산화물층(Co_3O_4)은 자성층(327)의 박막 격자 상수는 제1 및 제2 비자성 도전층(323, 329)의 격자 상수(R)와 유사하게 될 수 있어, 증가된 계면 스핀 궤도 결합(Interfacial Spin-Orbit-Coupling) 강도를 나타낼 수 있다. 이는 응력이 최대화된 강자성층이 비자성층과 헤테로 접합인 스핀-궤도 결합이 큰 Co/Pd 전자밴드 구조를 형성하고, 수소를 조사하여 전하 개수를 변화시켜 페르미 에너지의 높낮이를 조절함으로써 스핀홀 전도도가 최대화되도록 조절할 수 있음을 의미한다. 여기서 스핀홀 전도도는 스핀 전류를 의미하며, 스핀홀 전도도는 스핀 궤도 토크에 비례한다.
- [0043] 한편 스핀 전류 생성층(320)은 기관(310)에 제1 비자성 도전층(323)이 결정 상태로 증착될 수 있도록 계면 상태를 제공하는 시드층(321)을 더 포함할 수 있다.
- [0044] 자기터널 접합층(330)은 스핀-전류 생성층(320) 상에 배치되어 스핀 전류에 의해 자기 모멘트의 방향이 가변되는 자유 자성층(331)과 자유 자성층(331) 상에 배치되는 터널 절연층(333) 및 터널 절연층(333) 상에 배치되는 고정 자성층(335)을 포함한다.
- [0045] 자유 자성층(331)은 스핀-전류 생성층(320)에서 생성된 스핀 전류에 의해 자기 모멘트의 방향이 가변된다. 자유 자성층(331)에서 자기 모멘트의 방향이 가변되면, 고정 자성층(335)의 자기 모멘트와의 방향성 차이로 인해, 자기터널 접합층(330)의 터널 저항(Tunneling resistance)이 가변된다. 즉 자기터널 접합층(330)의 저항의 변화는 자기터널 접합층(330)을 통과하는 스핀 전류에 의해 가변되는 자기터널 접합층(330)의 터널 자기저항(TMR)의 변화에 기인한 것으로, 스핀 전류 생성층(320)에 전류가 인가되지 않더라도 유지된다.
- [0046] 여기서 자기터널 접합층(330)의 터널 자기저항(TMR)을 증가시키기 위하여, 자유 자성층(331)과 고정 자성층(335)은 CoFeB로 구현될 수 있으며, 터널 절연층(333)은 MgO 로 구현될 수 있다. 자유 자성층(331) 및 고정 자성층(335)은 수직자기 이방성을 가질 수 있다.
- [0047] 도4 는 자성/비자성 다층 박막 메모리의 전류에 따른 홀 저항 변화의 특성으로, (a)는 스핀 전류 생성층(320)에

인가된 전류(I_x) 또는 전류 밀도와 수소이온 조사에 의해 생성된 자유 자성층(331)이 가해진 외부 자기장(H_z)에 따른 홀 저항(R_H)의 변화를 나타낸다. (a)에 도시된 바와 같이, 자유 자성층(331)의 홀 저항(R_H)은 가해진 외부 자기장(H_z)이 없거나 또는 외부 자기장(H_z)에 따라 전류(I_x)에 대한 홀 저항(R_H)의 변화 그래프가 가변된다. 그러나 (b)에 도시된 바와 같이 스핀 전류 생성층(320)은 가해진 외부 자기장(H_z)이 일정(여기서는 1170 Oe)한 경우, 홀 저항(R_H)은 스핀 전류 생성층(320)에 인가된 전류(I_x)에 따라 변화한다. 특히 (b)에서 전류(I_x)가 I_{min} (여기서는 일예로 26mA) 이하인 경우에 홀 저항(R_H)은 거의 변화하지 않고, 전류(I_x)가 I_{min} 에서 I_{max} 인 구간에서는 홀 저항(R_H)이 전류(I_x)에 비례하여 선형적으로 변화하며, 다시 I_{max} (여기서는 일예로 32mA) 이상인 구간에서는 거의 변화하지 않는 특성을 나타냄을 알 수 있다. 따라서 전류(I_x)가 I_{min} 에서 I_{max} 인 구간에서 홀 저항(R_H)이 전류(I_x)에 비례하여 선형적으로 변화하는 특징을 이용하여, (c)와 같이 자성/비자성 다층 박막 메모리는 시냅틱 가중치(W)를 홀 저항(R_H)의 형태로 저장할 수 있다.

[0048] 기존에 PIM 구조의 인공 신경망 가속기에서 메모리 셀(MC)은 일반적으로 선형적으로 가변되는 아날로그 값을 저장할 수 없어, 디지털 값을 저장하도록 구성되었다. 이 경우, 인공 신경망의 부동 소수점 포맷의 다수의 가중치(W)를 저장하기 위해 대량의 데이터 저장공간이 필요하고, 이에 메모리 셀 어레이(110)에 포함되어야 하는 메모리 셀(MC)의 개수가 대량으로 요구되었다.

[0049] 그러나 본 실시예에 따른 인공 신경망 가속기는 메모리 셀(MC)이 자성/비자성 다층 박막 메모리로 구현되고, (b)에 도시된 바와 같이, 홀 저항(R_H)이 전류(I_x)에 비례하여 선형적으로 변화하는 I_{min} 에서 I_{max} 인 구간을 이용하여 따라 메모리 셀(MC) 각각이 아날로그적인 값을 갖는 가중치(W)를 저장할 수 있다. 따라서 메모리 셀(MC)의 개수를 크게 줄일 수 있다.

[0050] 뿐만 아니라, 메모리 셀(MC)에 연산 데이터에 대응하는 연산 전류(I_0)를 인가하여, 메모리 셀(MC)에서 가중치(W)와 연산 데이터의 곱셈 연산이 직접 수행될 수 있다. 이는 인공 신경망에서 요구되는 곱셈 연산을 수행하기 위해, 가중치(W)와 연산 데이터를 각각 연산 프로세서로 로드할 필요가 없도록 하여 연산 속도를 크게 향상시킬 수 있으며 전력 소모를 줄일 수 있도록 한다.

[0051] 다시 도2 를 참조하면, 제어부(120)는 고속 인공 신경망 가속기(100)와 학습부(200) 사이의 인터페이스를 수행하고, 연산 제어부(130)를 제어하여, 메모리 셀 어레이(110)에 가중치(W)를 저장한다. 그리고 가중치(W)가 저장된 메모리 셀(MC)에 연산 데이터에 대응하는 연산 전류(I_0)가 전달되도록 하여 연산 데이터와 가중치(W)의 곱셈 연산이 수행되도록 할 수 있다. 또한 제어부(120)는 보조 연산부(140)를 제어하여, 연산 데이터와 가중치(W)의 곱셈 연산 결과에 대해 덧셈 연산 등의 추가 연산을 수행할 수 있다.

[0052] 연산 제어부(130)는 쓰기 동작 시에 제어부(120)의 제어에 따라 메모리 셀 어레이(110)에서 가중치(W)를 저장할 메모리 셀(MC)을 선택하고, 선택된 메모리 셀(MC)에 가중치(W)에 대응하는 쓰기 전류(I_w)를 생성하여 쓰기 라인(WL)을 통해 전달함으로써, 메모리 셀(MC)에 가중치(W)를 저장한다.

[0053] 연산 제어부(130)는 메모리 셀(MC)의 쓰기 동작 시에 쓰기 라인(WL)과 비트 라인(BL)을 활성화하고, 쓰기 라인(WL)을 통해 기지정된 문턱값 이상의 쓰기 전류(I_w)를 메모리 셀(MC)로 공급한다. 이때 읽기 라인(RL)은 비활성화되며, 문턱값은 도4 의 최소 전류(I_{min})일 수 있다.

[0054] 상기한 바와 같이, 쓰기 라인(WL)을 통해 쓰기 전류(I_w)가 공급되면, 스핀 전류 생성층(320)은 스핀 전류를 생성하고, 이에 자기터널 접합층(330)의 터널 자기저항(TMR)이 가변된다. 가변된 공급터널 자기저항(TMR)은 쓰기 전류(I_w)가 인가되지 않아도 유지되어 메모리 셀(MC)의 가중치(W)로서 저장된다.

[0055] 한편, 연산 제어부(130)는 메모리 셀(MC)의 연산 동작 시에 쓰기 라인(WL)을 비활성화하고, 읽기 라인(RL)과 비트 라인(BL)을 활성화한다. 연산 제어부(130)는 활성화된 비트 라인(BL)으로 연산 데이터에 대응하는 크기를 갖는 연산 전류(I_0)를 공급한다. 그리고 연산 데이터와 가중치(W)와의 곱셈 연산 결과로서 읽기 라인(RL)의 전압(또는 전류)를 감지한다. 여기서 연산 전류(I_0)는 문턱값 미만의 값을 가질 수 있다.

[0056] 메모리 셀(MC)의 터널 자기저항(TMR)이 가중치(W)로서 유지되고 있으므로, 연산 데이터에 대응하는 연산 전류(I_0)가 인가되면, 연산 전류(I_0)를 터널 자기저항(TMR)에 따라 감쇄하여 읽기 라인(RL)으로 전달한다. 연산 전

류(Io)가 연산 데이터에 대응하는 전류값을 가지고, 메모리 셀(MC)에 가중치(W)에 대응하는 저항값이 저장되므로, 읽기 라인(RL)의 전압은 $V = I \times R$ 의 공식에 따라 연산 데이터와 가중치(W)의 곱에 대응하는 값으로 출력된다. 즉 본 실시예에서 메모리 셀(MC)은 데이터를 터널 자기저항(TMR)으로 저장하는 멤리스터 소자로 구현되어 가중치(W)를 저장할 뿐만 아니라, 연산 데이터와 가중치(W)와의 곱셈 연산을 직접 수행하여 출력할 수 있다. 따라서 기존에 범용 연산 프로세서와 메모리가 구분된 구조로 구현되는 인공 신경망 가속기에 비해, 곱셈 연산을 위해 가중치를 연산 프로세서로 전달할 필요가 없어 고속, 저전력으로 고성능의 인공 신경망 가속기를 구현할 수 있다.

[0057] 연산 제어부(130)는 제어부(120)로부터 가중치(W) 또는 연산 데이터 인가되면, 디지털 데이터인 가중치(W) 또는 연산 데이터를 아날로그 전류로 변환하는 D/A 변환기(미도시)와 전압(또는 전류)로 출력되는 연산 결과를 디지털 데이터로 변환하기 위한 A/D 변환기(미도시)를 포함할 수 있다.

[0058] 상기에서는 비트 라인(BL)을 통해 연산 전류(Io)가 메모리 셀(MC)로 인가되고, 읽기 라인(RL)에서 연산 데이터와 가중치(W)의 곱셈 연산의 결과인 전압(또는 전류)를 측정하는 것으로 설명하였으나, 경우에 따라서는 읽기 라인(RL)을 통해 메모리 셀(MC)로 연산 전류(Io)가 공급되고, 비트 라인(BL)에서 연산 결과를 측정하도록 구성될 수도 있다.

[0059] 한편, 연산 제어부(130)는 연산 동작과 유사하게 비트 라인(BL) 또는 읽기 라인(RL) 중 하나로 기지정된 크기의 읽기 전류(I_R)를 공급하고, 나머지 하나에서 전압(또는 전류)를 측정하여 메모리 셀(MC)에 저장된 가중치(W)를 판독할 수 있다.

[0060] 보조 연산부(140)는 인공 신경망에서 요구되는 연산 중 곱셈 연산을 제외한 나머지에 대한 연산을 수행한다. 상기한 바와 같이, 인공 신경망, 특히 CNN의 경우, 다수의 곱셈 연산이 요구되지만 곱셈 연산 이외에도 덧셈 연산 또는 평균값 연산 등이 요구된다. 이러한 추가적으로 요구되는 연산을 수행하기 위해, 인공 신경망 가속기(100)는 보조 연산부(140)를 포함할 수 있다. 도2에서는 설명의 편의를 위하여 보조 연산부(140)를 별도의 구성으로 도시하였으나 보조 연산부(140)는 연산 제어부(130)에 포함될 수 있으며, 경우에 따라서는 제어부(120)에 포함될 수도 있다.

[0061] 또한 연산 제어부(130)는 메모리 셀 어레이(110)의 다수의 메모리 셀(MC)에서 수행된 연산 결과인 전압을 직렬로 연결하여 곧바로 덧셈 연산에 대한 결과를 획득하도록 구성될 수도 있다.

[0062] 그리고 상기에서는 메모리 셀 어레이(110)에 가중치(W)가 저장되는 것으로 설명하였으나, 입력 데이터 또는 연산 데이터도 메모리 셀 어레이(110)의 메모리 셀(MC)에 저장될 수 있으며, 연산 결과인 연산 결과 데이터 또한 메모리 셀 어레이(110)의 메모리 셀(MC)에 저장될 수도 있다.

[0063] 학습부(200)는 학습용 데이터를 인공 신경망 가속기(100)로 전달하고, 학습용 데이터에 대한 연산 결과를 인가받아 오차를 판별하고, 판별된 오차에 따라 메모리 셀 어레이(110)에 저장된 가중치(W)를 업데이트하여 제어부(120)로 전달한다.

[0064] 상기에서는 학습부(200)를 인공 신경망 가속기(100)와 별도로 구성되는 것으로 도시하였으나, 학습부(200)는 인공 신경망 가속기(100)에 포함되어 구성될 수도 있다. 일반적으로 학습부(200)는 인공 신경망을 학습시켜 최적의 가중치(W)를 획득하기 위해 이용되며, 학습이 완료된 인공 신경망에서는 이용되지 않으므로, 인공 신경망 가속기(100)와 별도로 구성된다. 인공 신경망의 학습 과정은 상기한 바와 같이, 최적의 가중치(W)를 획득하기 위해 많은 반복 연산을 수행하게 되며, 이 과정에서 대량의 데이터 전송이 필요하게 된다. 즉 학습에 매우 긴 시간을 요구하게 된다.

[0065] 그러나 학습부(200)가 인공 신경망 가속기(100)에 포함되어 구성되는 경우, 이러한 학습 시간을 획기적으로 줄일 수 있다. 이 경우 제어부(120) 또는 보조 연산부(140)가 학습부(200)의 기능을 수행하도록 구성될 수 있다. 제어부(120) 또는 보조 연산부(140)가 학습부(200)의 기능을 함께 수행하도록 구성되는 경우, 인공 신경망 가속기(100)는 학습 시에도 매우 빠른 속도로 학습을 수행할 수 있으며 전력 소모를 크게 줄일 수 있다.

[0066] 도5는 도2의 메모리 어레이에서 메모리 셀의 구조의 다른 예를 나타낸다.

[0067] 도5에 도시된 자성/비자성 다층 박막 메모리 소자에서 자기터널 접합층(530)은 도3의 자기터널 접합층(330)과 동일한 구조를 갖는다. 즉 자기터널 접합층(530)은 스핀 전류 생성층(520) 상에 배치되어 스핀 전류에 의해 자기 모멘트의 방향이 가변되는 자유 자성층(531)과 자유 자성층(531) 상에 배치되는 터널 절연층(533) 및 터널 절연층(533) 상에 배치되는 고정 자성층(535)을 포함한다.

- [0068] 그러나 도5 에 도시된 자성/비자성 다층 박막 메모리 소자에서 스핀 전류 생성층(520)은 도3 의 자성/비자성 다층 박막 메모리 소자와 상이한 구조를 갖는다.
- [0069] 도3 의 자성/비자성 다층 박막 메모리 소자에서 스핀 전류 생성층(320)이 기판 상에 배치된 제1 비자성 도전층(323)과 제1 비자성 도전층(323) 상에 정렬되어 배치된 제2 비자성 도전층(329) 및 제1 비자성 도전층(323)과 제2 비자성 도전층(329) 사이에 배치되는 자성층(327)을 포함하도록 구성되었다. 그리고 도5 의 자성/비자성 다층 박막 메모리 소자의 스핀 전류 생성층(520) 또한 제1 비자성 도전층(523)과 제2 비자성 도전층(529) 사이에 자성층(527)을 포함한다. 그러나 도5 에서 스핀 전류 생성층(520)은 제1 비자성 도전층(523)과 자성층(527) 사이에 교대로 배치되는 다수의 보조 자성층(524a ~ 524n) 및 다수의 보조 비자성 도전층(525a ~ 525n)을 더 포함하는 다층 박막 구조를 가질 수 있다. 그리고 다수의 보조 비자성 도전층(525a ~ 525n) 중 적어도 하나에는 쓰기 라인(WL)을 통해 공급되는 전류에 따라 스핀 전류가 생성될 수 있다.
- [0070] 자성층(527)과 다수의 보조 자성층(524a ~ 524n)은 인접한 비자성 도전층(523, 529) 및 보조 비자성 도전층(525a ~ 525n)에 의해 인장 변형력을 받을 수 있으며, 이에 따라 도5 의 자성/비자성 다층 박막 메모리 소자는 도3 의 자성/비자성 다층 박막 메모리 소자에 비해 동일한 쓰기 전류가 인가되더라도 증가된 스핀-궤도 결합 강도를 나타낼 수 있다. 즉 적은 전류로 메모리 셀(MC)의 터널 자기저항(TMR)을 더욱 용이하게 가변할 수 있다.
- [0071] 도6 은 자성/비자성 다층 박막 메모리 소자로 구현된 메모리 셀의 3차원 적층 구조의 일예를 나타낸다.
- [0072] 최근 인공 신경망의 발전으로 인해, 3차원 인공 신경망에 대한 연구가 수행되고 있다. 3차원 인공 신경망은 다수의 프레임으로 구성된 동영상의 변화 등을 분석하기 위해 주로 이용되고 있다. 즉 연속하는 2차원 영상인 다수의 프레임 사이의 변화나 특징을 추출하기 위해 이용된다. 이러한 3차원 인공 신경망을 위한 연산을 수행하기 위해서는 다수의 2차원 행렬(또는 벡터)를 반복적으로 연산하도록 할 수도 있으나, 본 실시예에서는 가중치(W)와 연산 데이터 간의 연산을 직접 수행할 수 있는 다수의 메모리 셀(MC)을 3차원 구조로 적층하여 구성함으로써, 3차원 연산을 일괄적으로 수행할 수 있도록 할 수도 있다. 이렇게 다수의 메모리 셀(MC)을 3차원 구조로 적층하여 구성하는 경우, 3차원 인공 신경망의 구조를 그대로 다수의 메모리 셀(MC)에 구현할 수 있어, 인공 신경망의 구조를 분석하기에 용이할 뿐만 아니라 고속 연산이 가능하다. 또한 3차원 적층 구조가 가능함에 따라 초소형 대용량의 인공 신경망 가속기를 구현할 수 있다.
- [0073] 기존의 멤리스터 소자 중 ReRAM 메모리 소자 자체는 크로스 바 어레이를 수직방향으로 집적하여 3차원 적층이 용이한 구조이지만, 주변회로로 인해 3차원 적층이 불가능하다는 한계가 있다. 그에 비해 본 실시예에 따른 자성/비자성 다층 박막 메모리 소자는 터널 자기저항(TMR)이 가역적이고 선형적이며, 점진적으로 증가 또는 감소될 수 있어 보상을 위한 주변 회로를 최소화하거나 제거할 수 있다. 따라서 3차원 적층 구조로 구현하기 용이하다.
- [0074] 도6 에서는 일예로 자성/비자성 다층 박막 메모리 소자가 2단으로 적층된 3차원 메모리 셀 구조를 나타내고 있다. 도6 을 참조하면, 기판(410) 상에 배치되는 제1 자성/비자성 다층 박막 메모리 소자는 도3 의 자성/비자성 다층 박막 메모리 소자와 동일하게 구성될 수 있다. 그리고 제1 자성/비자성 다층 박막 메모리 소자의 상부면에는 읽기 라인(RL)과 선택 트랜지스터(TR)가 형성되는 데이터 전달층(440a)이 형성될 수 있으며, 데이터 전달층(440a) 상에는 절연층(450)이 형성될 수 있다. 그리고 절연층(450) 상에 제1 자성/비자성 다층 박막 메모리 소자와 동일한 구조로 제2 자성/비자성 다층 박막 메모리 소자가 형성될 수 있다.
- [0075] 도6 에서는 간략한 일예로 2단 적층된 3차원 메모리 셀 구조를 도시하였으나, 본 실시예는 이에 한정되지 않는다. 즉 3단 이상으로 적층 가능하다.
- [0076] 도7 은 본 발명의 일 실시예에 따른 자성/비자성 다층 박막 메모리를 이용한 고속 인공 신경망 가속기의 운용 방법을 나타낸다.
- [0077] 도2 내지 도6 을 참조하여, 도7 의 고속 인공 신경망 가속기의 운용 방법을 설명하면, 우선 학습 과정 동안 학습을 통해 최종적으로 업데이트 된 가중치(W)가 메모리 셀 어레이(110)의 다수의 메모리 셀(MC) 중 기지정된 메모리 셀에 저장된다(S10). 여기서 다수의 메모리 셀(MC) 각각은 자성/비자성 다층 박막 메모리 소자로 구현될 수 있으며, 가중치(W)는 아날로그 값을 갖는 터널 자기저항(TMR)의 형태로 메모리 셀(MC)에 저장될 수 있다. 특히 다수의 메모리 셀(MC) 각각은 쓰기 라인(WL)을 통해 저장할 가중치(W)에 대응하고 문턱값 이상으로 인가되는 쓰기 전류(I_w)에 대응하여 터널 자기저항(TMR)이 선형적으로 가변될 수 있다. 그리고 가중치(W)가 저장되는 메모리 셀(MC)의 개수는 인공 신경망의 구조 및 크기에 따라 다양하게 조절될 수 있다.

- [0078] 다수의 메모리 셀(MC)에 가중치(W)가 아날로그적 저항값의 형태로 저장되면, 가중치(W)와 연산이 수행되어야 하는 연산 데이터를 획득한다(S20). 여기서 연산 데이터는 인공 신경망의 입력 데이터이거나, 이전 레이어에서 연산된 결과 데이터일 수 있다. 그리고 연산 데이터는 인공 신경망 가속기의 외부로부터 전달된 데이터일 수 있으며, 메모리 셀 어레이(110)의 다수의 메모리 셀(MC) 중 가중치(W)가 저장되지 않은 메모리 셀(MC)에 기저장된 데이터일 수도 있다. 만일 연산 데이터가 메모리 셀(MC)에 기저장된 데이터인 경우, 인공 신경망 가속기는 연산 데이터가 저장된 메모리 셀(MC)로 문턱값 미만의 기지정된 전류값을 갖는 읽기 전류를 공급하고 메모리 셀(MC)에서 출력되는 전압(또는 전류)을 측정함으로써, 연산 데이터를 획득할 수 있다.
- [0079] 연산 데이터가 획득되면, 연산 데이터를 연산 전류(I_o)로 변환한다(S30). 여기서 연산 전류(I_o)는 연산 데이터에 대응하는 전류값을 갖되, 기지정된 문턱값 미만의 전류값을 갖는다.
- [0080] 그리고 변환된 연산 전류(I_o)를 가중치(W)가 저장된 메모리 셀(MC)로 공급하여 메모리 셀(MC)을 이용하여 가중치(W)와 연산 데이터의 연산을 수행한다(S40). 가중치(W)에 대응하는 저항값을 갖는 메모리 셀(MC)에 연산 전류(I_o)가 공급되면, 메모리 셀(MC)에서 가중치(W)와 연산 데이터의 곱에 대응하는 전압(또는 전류)이 출력된다. 즉 메모리 셀(MC)에서 출력되는 전압(또는 전류)을 감지하여, 이용하여 가중치(W)와 연산 데이터의 곱셈 연산 결과를 획득할 수 있다.
- [0081] 그리고 가중치(W)와 연산 데이터의 곱셈 연산 결과에 대해 인공 신경망의 구성에 따라 추가적인 후처리 연산을 수행한다(S50). 여기서 후처리 연산은 덧셈 연산, 평균 연산 등이 포함될 수 있다.
- [0082] 이후 후처리 연산의 결과를 다시 메모리 셀 어레이(110)의 메모리 셀(MC)에 저장되거나 출력될 있다(S60). 저장된 연산 결과는 인공 신경망의 다음 레이어의 연산을 위한 연산 데이터로 다시 이용될 수 있다. 여기서 연산 결과는 가중치(W)를 메모리 셀에 저장하는 경우와 동일하게 문턱값 이상의 연산 데이터에 대응하는 전류값을 갖는 쓰기 전류(I_w)를 생성하여 메모리 셀(MC)로 공급함으로써, 저장될 수 있다.
- [0083] 도8 은 도7 의 학습 가중치 저장 단계를 상세하게 나타낸다.
- [0084] 여기서는 도2 에서 별도로 도시된 학습부(200)가 인공 신경망 가속기(100)에 포함되어 구성된 것으로 가정하여 설명한다.
- [0085] 도8 을 참조하면, 학습 가중치 저장 단계(S10)에서는 우선 기지정된 학습 가중치(W)를 지정된 메모리 셀(MC)에 저장한다(S11). 여기서 학습 가중치(W)는 인공 신경망의 학습 과정에 획득된 가중치로서 학습을 통해 계속적으로 업데이트 될 수 있으며, 초기값은 일예로 1일 수 있다. 여기서 학습 가중치(W)에 대응하는 문턱값 이상의 쓰기 전류(I_w)를 메모리 셀(MC)로 전달함으로써, 메모리 셀(MC)에 이에 대응하는 터널 자기저항(TMR)이 저장될 수 있다.
- [0086] 그리고 학습을 수행하기 위한 학습 데이터를 획득한다(S12). 여기서 학습 데이터는 외부에서 인가되어 저장될 수 있으며, 학습 가중치(W)와 마찬가지로, 학습 데이터에 대응하는 전류를 메모리 셀(MC)로 전달하여 미리 저장될 수 있다. 그리고 학습 데이터가 메모리 셀(MC)에 이미 저장된 상태이면, 메모리 셀(MC)에 문턱값 미만의 읽기 전류(I_r)를 공급하고, 메모리 셀(MC)에서 출력되는 전류(또는 전압)을 감지하여 학습 데이터를 판별할 수 있다.
- [0087] 학습 데이터가 획득되면, 학습 데이터를 연산 전류(I_o)로 변환한다(S13). 그리고 변환된 연산 전류(I_o)를 학습 가중치(W)가 저장된 메모리 셀(MC)로 공급하여 학습 데이터와 학습 가중치 사이의 연산을 수행한다(S14). 여기서 메모리 셀 연산은 곱셈 연산이며, 학습 과정에서 가중치가 최종적으로 업데이트 되지 않은 상태의 연산으로 볼 수 있다. 메모리 셀 연산이 수행되면, 메모리 셀 연산 결과에 대한 기지정된 후처리 연산을 수행한다(S15).
- [0088] 여기서 학습 데이터 전류 변환 단계(S13)와 메모리 셀 연산 단계(S14) 및 연산 후처리 단계(S15)는 인공 신경망의 레이어의 수에 따라 반복적으로 수행될 수 있다. 그리고 인공 신경망의 연산 결과를 기지정된 검증 데이터 등과 비교하여 오차를 판별한다(S16). 오차가 판별되면, 판별된 오차를 이용하여 학습 가중치(W)를 업데이트한다(S17). 그리고 업데이트된 학습 가중치를 대응하는 쓰기 전류(I_w)로 변환한다(S18). 학습 가중치에 대응하는 쓰기 전류(I_w)를 메모리 셀(MC)로 공급하여 업데이트된 가중치(W)를 저장한다(S19).
- [0089] 상기한 바와 같이, 본 실시예에 따른 고속 인공 신경망 가속기 및 이의 운용 방법은 인공 신경망의 가중치 및 연산 데이터 중 적어도 하나를 아날로그적 저항값의 형태로 저장할 수 있는 자성/비자성 다층 박막 메모리를 메모리 셀로 이용하여 메모리 칩 내부에 연산 로직이 이식된 고효율 PIM 구조의 인공 신경망 가속기를 구현함으로

써, 가중치 또는 연산 데이터를 저장하기 위한 메모리 용량을 크게 줄일 수 있다. 그리고 가중치가 저장된 메모리 셀에 연산 데이터를 전달하여, 메모리 셀 각각에서 가중치와 연산 데이터 사이의 연산 결과가 출력되도록 함으로써, 연산을 위한 데이터 전송을 최소화할 수 있을 뿐만 아니라 대규모 병렬 연산을 간단하게 수행할 수 있어, 고속 고효율의 연산을 수행할 수 있도록 하며, 전력 소모를 크게 줄일 수 있다.

[0090] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0091] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

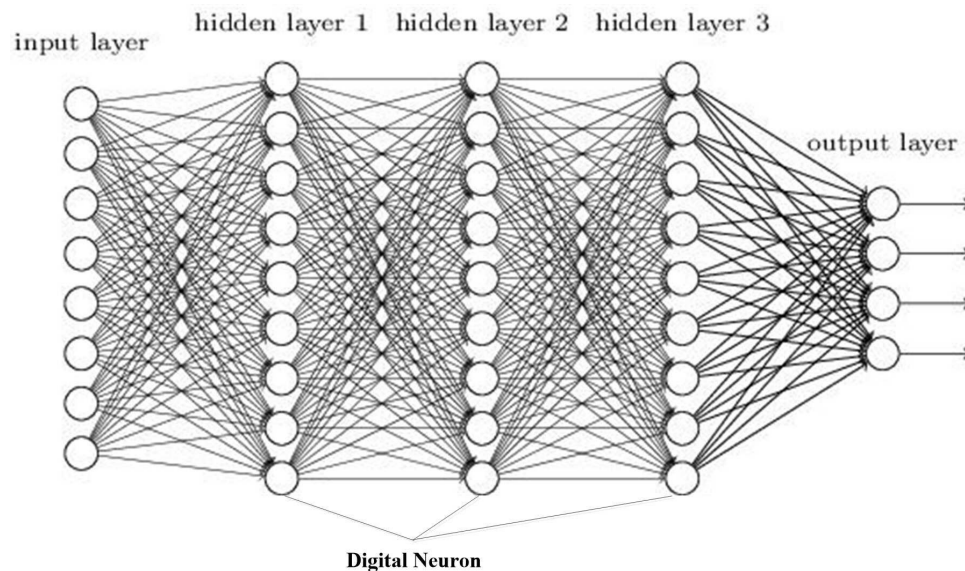
[0092] 100: 인공 신경망 가속기 200: 학습부

110: 메모리 셀 어레이 120: 제어부

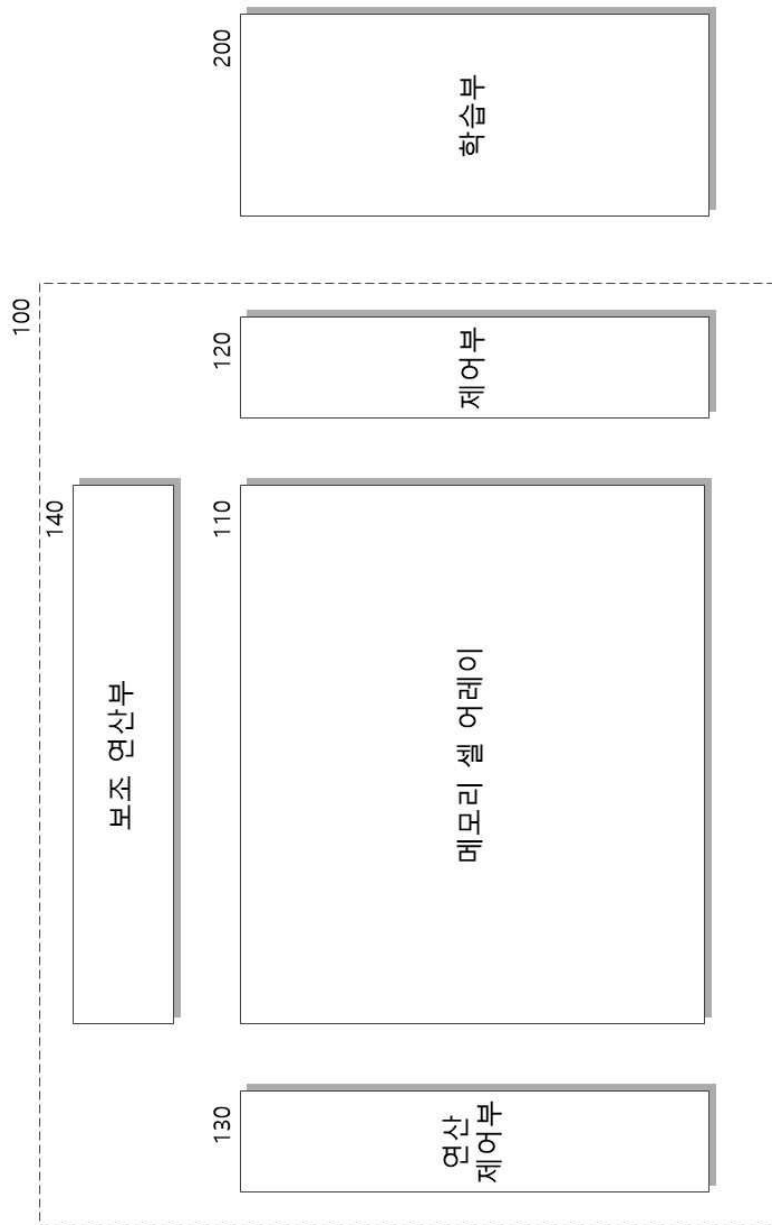
130: 연산 제어부 140: 보조 연산부

도면

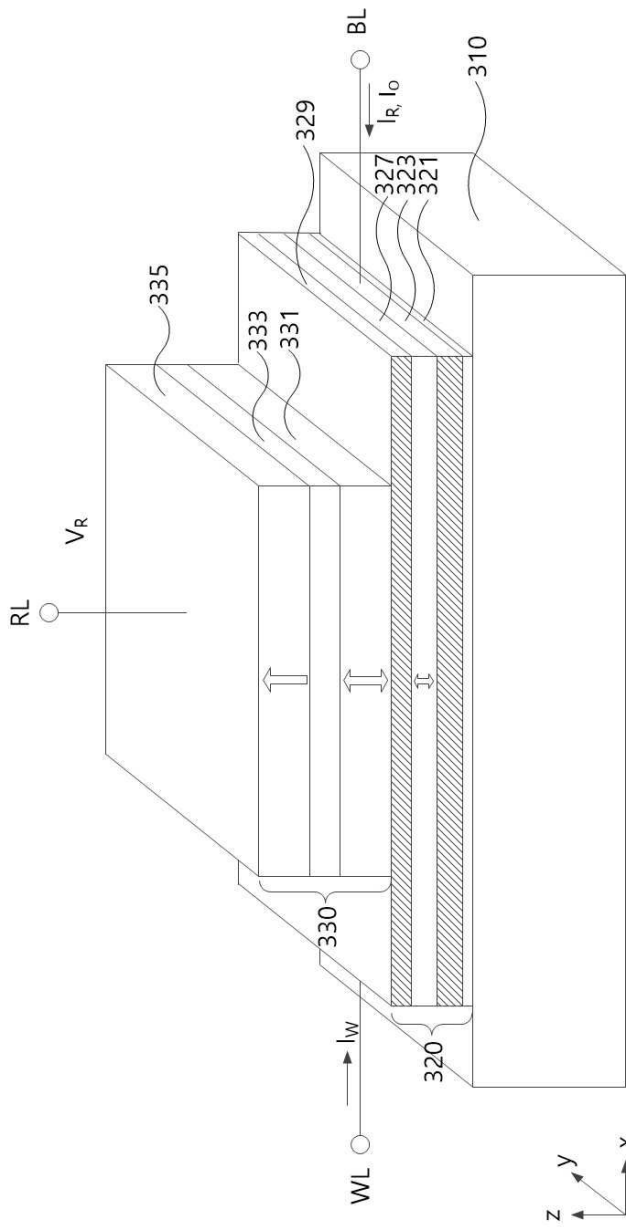
도면1



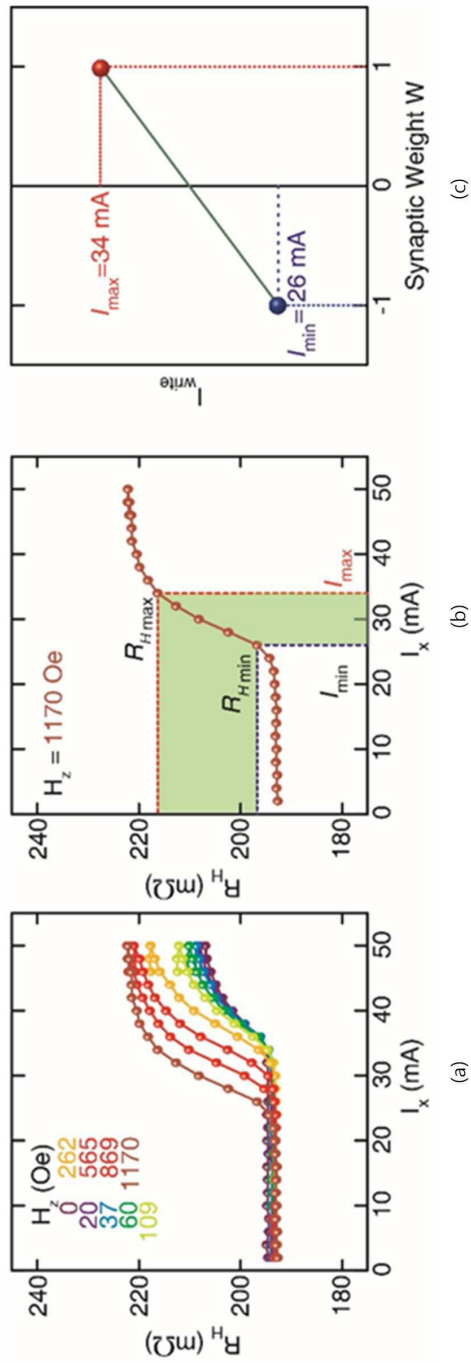
도면2



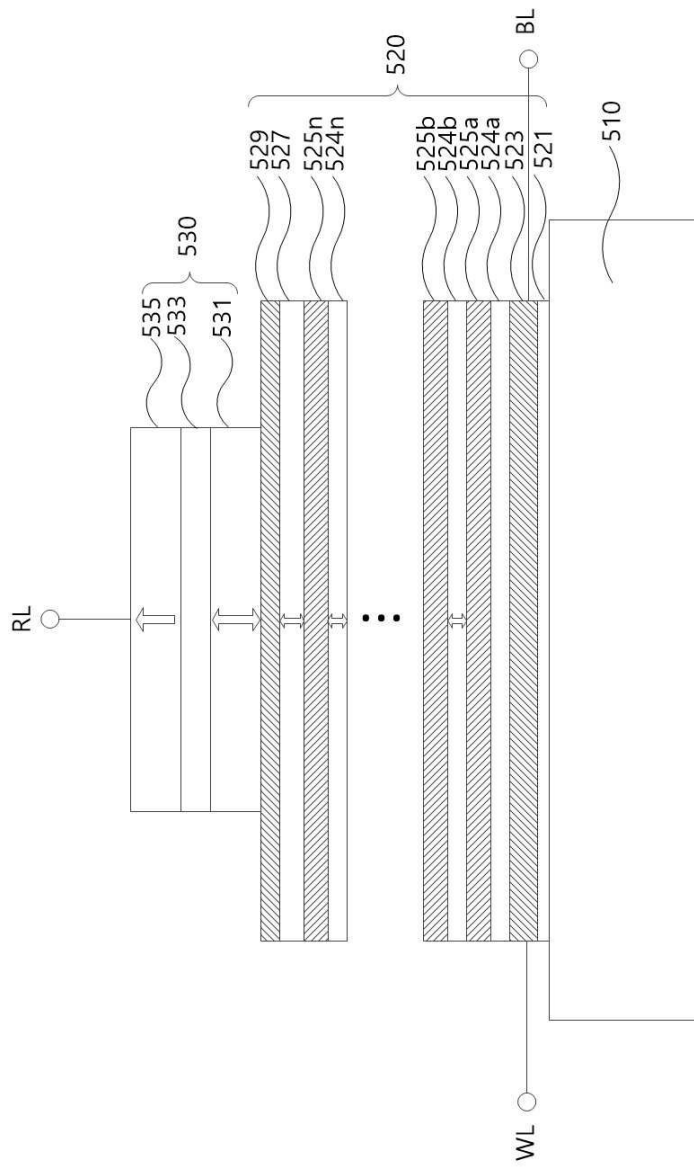
도면3



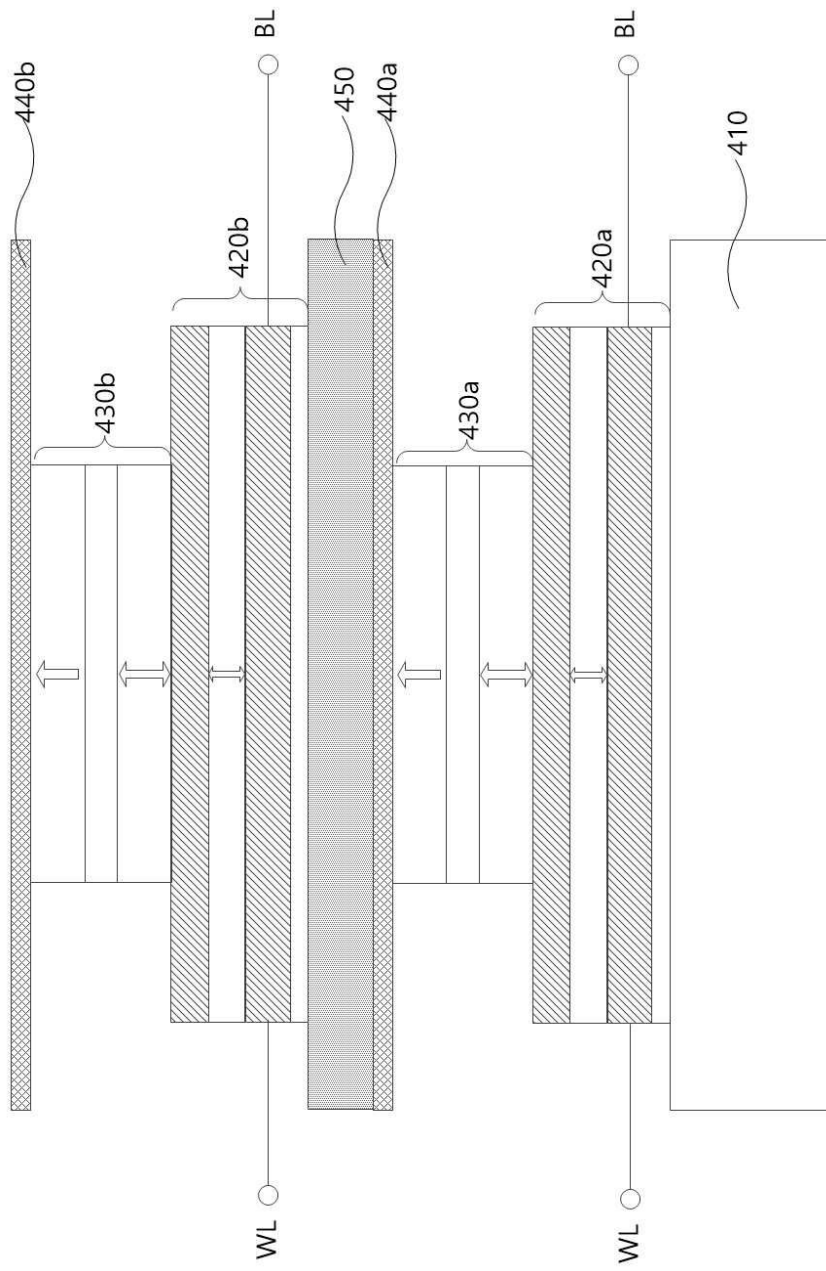
도면4



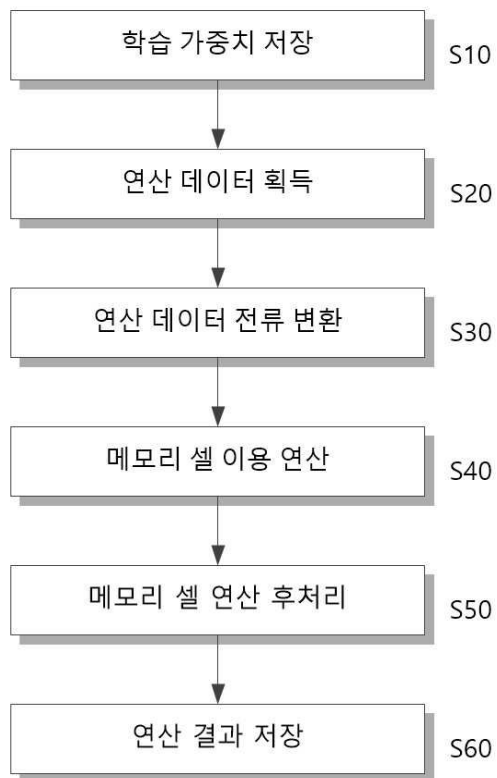
도면5



도면6



도면7



도면8

