



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0062444
(43) 공개일자 2024년05월09일

- | | |
|--|--|
| <p>(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) G06F 17/16 (2006.01)
G06N 3/063 (2023.01) G06N 3/10 (2006.01)</p> <p>(52) CPC특허분류
G06N 3/084 (2023.01)
G06F 17/16 (2013.01)</p> <p>(21) 출원번호 10-2022-0143557
(22) 출원일자 2022년11월01일
심사청구일자 2022년11월01일</p> | <p>(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50</p> <p>(72) 발명자
이종은
울산광역시 울주군 언양읍 유니스트길 50
이수길
울산광역시 울주군 언양읍 유니스트길 50</p> <p>(74) 대리인
특허법인 무한</p> |
|--|--|

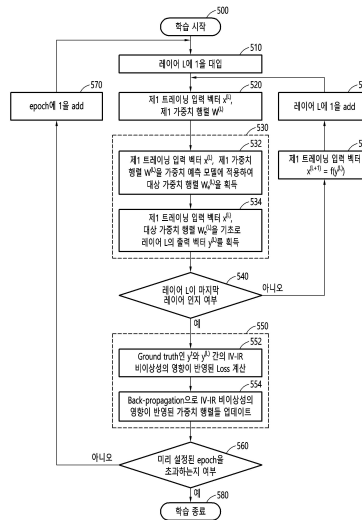
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 저항성 랜덤 액세스 메모리 크로스바 어레이의 비이상성을 극복하기 위한 전자 장치 및 방법

(57) 요약

일 실시예에 따른 전자 장치는, 컴퓨터로 실행 가능한 명령어들(computer-executable instructions)이 저장된 메모리; 및 상기 메모리에 액세스(access)하여 상기 명령어들을 실행하는 프로세서를 포함하고, 상기 명령어들은, 제1 트레이닝 입력 벡터로부터 제1 확장 입력 데이터를 상기 심층 신경망의 제1 가중치 행렬의 차원에 기초하여 생성하고, 상기 제1 가중치 행렬 및 제1 확장 입력 데이터의 결합에 기초하여 생성된 텐서 데이터를 가중치 예측 모델에 적용함으로써, 상기 제1 가중치 행렬이 RCA(ReRAM crossbar arrays) 회로의 아날로그 비이상성(analog non-ideality)에 따라 변화된 대상 가중치 행렬을 획득하고, 상기 제1 트레이닝 입력 벡터 및 상기 대상 가중치 행렬에 기초한 순전파(forward propagation)를 포함하는 연산을 통해 계산된 상기 심층 신경망의 손실을 이용하여 역전파(back propagation)시키면서 상기 심층 신경망을 트레이닝시킬 수 있다.

대표도 - 도5



(52) CPC특허분류

G06N 3/063 (2013.01)

G06N 3/10 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711165109
과제번호	2020R1A2C2015066
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	RRAM 기반 뉴럴넷 시스템 설계를 위한 고속 시뮬레이션 기술 연구
기 여 율	60/100
과제수행기관명	울산과학기술원
연구기간	2022.03.01 ~ 2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711139247
과제번호	2021-0-02052-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	스마트 모빌리티를 위한 인공지능 시스템반도체 핵심 기술 개발 및 인력 양성
기 여 율	40/100
과제수행기관명	인하대학교 산학협력단
연구기간	2021.07.01 ~ 2021.12.31

명세서

청구범위

청구항 1

프로세서에 의하여 수행되는 심층 신경망의 학습 방법에 있어서,

제1 트레이닝 입력 벡터로부터 제1 확장 입력 데이터를 상기 심층 신경망의 제1 가중치 행렬의 차원에 기초하여 생성하는 단계;

상기 제1 가중치 행렬 및 제1 확장 입력 데이터의 결합에 기초하여 생성된 텐서 데이터를 가중치 예측 모델에 적용함으로써, 상기 제1 가중치 행렬이 RCA(ReRAM crossbar arrays) 회로의 아날로그 비이상성(analog non-ideality)에 따라 변화된 대상 가중치 행렬을 획득하는 단계; 및

상기 제1 트레이닝 입력 벡터 및 상기 대상 가중치 행렬에 기초한 순전파(forward propagation)를 포함하는 연산을 통해 계산된 상기 심층 신경망의 손실을 이용하여 역전파(back propagation)시키면서 상기 심층 신경망을 트레이닝시키는 단계

를 포함하는 심층 신경망 학습 방법.

청구항 2

제1항에 있어서,

상기 제1 확장 입력 데이터를 생성하는 단계는,

상기 제1 트레이닝 입력 벡터의 차원의 개수만큼, 상기 제1 트레이닝 입력 벡터와 동일한 복제 벡터를 생성하는 단계; 및

상기 복수의 복제 벡터들을, 상기 제1 가중치 행렬의 차원에 기초하여, 미리 결정된(predetermined) 방향으로 결합하여 상기 제1 확장 입력 데이터를 생성하는 단계

를 포함하는 심층 신경망 학습 방법.

청구항 3

제1항에 있어서,

제2 트레이닝 입력 벡터 및 트레이닝 가중치 행렬을, 상기 RCA 회로 기반의 심층 신경망 가속기를 구현한 하드웨어 시뮬레이션에 적용함으로써, 상기 RCA 회로의 아날로그 비이상성에 따라 변화된 가중치를 나타내는 실수 값 가중치 행렬을 획득하는 단계; 및

상기 제2 트레이닝 입력 벡터 및 상기 트레이닝 가중치 행렬로부터 상기 실수 값 가중치 행렬을 출력하도록, 상기 가중치 예측 모델을 트레이닝시키는 단계;

를 더 포함하는 심층 신경망 학습 방법.

청구항 4

제1항에 있어서,

상기 대상 가중치 행렬을 획득하는 단계는,

상기 제1 트레이닝 입력 벡터의 원소 데이터가 이진(binary) 데이터를 포함하는 경우에 기초하여, 적어도 2개 이상의 트레이닝 입력 벡터들을 포함하는 트레이닝 입력 벡터 집합에서 트레이닝 입력 벡터들의 평균 데이터를

획득하는 단계; 및

상기 획득된 평균 데이터에 기초하여, 상기 트레이닝 입력 벡터 집합의 분포와 동일한 분포를 따르는 원소 데이터들을 포함하는 무작위(random) 입력 벡터를 생성하는 단계

를 포함하는 심층 신경망 학습 방법.

청구항 5

제1항에 있어서,

상기 RCA 회로의 아날로그 비이상성으로 인해 유발되는 오차가 보상되는 방법으로 학습된 심층 신경망의 가중치 행렬을 추출하는 단계;

상기 추출된 가중치 행렬을 상기 RCA 회로의 저항성 소자로 매핑(mapping)하는 단계;

상기 심층 신경망에 적용되는 입력 벡터를 상기 RCA 회로의 입력 비트 라인으로 매핑하는 단계; 및

상기 RCA 회로에 의해, 상기 입력 벡터 및 상기 학습된 가중치 행렬의 행렬-벡터 곱 연산에 의해 출력 벡터를 획득하는 단계

를 더 포함하는 방법.

청구항 6

하드웨어와 결합되어 제1항 내지 제5항 중 어느 하나의 항의 방법을 실행시키기 위하여 컴퓨터 판독 가능한 기록매체에 저장된 컴퓨터 프로그램.

청구항 7

심층 신경망의 학습 전자 장치에 있어서,

컴퓨터로 실행 가능한 명령어들(computer-executable instructions)이 저장된 메모리; 및

상기 메모리에 액세스(access)하여 상기 명령어들을 실행하는 프로세서

를 포함하고,

상기 명령어들은,

제1 트레이닝 입력 벡터로부터 제1 확장 입력 데이터를 상기 심층 신경망의 제1 가중치 행렬의 차원에 기초하여 생성하고,

상기 제1 가중치 행렬 및 제1 확장 입력 데이터의 결합에 기초하여 생성된 텐서 데이터를 가중치 예측 모델에 적용함으로써, 상기 제1 가중치 행렬이 RCA(ReRAM crossbar arrays) 회로의 아날로그 비이상성(analog non-ideality)에 따라 변화된 대상 가중치 행렬을 획득하고,

상기 제1 트레이닝 입력 벡터 및 상기 대상 가중치 행렬에 기초한 순전파(forward propagation)를 포함하는 연산을 통해 계산된 상기 심층 신경망의 손실을 이용하여 역전파(back propagation)시키면서 상기 심층 신경망을 트레이닝시키는

전자 장치.

청구항 8

제7항에 있어서,

상기 프로세서는,

상기 제1 트레이닝 입력 벡터의 차원의 개수만큼, 상기 제1 트레이닝 입력 벡터와 동일한 복제 벡터를 생성하고,

상기 복수의 복제 벡터들을, 상기 제1 가중치 행렬의 차원에 기초하여, 미리 결정된(predetermined) 방향으로 결합하여 상기 제1 확장 입력 데이터를 생성하는

전자 장치.

청구항 9

제7항에 있어서,

상기 프로세서는,

제2 트레이닝 입력 벡터 및 트레이닝 가중치 행렬을, 상기 RCA 회로 기반의 심층 신경망 가속기를 구현한 하드웨어 시뮬레이션에 적용함으로써, 상기 RCA 회로의 아날로그 비이상성에 따라 변화된 가중치를 나타내는 실수 값 가중치 행렬을 획득하고,

상기 제2 트레이닝 입력 벡터 및 상기 트레이닝 가중치 행렬로부터 상기 실수 값 가중치 행렬을 출력하도록, 상기 가중치 예측 모델을 트레이닝 시키는

전자 장치.

청구항 10

제9항에 있어서,

상기 프로세서는,

상기 제1 트레이닝 입력 벡터의 원소 데이터가 이진(binary) 데이터를 포함하는 경우에 기초하여, 적어도 2개 이상의 트레이닝 입력 벡터들을 포함하는 트레이닝 입력 벡터 집합에서 트레이닝 입력 벡터들의 평균 데이터를 획득하고,

상기 획득된 평균 데이터에 기초하여, 상기 트레이닝 입력 벡터 집합의 분포와 동일한 분포를 따르는 원소 데이터를 포함하는 무작위(random) 입력 벡터를 생성하는

전자 장치.

청구항 11

제7항에 있어서,

상기 프로세서는,

상기 RCA 회로의 아날로그 비이상성으로 인해 유발되는 오차가 보상되는 방법으로 학습된 심층 신경망의 가중치 행렬을 추출하고,

상기 추출된 가중치 행렬을 상기 RCA 회로의 저항성 소자로 매핑(mapping)하고,

상기 심층 신경망에 적용되는 입력 벡터를 상기 RCA 회로의 입력 비트 라인으로 매핑하고,

상기 RCA 회로에 의해, 상기 입력 벡터 및 상기 학습된 가중치 행렬의 행렬-벡터 곱 연산에 의해 출력 벡터를 획득하는

전자 장치.

발명의 설명

기술 분야

[0001] 이하, 저항성 랜덤 액세스 메모리 크로스바 어레이의 비이상성을 극복하는 방법 및 장치에 관한 기술이 제공된다.

배경 기술

[0002] 심층 신경망을 이용한 행렬-벡터 곱 연산은 심층 신경망 연산에서 핵심 작업이다. 상기 연산과 관련하여, RCA(ReRAM crossbar arrays) 회로는 행렬-벡터 곱 연산을 빠르고 효율적으로 구현할 수 있는 회로로서, 심층 신경망 가속기를 위한 하드웨어로서 사용되고 있다.

[0003] 구체적으로, RCA 회로는 아날로그 소자를 2차원의 격자 형태로 배열한 구조를 포함할 수 있다. 예를 들어, 심층 신경망의 가중치 행렬을 RCA 회로의 각 저항성 소자의 컨덕턴스 값으로 매핑하고, 심층 신경망의 입력 벡터의 값을 RCA 회로에 전압으로 인가하면, 행렬-벡터 곱의 결과는 전류 값의 형태를 가지는 출력 값으로 나타낼 수 있다.

[0004] 다만, RCA 회로는 아날로그 소자의 비이상성(nonideality)으로 인해 연산 결과가 부정확한 결과를 도출하는 문제점을 가진다. 예를 들어, RCA 회로의 아날로그 소자들은 선 저항(wire resistance) 등의 영향으로, 아날로그 소자에 인가되는 전압이 이상적인 기대 값 보다 낮아지는 전압 강하(IR drop) 또는 I-V 비선형성의 문제들을 가진다. 그러므로, RCA 회로의 비이상성으로 인해 도출되는 결과는 상기 심층 신경망의 성능을 떨어뜨리는 문제점이 될 수 있다.

[0005] 따라서, 심층 신경망의 학습과정에서 RCA 회로에 대한 비이상성의 영향을 미리 예측함으로써 상기 문제점을 해결하는 방법이 요구된다.

[0006] 위에서 설명한 배경기술은 발명자가 본원의 개시 내용을 도출하는 과정에서 보유하거나 습득한 것으로서, 반드시 본 출원 전에 일반 공중에 공개된 공지기술이라고 할 수는 없다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0007] 일 실시예에 따른 전자 장치는, 컴퓨터로 실행 가능한 명령어들(computer-executable instructions)이 저장된 메모리; 및 상기 메모리에 액세스(access)하여 상기 명령어들을 실행하는 프로세서를 포함하고, 상기 명령어들은, 제1 트레이닝 입력 벡터로부터 제1 확장 입력 데이터를 상기 심층 신경망의 제1 가중치 행렬의 차원에 기초하여 생성하고, 상기 제1 가중치 행렬 및 제1 확장 입력 데이터의 결합에 기초하여 생성된 텐서 데이터를 가중치 예측 모델에 적용함으로써, 상기 제1 가중치 행렬이 RCA(ReRAM crossbar arrays) 회로의 아날로그 비이상성(analog non-ideality)에 따라 변화된 대상 가중치 행렬을 획득하고, 상기 제1 트레이닝 입력 벡터 및 상기 대상 가중치 행렬에 기초한 순전파(forward propagation)를 포함하는 연산을 통해 계산된 상기 심층 신경망의 손실을 이용하여 역전파(back propagation)시키면서 상기 심층 신경망을 트레이닝시킬 수 있다.

[0008] 상기 프로세서는, 상기 제1 트레이닝 입력 벡터의 차원의 개수만큼, 상기 제1 트레이닝 입력 벡터와 동일한 복제 벡터를 생성하고, 상기 복수의 복제 벡터들을, 상기 제1 가중치 행렬의 차원에 기초하여, 미리 결정된(predetermined) 방향으로 결합하여 상기 제1 확장 입력 데이터를 생성할 수 있다.

[0009] 상기 프로세서는, 제2 트레이닝 입력 벡터 및 트레이닝 가중치 행렬을, 상기 RCA 회로 기반의 심층 신경망 가속기를 구현한 하드웨어 시뮬레이션에 적용함으로써, 상기 RCA 회로의 아날로그 비이상성에 따라 변화된 가중치를 나타내는 실수 값 가중치 행렬을 획득하고, 상기 제2 트레이닝 입력 벡터 및 상기 트레이닝 가중치 행렬로부터 상기 실수 값 가중치 행렬을 출력하도록, 상기 가중치 예측 모델을 트레이닝시킬 수 있다.

[0010] 상기 프로세서는, 상기 제1 트레이닝 입력 벡터의 원소 데이터가 이진(binary) 데이터를 포함하는 경우에 기초하여, 적어도 2개 이상의 트레이닝 입력 벡터들을 포함하는 트레이닝 입력 벡터 집합에서 트레이닝 입력 벡터들의 평균 데이터를 획득하고, 상기 획득된 평균 데이터에 기초하여, 상기 트레이닝 입력 벡터 집합의 분포와 동

일한 분포를 따르는 원소 데이터들을 포함하는 무작위(random) 입력 벡터를 생성할 수 있다.

[0011] 상기 프로세서는, 상기 RCA 회로의 아날로그 비이상성으로 인해 유발되는 오차가 보상되는 방법으로 학습된 심층 신경망의 가중치 행렬을 추출하고, 상기 추출된 가중치 행렬을 상기 RCA 회로의 저항성 소자로 매핑(mapping)하고, 상기 심층 신경망에 적용되는 입력 벡터를 상기 RCA 회로의 입력 비트 라인으로 매핑하고, 상기 RCA 회로에 의해, 상기 입력 벡터 및 상기 학습된 가중치 행렬의 행렬-벡터 곱 연산에 의해 출력 벡터를 획득할 수 있다.

발명의 효과

도면의 간단한 설명

[0012] 도 1은 일 실시예에 따른 RCA(ReRAM crossbar arrays) 회로의 구조를 도시한 도면이다.

도 2는 일 실시예에 따른 행렬-벡터 곱 연산의 예시를 나타내는 도면이다.

도 3은 일 실시예에 따른 RCA 회로의 아날로그 비이상성으로 인해 왜곡된 전류 값을 나타내는 도면이다.

도 4는 일 실시예에 따른 가중치 예측 모델을 활용하여 심층 신경망을 학습하는 방법을 도시한 흐름도이다.

도 5는 일 실시예에 따른 가중치 예측 모델을 통해 획득된 가중치 행렬에 기초한 심층 신경망 학습 단계를 도시한 흐름도이다.

도 6a 및 도 6b는 일 실시예에 따른 가중치 예측 모델에서 대상 가중치 행렬을 획득하는 방법을 도시한 도면이다.

도 7은 일 실시예에 따른 RCA 회로 및 가중치 예측 모델의 관계를 도시한 도면이다.

도 8a 및 도 8b는 일 실시예에 따른 가중치 예측 모델의 학습 방법을 도시한 도면이다.

도 9는 일 실시예에 따른 가중치 예측 모델을 심층 신경망에 적용하여 심층 신경망을 학습하는 방법을 도시한 도면이다.

도 10은 일 실시예에 따른 RCA 단위 블록 별로 대상 가중치 행렬을 적용하여 심층 신경망을 학습하는 방법을 도시한 도면이다.

도 11a 내지 도 11c는 일 실시예에 따른 학습된 심층 신경망의 가중치 행렬을 RCA 회로에 적용하는 방법을 도시한 도면이다.

도 12는 일 실시예에 따른 RCA 회로 아날로그 소자의 비이상성 특성을 하드웨어 시뮬레이션에 의해 결과를 도시한 도면이다.

도 13은 일 실시예에 따른 IV-IR시나리오의 16가지 예측 결과를 비교한 그래프를 도시한 도면이다.

도 14는 일 실시예에 따른 심층 신경망의 트레이닝 시간을 비교한 그래프를 도시한 도면이다.

도 15는 일 실시예에 따른 학습된 심층 신경망의 가중치 행렬을 RCA 회로에 적용했을 때 정확도(accuracy)를 측정한 결과를 도시한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0013] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예에만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0014] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0015] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

- [0016] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 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.
- [0017] 본 문서에서, "A 또는 B", "A 및 B 중 적어도 하나", "A 또는 B 중 적어도 하나", "A, B 또는 C", "A, B 및 C 중 적어도 하나", 및 "A, B, 또는 C 중 적어도 하나"와 같은 문구들 각각은 그 문구들 중 해당하는 문구에 함께 나열된 항목들 중 어느 하나, 또는 그들의 모든 가능한 조합을 포함할 수 있다.
- [0018] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 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.
- [0019] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.
- [0020] 도 1은 일 실시예에 따른 RCA(ReRAM crossbar arrays) 회로의 구조를 도시한 도면이다.
- [0021] 일 실시예에 따른 RCA 회로(100)는 심층 신경망 가속기를 위한 하드웨어로서 저항성 랜덤 액세스 메모리 크로스바 어레이 구조일 수 있다. RCA 회로(100)는 입력 라인(110), 출력 라인(120), 및 아날로그 소자(130)를 포함할 수 있다. 예를 들어, RCA 회로(100)는 하나의 저항성 소자 어레이 셀(cell)을 포함할 수 있다. 입력 라인(110)은 RCA 회로(100)에서 제1 통로(115)를 통해 아날로그 소자(130)로 입력 데이터를 전달하는 라인(예: 비트 라인(bit line))을 나타낼 수 있다. 출력 라인(120)은 RCA 회로(100)에서 제2 통로(135)를 통해 아날로그 소자(130)로부터 전달되는 출력 데이터를 출력하는 라인을 나타낼 수 있다. 아날로그 소자(130)는 심층 신경망 연산을 위한 가중치에 대응하는 저항을 가지는 소자로서, 비휘발성(non-volatile) 랜덤 액세스 메모리로 구현될 수 있다. 예를 들어, 아날로그 소자(130)는 유전체의 고체-상태 재료의 저항을 변경함으로써 작동하는 멤리스터(memristor)인 저항성 소자일 수 있다.
- [0022] 심층 신경망(deep neural network) 연산은, 심층 신경망의 임의의 레이어에서 수신되는 입력 값들에 가중치들(예: 연결 가중치들)을 개별적으로 곱하는 연산 및 곱들을 합산하는 연산을 포함할 수 있다. 전술한 곱셈 누적 연산(multiplication and accumulation, MAC)은 심층 신경망 연산에서 상당한 부분을 차지하며, 행렬-벡터 연산이라고도 나타낼 수 있다. 일 실시예에 따른 RCA 회로(100)는 행렬-벡터 연산을 수행할 수 있다. 예를 들어, 행렬-벡터 연산은 행렬을 구성하는 열 벡터에 벡터의 각 성분을 선형조합(linear combination)하는 연산을 나타낼 수 있다.
- [0023] 일 실시예에 따른 행렬-벡터 곱 연산은 수학식 1에 의해 표현될 수 있다:

수학식 1

$$Y_k = \sum_{i=1}^n W_{ik} * X_i$$

[0024]

[0025] 여기서, X_i 은 입력 벡터의 i번째 원소 데이터를 나타낼 수 있고, W_{ik} 는 행렬의 i번째 행(row) 및 k번째 열(column)의 원소 데이터를 나타낼 수 있으며, Y_k 는 출력 벡터의 k번째 원소 데이터를 나타낼 수 있다.

[0026] RCA 회로(100)에 의해 수행되는 행렬-벡터 연산에서, 행렬은 심층 신경망의 노드들을 연결하는 연결선의 가중치 값들을 가지고, 벡터는 입력 값을 가질 수 있다. RCA 회로(100)는 입력 라인(110)을 통해 입력 값에 대응하는

신호(예: 입력 전압(voltage))을 수신할 수 있다. RCA 회로(100)의 아날로그 소자(130)는 가중치 값에 대응하는 소자 값(예: 저항 값 또는 컨덕턴스(conductance) 값)을 가질 수 있다. RCA 회로(100)는 출력 라인(120)을 통해 전술한 행렬-벡터 연산의 결과에 대응하는 신호(예: 출력 전류(current))를 제공할 수 있다.

[0027] 다만, RCA 회로(100)의 아날로그 소자(130)는 아날로그 특성(예: 선 저항(wire resistance) 등)의 영향으로, 아날로그 소자(130)에 인가되는 값이 이상적인 기대 값과 달라질 수 있다. RCA 회로(100)의 아날로그 소자(130)는 아날로그 특성에 따른 전압 강하(IR drop)로 인해, 비이상성(nonideality)을 가질 수 있다. 아날로그 소자(130)의 비이상성으로 인해, 아날로그 소자(130)는 의도된 가중치에 대응하는 컨덕턴스와 다른 컨덕턴스를 나타낼 수 있다. 따라서, 아날로그 소자(130)에서 가중치 왜곡이 발생하고, 행렬-벡터 연산의 결과에서도 왜곡이 발생할 수 있다.

[0028] 도 2는 일 실시예에 따른 행렬-벡터 곱 연산의 예시를 나타내는 도면이다.

[0029] 일 실시예에 따른 전자 장치는 입력 벡터(210) 및 입력 행렬(230)의 곱 연산을 통해 출력 벡터(220)를 획득할 수 있다. 예를 들어, 전자 장치는 입력 벡터(210)의 첫번째 원소와 입력 행렬(230)의 첫번째 열에 있는 모든 원소들을 곱하여 더한 값을 출력 벡터(220)의 첫번째 원소로서 출력 벡터(220)를 획득할 수 있다.

[0030] 일 실시예에 따른 전자 장치는 RCA 회로(예: 도 1의 RCA 회로(100))에서 상술한 행렬-벡터 곱 연산을 수행할 수 있다. 예를 들어, 입력 벡터(210)는 RCA 회로의 입력 라인(예: 도 1의 입력 라인(110))에 입력 데이터로 인가되는 벡터를 나타낼 수 있다. 입력 행렬(230)은 RCA 회로의 아날로그 소자(예: 도 1의 아날로그 소자(130))의 저항성 소자에 매핑(mapping)되는 행렬을 나타낼 수 있다. 출력 벡터(220)는 RCA 회로의 출력 라인(예: 도 1의 출력 라인(120))의 출력 데이터인 벡터를 나타낼 수 있다.

[0031] 도 3은 일 실시예에 따른 RCA 회로의 아날로그 비이상성으로 인해 왜곡된 전류 값을 나타내는 도면이다.

[0032] 일 실시예에 따른 기대 전류(300)는 수학식 2에 의해 표현될 수 있다:

수학식 2

$$I_k = \sum_{i=1}^n G_{ik} * V_i$$

[0033]

[0034] 여기서, V_i 는 전압 값을 포함하는 전압 벡터에서 i 번째 전압 값을 나타낼 수 있고, G_{ik} 는 컨덕턴스 값을 포함하는 컨덕턴스 행렬에서 (i,k) 의 컨덕턴스 값을 나타낼 수 있다. 따라서, 전류는 전압 및 컨덕턴스 곱의 결과를 나타내므로, I_k 는 전류 값을 포함하는 전류 벡터에서 k 번째 전류 값을 나타낼 수 있다.

[0035] 일 실시예에 따른 전자 장치는 RCA 회로(예: 도 1의 RCA 회로(100))에서 아날로그 소자의 비이상성으로 인해 왜곡된 전류(310)를 획득할 수 있다. 왜곡된 전류(310)는 수학식 2에 의해 획득된 기대 전류(300) 값과 다른 전류를 나타낼 수 있다. 예를 들어, RCA 회로는, 인가된 전압 및 컨덕턴스에 비해, 기대 전류(300)와 다른 왜곡된 전류(310)를 출력할 수 있다. 구체적으로, 전자 장치는 RCA 회로에서 컨덕턴스 행렬이 매핑되는 아날로그 소자(예: 도 1의 아날로그 소자(130))는 아날로그 특성에 따른 전압 강하로 인해 의도된 가중치에 대응하는 컨덕턴스와 다른 컨덕턴스를 가지는 특성으로 인해, 수학식 2에서 예측된 기대 전류(300)와 다른 전류인 왜곡된 전류(310)를 획득할 수 있다. 이하, RCA 회로의 아날로그 소자에서 비이상성으로 인해 발생된 가중치 왜곡을 해결하기 위한 방법을 후술한다.

[0036] 도 4는 일 실시예에 따른 가중치 예측 모델을 활용하여 심층 신경망을 학습하는 방법을 도시한 흐름도이다.

[0037] 단계(410)에서, 전자 장치는 제1 가중치 행렬 및 제1 확장 입력 데이터를 결합하여 제1 텐서(tensor) 데이터를 생성할 수 있다. 예를 들어, 제1 가중치 행렬은 심층 신경망의 복수의 레이어 중 하나의 레이어의 가중치 행렬을 나타낼 수 있다. 제1 확장 입력 데이터는 제1 트레이닝 입력 벡터와 동일한 복수의 벡터들을 포함하는 행렬을 나타낼 수 있다. 여기서, 제1 트레이닝 입력 벡터는 심층 신경망의 학습을 위한 트레이닝 데이터를 나타낼 수 있다. 텐서 데이터는, 제1 가중치 행렬 및 제1 확장 입력 데이터를 미리 결정된 차원(dimension)의 열(column)으로 결합한 데이터를 나타낼 수 있다. 텐서 데이터를 생성하는 방법은 하기 도 6b에서 후술한다.

[0038] 일 실시예에 따른 전자 장치는, 제1 트레이닝 입력 벡터의 차원 및 제1 가중치 행렬의 차원에 기초하여, 제1 확장 입력 데이터를 생성할 수 있다. 예를 들어, 제1 트레이닝 입력 벡터의 차원이 N차원인 경우와 제1 가중치 행렬의 차원이 $N \times N$ 차원인 경우에, 전자 장치는 복제 벡터를 N개 생성할 수 있다. 복제 벡터는 제1 트레이닝 입력 벡터와 동일한 벡터를 나타낼 수 있다. 이후, 전자 장치는 복수의 복제 벡터를 제1 가중치 행렬의 차원과 동일한 $N \times N$ 차원으로, N개의 복제 벡터를 미리 결정된(predetermined) 방향으로 결합하여 제1 확장 입력 데이터를 생성할 수 있다. 즉, 제1 확장 입력 데이터는 제1 트레이닝 입력 벡터를 복제한 복수의 복제 벡터를 결합한 행렬을 나타낼 수 있다. 결과적으로, 전자 장치는 복수의 복제 벡터를 가중치 행렬의 차원을 가지는 행렬의 형태를 가지는 제1 확장 입력 데이터를 생성할 수 있다. 전자 장치는, 제1 확장 입력 데이터를 생성함으로써, 제1 가중치 행렬의 열(column) 차원의 가중치 데이터에 동일한 제1 트레이닝 입력 벡터의 데이터를 보존할 수 있다. 제1 확장 입력 데이터를 생성하는 방법은 하기 도 6b에서 후술한다.

[0039] 단계(420)에서, 전자 장치는 텐서 데이터를 가중치 예측 모델에 적용하여 대상 가중치 행렬을 획득할 수 있다. 예를 들어, 가중치 예측 모델은 RCA 회로(예: 도 1의 RCA 회로(100))의 비이상성으로 인해 왜곡된 가중치를 모델링(modeling)한 신경망을 나타낼 수 있다. 또한, 가중치 예측 모델은 수학적 3에 의해 표현될 수 있다:

수학적 3

$$g(W, x) = W_e$$

[0040]

[0041] 여기서, W 는 심층 신경망의 가중치 행렬을 나타낼 수 있고, x 는 트레이닝 입력 벡터를 나타낼 수 있고, W_e 는 대상 가중치 행렬을 나타낼 수 있다. 또한, $g(W, x)$ 는 가중치 예측 모델을 나타낼 수 있다. 예를 들어, $g(W, x)$ 는 가중치 행렬 및 트레이닝 입력 벡터의 결합을 통해 획득된 텐서 데이터를 입력 데이터로서, 대상 가중치 행렬을 출력하는 가중치 예측 모델을 나타내는 함수일 수 있다.

[0042] 전술한 바와 같이, 심층 신경망의 가중치 행렬의 가중치 값들에 대응하는 소자 값(예: 컨덕턴스 값 또는 저항 값)이 설정된 RCA 회로에 입력 벡터에 대응하는 입력 신호(예: 입력 전압)를 인가하는 경우, RCA 회로의 아날로그 소자들에서 의도된 저항 값과 다른 저항 값이 관측될 수 있다. 다시 말해, RCA 회로에서 아날로그 특성으로 인해 가중치 왜곡이 발생할 수 있다. 가중치 예측 모델은 텐서 데이터를 입력 받아, RCA 회로의 비이상성에 따라 변화된 또는 왜곡된 소자 값들에 대응하는 대상 가중치 행렬을 출력할 수 있다. 대상 가중치 행렬은 RCA 회로의 아날로그 소자(예: 도 1의 아날로그 소자(130))의 아날로그 특성으로 인해 왜곡된 가중치 값들을 포함하는 행렬을 나타낼 수 있다.

[0043] 단계(430)에서, 전자 장치는 대상 가중치 행렬 및 제1 트레이닝 입력 벡터를 심층 신경망에 적용할 수 있다.

[0044] 일 실시예에 따른 심층 신경망은 기계 학습을 통해 생성될 수 있다. 심층 신경망은, 복수의 인공 신경망 레이어들을 포함할 수 있다. 참고로, 심층 신경망은 트레이닝 입력 및 해당 트레이닝 입력에 매핑된 트레이닝 타겟(target)의 쌍을 포함하는 트레이닝 데이터에 기초하여 트레이닝 될 수 있다. 트레이닝 중의 심층 신경망은 트레이닝 입력에 응답하여 임시 출력을 생성할 수 있고, 임시 출력 및 트레이닝 타겟간의 손실이 최소화되도록 트레이닝 될 수 있다. 트레이닝 과정 동안 심층 신경망의 파라미터(예를 들어, 뉴럴 네트워크에서 노드들/레이어들 간의 연결 가중치)가 손실에 따라 업데이트될 수 있다. 이러한 학습은, 예를 들어, 기계학습 모델이 수행되는 전자 장치 자체에서 수행될 수 있고, 별도의 서버를 통해 수행될 수도 있다. 학습이 완료된 기계 학습 모델은 메모리에 저장될 수 있다.

[0045] 일 실시예에 따른 전자 장치는, 수학적 4와 같이 대상 가중치 행렬 및 제1 트레이닝 입력 벡터를 심층 신경망에 적용할 수 있다:

수학식 4

$$y = f(W, x) = g(W, x)x$$

[0046]

[0047]

여기서, W 는 심층 신경망의 가중치 행렬을 나타낼 수 있고, x 는 제1 트레이닝 입력 벡터를 나타낼 수 있다. 또한, $f(W, x)$ 는 심층 신경망의 하나의 레이어에서 가중치 행렬 및 트레이닝 입력 벡터의 연산을 나타낼 수 있다. 구체적으로, 전자 장치는 가중치 행렬 대신, $g(W, x)$ 의 연산을 수행하는 가중치 예측 모델을 통해 획득된 대상 가중치 행렬을 심층 신경망에 적용할 수 있다.

[0048]

예를 들어, 전자 장치는 트레이닝 입력인 제1 트레이닝 입력 벡터 및 대상 가중치 행렬에 기초한 임시 출력을 생성할 수 있다. 여기서, 임시 출력은 입력 벡터 및 대상 가중치 행렬의 행렬-벡터 연산의 결과일 수 있다. 전자 장치는, 생성된 임시 출력 및 트레이닝 타겟 간의 손실이 최소화되도록, 심층 신경망을 트레이닝시킬 수 있다. 여기서, 트레이닝 타겟은 제1 트레이닝 입력 벡터에 대응하는 출력일 수 있다. 전자 장치는, 트레이닝 과정 동안 심층 신경망의 복수의 레이어(예를 들어, 제1 가중치 행렬을 포함하는 레이어)들의 파라미터(예를 들어, 가중치 행렬)를 획득된 손실에 따라 업데이트를 시킬 수 있다. 전자 장치는, 제1 가중치 행렬에 기초하여 임시 출력을 생성하는 대신 대상 가중치 행렬에 기초하여 임시 출력을 생성함으로써, RCA 회로의 비이상성에 따른 왜곡이 보상된 가중치 값들을 가질 수 있도록 심층 신경망을 트레이닝시킬 수 있다.

[0049]

도 5는 일 실시예에 따른 가중치 예측 모델을 통해 획득된 가중치 행렬에 기초한 심층 신경망 학습 단계를 도시한 흐름도이다.

[0050]

일 실시예에 따른 전자 장치는 가중치 예측 모델에서 획득된 대상 가중치 행렬을, 심층 신경망의 트레이닝에 적용함으로써 심층 신경망을 트레이닝할 수 있다. 예를 들어, 심층 신경망은 복수의 레이어들을 포함할 수 있고, 각 레이어마다 가중치 행렬을 포함할 수 있다.

[0051]

단계(530)에서, 전자 장치는 심층 신경망의 복수의 레이어 중 하나의 레이어의 출력 벡터(예: 임시 출력)를 획득할 수 있다. 예를 들어, 단계(532)에서, 전자 장치는 제1 트레이닝 입력 벡터 및 제1 가중치 행렬을 가중치 예측 모델에 적용하여, 제1 가중치 행렬을 포함하는 레이어의 대상 가중치 행렬을 획득할 수 있다. 참고로, 제1 트레이닝 입력 벡터는 심층 신경망의 트레이닝을 위한 트레이닝 입력 벡터 또는 이전 레이어의 출력 벡터를 포함할 수 있다. 단계(534)에서, 전자 장치는 제1 트레이닝 입력 벡터 및 대상 가중치 행렬을 행렬에 기초한 순전파(forward propagation)를 포함하는 연산을 통해 출력 벡터를 획득할 수 있다. 달리 말해, 전자 장치는 제1 가중치 행렬 대신 가중치 예측 모델로부터 획득된 대상 가중치 행렬에 기초한 순전파를 포함하는 연산을 통해 심층 신경망을 트레이닝시킬 수 있다. 전술한 바에 따라 트레이닝된 심층 신경망은 RCA 회로의 비이상성에 따른 왜곡이 보상된 가중치 값들을 가질 수 있다. 따라서 심층 신경망의 가중치 값들에 대응하는 소자 값이 설정된 RCA 회로는, 동작 동안 의도된 소자 값을 가지고 행렬-벡터 연산을 수행할 수 있다.

[0052]

이후, 단계(540)에서, 전자 장치는 전술한 연산의 레이어가 마지막 레이어가 아닌 경우, 다음 레이어의 제1 트레이닝 입력 벡터를 획득된 출력 벡터로 적용함으로써, 다음 레이어의 순전파를 포함하는 연산을 통해 출력 벡터를 획득할 수 있다. 또한, 전자 장치는 전술한 연산의 레이어가 마지막 레이어인 경우, 역전파를 통해 심층 신경망을 트레이닝시킬 수 있다.

[0053]

단계(550)에서, 전자 장치는, 출력 벡터와 트레이닝 타겟간의 차이에 기초하여 심층 신경망을 트레이닝시킬 수 있다. 예를 들어, 전자 장치는 출력 벡터 및 트레이닝 타겟의 차이를 통해 심층 신경망의 손실을 계산할 수 있다. 이후, 전자 장치는 계산된 손실을 이용하여 역전파 시키면서 심층 신경망을 트레이닝시킬 수 있다.

[0054]

단계(560)에서, 전자 장치는 학습 횟수(예: epoch)가 미리 설정된 학습 횟수를 초과하는지 여부를 판단하여 학습을 종료할 수 있다.

[0055]

도 6a 및 도 6b는 일 실시예에 따른 가중치 예측 모델에서 대상 가중치 행렬을 획득하는 방법을 도시한 도면이다.

[0056]

일 실시예에 따른 전자 장치는 가중치 행렬(620a)을 가중치 예측 모델(640a)에 적용하여 대상 가중치 행렬

(650a)을 획득할 수 있다. 전자 장치는, 획득된 대상 가중치 행렬(650a) 및 트레이닝 벡터 x 를 수학적식(660a)에 의해 출력 벡터 y 를 획득할 수 있다.

[0057] 일 실시예에 따른 전자 장치는 텐서 데이터(630)를 가중치 예측 모델(640b)에 적용하여 대상 가중치 행렬(650b)을 획득할 수 있다. 전자 장치는 확장 입력 데이터(610) 및 가중치 행렬(620b)을 결합하여 텐서 데이터(630)를 획득할 수 있다. 도 6b에서는 설명의 편의를 위하여 텐서 데이터의 첫번째 행렬을 가중치 행렬(620b)로 도시하였으나, 이로 한정하는 것은 아니고, 전자 장치는 텐서 데이터(630)의 첫번째 행렬이 확장 입력 데이터(610)인 텐서 데이터(630)를 획득할 수 있다. 이후, 전자 장치는 텐서 데이터(630)를 가중치 예측 모델(640b)에 적용하여 대상 가중치 행렬(650b)을 획득할 수 있다.

[0058] 일 실시예에 따른 전자 장치는 트레이닝 입력 벡터(605) 및 대상 가중치 행렬(650b)을 RCA 회로에 적용할 수 있다. 여기서, 전자 장치는 트레이닝 입력 벡터(605)를 RCA 회로의 입력 라인(660b)에 적용할 수 있고, 대상 가중치 행렬(650b)을 RCA 회로의 아날로그 소자(670b)에 적용할 수 있다.

[0059] 다만, 도 6b에서는 설명의 편의를 위해 대상 가중치 행렬(650b)이 RCA 회로에 적용되는 것을 도시하였으나 이로 한정하는 것은 아니다. 전자 장치는, 대상 가중치 행렬을 적용하여 학습된 심층 신경망의 가중치 행렬을 RCA 회로에 적용할 수 있다. 학습된 심층 신경망의 가중치 행렬을 RCA 회로에 적용하는 방법은 하기 도 11c에서 후술한다.

[0060] 도 7은 일 실시예에 따른 RCA 회로 및 가중치 예측 모델의 관계를 도시한 도면이다.

[0061] 일 실시예에 따른 전자 장치는 가중치 예측 모델(700) 및 심층 신경망(730)에 기초하여 RCA 회로에서 최종 출력 전류를 획득할 수 있다. 예를 들어, 심층 신경망(730)은 가중치 예측 모델(700)에서 출력한 대상 가중치 행렬(720)에 기초하여 트레이닝된 신경망을 나타낼 수 있다. 가중치 예측 모델(700)은 텐서 데이터(710)를 입력 데이터로 하여, 대상 가중치 행렬(720)을 출력하는 신경망을 나타낼 수 있다. 전자 장치는, 트레이닝된 심층 신경망(730)의 가중치를 RCA 회로의 아날로그 소자에 매핑하여 컨덕턴스(750)를 RCA 회로에 제공할 수 있다. 구체적으로, 전자 장치는 대상 가중치 행렬(720)을 통해 RCA 회로의 비이상성에 따른 왜곡이 보상된 가중치 값들을 가질 수 있도록 트레이닝된 심층 신경망(730)의 가중치를 RCA 회로에 제공할 수 있다. 전자 장치는, 심층 신경망의 입력 벡터를 RCA 회로의 입력 라인을 통해 전압(740)을 인가할 수 있다. 결과적으로, 전자 장치는 심층 신경망의 연산 중 행렬-벡터 곱 연산을 RCA 회로의 비이상성에 따른 왜곡이 보상된 가중치를 가지는 RCA 회로에서 수행할 수 있다. 트레이닝된 심층 신경망(730)을 RCA 회로에 적용하여 최종 출력 전류에 대응하는 결과를 획득하는 방법은 하기 도 11c에서 후술한다.

[0062] 도 8a 및 도 8b는 일 실시예에 따른 가중치 예측 모델의 학습 방법을 도시한 도면이다.

[0063] 일 실시예에 따른 전자 장치는 심층 신경망의 트레이닝에 사용되는 가중치 예측 모델(840a)을 트레이닝시킬 수 있다. 가중치 예측 모델(840a)은, 복수의 인공 신경망 레이어들을 포함할 수 있다. 참고로, 가중치 예측 모델(840a)은 트레이닝 입력 및 해당 트레이닝 입력에 매핑된 트레이닝 타깃의 쌍을 포함하는 트레이닝 데이터에 기초하여 트레이닝 될 수 있다. 여기서, 트레이닝 입력은 트레이닝 입력 벡터(800a) 및 트레이닝 가중치 행렬(810a)을 포함할 수 있다. 트레이닝 입력 벡터(800a)의 원소 및 트레이닝 가중치 행렬(810a)의 원소들은 무작위 이진(random binary) 값들을 나타낼 수 있다. 트레이닝 타깃은 실수 값 가중치 행렬(830a)을 포함할 수 있다. 실수 값 가중치 행렬(830a)은 트레이닝 입력이 RCA 회로에 적용되었을 때, RCA 회로의 비이상성으로 인해 RCA 회로에서 발생된 가중치 행렬을 나타낼 수 있다. 트레이닝 중의 가중치 예측 모델(840a)은 트레이닝 입력에 응답하여 임시 출력을 생성할 수 있고, 임시 출력 및 트레이닝 타깃 간의 손실(860a)이 최소화되도록 트레이닝 될 수 있다. 여기서, 임시 출력은 임시 가중치 행렬(850a)을 포함할 수 있다. 트레이닝 과정 동안 가중치 예측 모델(840a)의 파라미터(예를 들어, 뉴럴 네트워크에서 노드들/레이어들 간의 연결 가중치)가 손실(860a)에 따라 업데이트될 수 있다. 이러한 학습은, 예를 들어, 가중치 예측 모델(840a)이 수행되는 전자 장치 자체에서 수행될 수 있고, 별도의 서버를 통해 수행될 수도 있다. 학습이 완료된 가중치 예측 모델(840a)은 메모리에 저장될 수 있다.

[0064] 일 실시예에 따른 전자 장치는, 트레이닝 입력 벡터(800a) 및 트레이닝 가중치 행렬(810a)을 하드웨어 시뮬레이션(820a)에 적용하여 실수 값 가중치 행렬(830a)을 획득할 수 있다. 구체적으로, 전자 장치는 트레이닝 입력 벡터(800a) 및 트레이닝 가중치 행렬(810a)을 통해 생성된 텐서 데이터(예: 도 6b의 텐서 데이터(630))를 하드웨어 시뮬레이션(820a)에 적용하여 실수 값 가중치 행렬(830a)을 획득할 수 있다. 예를 들어, 하드웨어 시뮬레이션(820a)은 심층 신경망 연산을 수행하는 RCA 회로의 동작의 시뮬레이션을 나타낼 수 있다. 본 명세서에서는

설명의 편의를 위해 SPICE(simulation program with integrated circuit emphasis) 시뮬레이션을 하드웨어 시뮬레이션(820a)으로 주로 설명한다.

일 실시예에 따른 실수 값 가중치 행렬(830a)은 수학식 5에 의해 획득될 수 있다:

수학식 5

$$W_e = YD^{-1}$$

여기서, Y 는 하드웨어 시뮬레이션(820a)을 통해 획득된 최종 출력 전류 벡터를 나타낼 수 있고, 또한 D^{-1} 는 트레이닝 입력 벡터(800a)가 하드웨어 시뮬레이션(820a)에서 전압 값으로 인가된 전압 벡터로부터 생성된 대각 행렬(diagonal matrix)을 나타낼 수 있다. 다시 말해, 실수 값 가중치 행렬(830a)은 하드웨어 시뮬레이션(820a)의 최종 출력 전류 및 하드웨어 시뮬레이션(820a)의 입력 전압에 기초하여, RCA 회로의 비이상성으로 인한 가중치 왜곡을 포함할 수 있다. 또한, 전자 장치는 가중치 왜곡을 포함하는 실수 값 가중치 행렬(830a)을 기초로 가중치 예측 모델(840a)을 트레이닝 시킴으로써, 아날로그 특성으로 인해 왜곡된 가중치 값들을 포함하는 행렬을 출력하는 가중치 예측 모델(840a)을 획득할 수 있다.

일 실시예에 따른 전자 장치는, 제1 트레이닝 입력 벡터(800b)의 원소들의 평균 데이터(810b)에 기초하여 생성된 제2 확장 입력 데이터(822b)를 가중치 예측 모델(840b)에 적용하여 가중치 예측 모델(840b)을 트레이닝 시킬 수 있다. 예를 들어, 제1 트레이닝 입력 벡터(800b)는 가중치 예측 모델(840b)을 트레이닝 시키기 위한 트레이닝 입력일 수 있다. 구체적으로, 제1 트레이닝 입력 벡터(800b)는 무작위(random)로 생성된 이진(binary) 데이터들을 포함할 수 있다. 제2 확장 입력 데이터(822b)는 제2 무작위 입력 벡터(812b)와 동일한 복수의 벡터들을 포함하는 행렬을 나타낼 수 있다. 여기서, 제2 무작위 입력 벡터(812b)는 제1 트레이닝 입력 벡터(800b)의 원소들의 평균 데이터(810b)에 기초하여 무작위로 생성된 이진 데이터들을 포함하는 벡터를 나타낼 수 있다. 구체적으로, 전자 장치는 제1 트레이닝 입력 벡터(800b)의 원소 데이터가 이진 데이터인 경우에 기초하여, 이진 데이터들의 평균 데이터(810b)를 계산할 수 있다. 평균 데이터(810b)는 제1 트레이닝 입력 벡터(800b)의 모든 원소들을 더한 값을 제1 트레이닝 입력 벡터(800b)의 원소들의 개수로 나눈 값일 수 있다. 다만, 평균 데이터(810b)를 이로 한정하는 것은 아니고, 평균 데이터(810b)는 제1 트레이닝 입력 벡터(800b)의 이진 데이터들 각각의 비율을 포함할 수 있다.

일 실시예에 따른 평균 데이터는 수학식 6과 같이 계산될 수 있다:

수학식 6

$$\mathbb{E}[\hat{X}] = \sum_x \mathbb{E}[x]/100$$

예를 들어, 여기서 $\mathcal{X}$ 는 트레이닝 입력 벡터를 나타낼 수 있고, $\mathbb{E}[x]/100$ 는 100개의 트레이닝 입력 벡터들의 평균 데이터를 나타낼 수 있다. 또한, $\hat{X}$ 는 무작위 입력 벡터를 나타낼 수 있고, $\mathbb{E}[\hat{X}]$ 는 무작위 입력 벡터의 평균 데이터를 나타낼 수 있다.

이후, 전자 장치는 계산된 평균 데이터(810b)에 기초하여, 무작위로 이진 데이터를 생성하여 제1 트레이닝 입력 벡터(800b)의 원소들 개수와 동일한 개수를 가지는 벡터를 생성할 수 있다. 예를 들어, 전자 장치는 제1 트레이닝 입력 벡터(800b)의 평균 데이터(810b)와 동일한 평균 데이터를 가지는 제2 무작위 입력 벡터(812b)를 생성할 수 있다. 전자 장치는, 제2 무작위 입력 벡터(812b)를 생성함으로써, 가중치 예측 모델(840b)이 입력 데이터(예: 제1 트레이닝 입력 벡터(800b))에 의존적이지 않게 트레이닝 시킬 수 있다. 참고로, 도 8b에서는 설명의 편의를 위해, 전자 장치가 하나의 제1 트레이닝 입력 벡터(800b)를 기준으로 평균 데이터(810b)를 계산하는 것으로 설명하였으나, 이로 한정하는 것은 아니다. 다른 예를 들어, 전자 장치는 복수의 트레이닝 입력 벡터들의 평균 데이터를 계산하여 무작위 입력 벡터를 생성할 수 있다. 즉, 전자 장치는 복수의 트레이닝 입력 벡터

들의 평균 데이터와 동일한 평균 데이터를 가지는 무작위 입력 벡터를 생성할 수 있다.

[0073] 일 실시예에 따른 전자 장치는, 제1 트레이닝 입력 벡터(800b)로부터 생성된 제2 확장 입력 데이터(822b) 및 가중치 행렬(830b)을 가중치 예측 모델(840b)에 적용하여 임시 가중치 행렬(852b)을 획득할 수 있다. 전자 장치는, 임시 가중치 행렬(852b) 및 실수 값 가중치 행렬(860b)의 차이를 통해 손실(872b)을 계산할 수 있다. 이후, 전자 장치는 가중치 예측 모델(840b)이 손실(872b)을 최소화되도록 트레이닝시킬 수 있다. 전자 장치는, 트레이닝 과정 동안 가중치 예측 모델(840b)의 파라미터(예를 들어, 가중치 예측 모델(840b)에서 노드들/레이어들 간의 연결 가중치)가 손실(872b)에 따라 업데이트 시킬 수 있다. 예를 들어, 전자 장치는, 제1 트레이닝 입력 벡터(800b)로부터 생성된 제2 확장 입력 데이터(826b) 및 가중치 행렬(830b)을 가중치 예측 모델(840b)에 적용하여 임시 가중치 행렬(856b)을 획득할 수 있다. 전자 장치는, 임시 가중치 행렬(856b) 및 실수 값 가중치 행렬(860b)의 차이를 통해 손실(876b)을 계산할 수 있다. 이후, 전자 장치는 가중치 예측 모델(840b)이 손실(876b)을 최소화되도록 트레이닝시킬 수 있다. 전자 장치는, 트레이닝 과정 동안 가중치 예측 모델(840b)의 파라미터(예를 들어, 가중치 예측 모델(840b)에서 노드들/레이어들 간의 연결 가중치)가 손실(876b)에 따라 업데이트 시킬 수 있다.

[0074] 참고로, 도 8b에서는 전자 장치가 제2 무작위 입력 벡터(예를 들어, 제2 무작위 입력 벡터(812b))를 생성함으로써 가중치 예측 모델(840b)이 입력 데이터에 의존적이지 않게 트레이닝 시키는 방법을 주로 설명하였으나, 이로 한정하는 것은 아니다. 예를 들어, 전자 장치는 제2 무작위 입력 벡터 및 가중치 행렬을, 상술한 방법에 의해 트레이닝된 가중치 예측 모델에 적용하여 대상 가중치 행렬을 획득할 수 있다. 여기서, 전자 장치는 복수 개의 입력 데이터(예를 들어, 제1 트레이닝 입력 벡터(800b))에 대해 각각 대상 가중치 행렬을 획득하지 않고, 무작위 데이터(예를 들어, 제2 무작위 입력 벡터로서 복수 개의 입력 데이터에 대한 대표성을 나타낼 수 있는 입력 벡터를 의미)를 생성함으로써 대상 가중치 행렬을 획득하는 시간을 감소시킬 수 있다. 또한, 전자 장치는 무작위 데이터(예를 들어, 제2 무작위 입력 벡터)에 기초하여 대상 가중치 행렬을 획득함으로써, 입력 데이터(예를 들어, 제1 트레이닝 입력 벡터(800b))에 대한 의존성이 제거된 대상 가중치 행렬을 심층 신경망에 적용할 수 있다. 이하, 도 9에서는 가중치 예측 모델로부터 획득된 대상 가중치 행렬을 심층 신경망의 트레이닝에 적용하는 방법을 후술한다.

[0075] 도 9는 일 실시예에 따른 가중치 예측 모델을 심층 신경망에 적용하여 심층 신경망을 학습하는 방법을 도시한 도면이다.

[0076] 일 실시예에 따른 전자 장치는 가중치 예측 모델(930)을 심층 신경망(910)의 트레이닝에 적용시켜, 심층 신경망(910)을 트레이닝 시킬 수 있다. 예를 들어, 가중치 예측 모델(930)은 사전에 트레이닝된 가중치 예측 모델을 나타낼 수 있다. 전자 장치는, 트레이닝 대상인 심층 신경망(910)의 가중치 행렬(920) 및 심층 신경망(910)의 트레이닝을 위한 트레이닝 입력인 입력 벡터(900)를, 사전에 트레이닝된 가중치 예측 모델(930)에 적용하여 대상 가중치 행렬(940)을 획득할 수 있다. 이후, 전자 장치는 대상 가중치 행렬(940) 및 입력 벡터(900)를, 가중치 행렬(920)이 포함된 레이어에서 행렬-벡터 곱 연산하여 출력 벡터(950)을 획득할 수 있다. 예를 들어, 출력 벡터(950)은 대상 가중치 행렬(940)이 적용되는 레이어의 임시 출력일 수 있다. 전자 장치는, 가중치 행렬(920)이 포함된 레이어가 심층 신경망(910)의 마지막 레이어인 경우에 기초하여, 출력 벡터(950) 및 입력 벡터(900)에 대응하는 트레이닝 타겟의 차이를 나타내는 손실(960)을 계산할 수 있다. 예를 들어, 손실(960)은 가중치 행렬(920)이 아닌 대상 가중치 행렬(940)을 이용하여 획득된 손실이므로, RCA 회로의 비이상성이 반영된 손실을 나타낼 수 있다. 그러므로, 전자 장치는 심층 신경망(910)의 파라미터(예를 들어, 노드들/레이어들 간의 연결 가중치)를 전술한 손실(960)에 따라 업데이트 시키므로, RCA 회로의 비이상성이 반영되도록 심층 신경망(910)을 트레이닝 시킬 수 있다. 참고로, 도 9에서는 설명의 편의를 위해 심층 신경망(910)의 복수의 레이어들을 생략하고 하나의 레이어의 가중치 행렬(920)을 가중치 예측 모델(930)에 적용하는 예시를 주로 설명하겠으나, 이로 한정하는 것은 아니다.

[0077] 도 10은 일 실시예에 따른 RCA 단위 블록 별로 대상 가중치 행렬을 적용하여 심층 신경망을 학습하는 방법을 도시한 도면이다.

[0078] 일 실시예에 따른 전자 장치는, 심층 신경망의 트레이닝 동안, RCA 단위 블록 별로 가중치 예측 모델로부터 예측된 대상 가중치 행렬을 적용할 수 있다. 심층 신경망의 크기가 RCA의 크기보다 큰 경우, 심층 신경망의 일부분에 대응하는 가중치(예: RCA 단위 블록에 대응하는 가중치)가 RCA에 설정될 수 있기 때문이다. RCA 단위 블록은 RCA에서 설정 가능한 가중치들의 개수에 기초하여 결정되는 블록으로서, 예를 들어, RCA에 64×64 개의 가중치들이 설정 가능한 경우 RCA 단위 블록은 64×64 크기의 블록일 수 있다. 예를 들어, 심층 신경망의 가중치

들이 100×100 개인 경우, 100×100 개의 가중치들 중 RCA 단위 블록인 64×64 에 대응하는 가중치들에 대해서 전술한 가중치 예측 모델에 기초한 대상 가중치 행렬이 예측될 수 있다. 전자 장치는 심층 신경망의 순전과 과정에서 RCA 단위 블록마다 전술한 가중치 예측 모델에 기초하여 대상 가중치 행렬을 반복적으로 예측하고, 예측된 대상 가중치 행렬을 이용하여 순전과 연산을 수행할 수 있다. 본 명세서에서는 설명의 편의를 위해 RCA 단위 블록이 레이어인 예시를 주로 설명하나, 이로 한정하는 것은 아니다.

[0079] 전자 장치는, 심층 신경망의 각 레이어(layer)별로 가중치 예측 모델의 대상 가중치 행렬을 적용하여 심층 신경망을 트레이닝 시킬 수 있다. 예를 들어, 전자 장치는 심층 신경망의 하나의 레이어에 해당하는 제1 RCA 단위 블록(1010), 제2 RCA 단위 블록(1020), 및 제n RCA 단위 블록(1040)에 대응하는 심층 신경망의 레이어에 제1 가중치 예측 모델(1016), 제2 가중치 예측 모델(1026), 및 제3 가중치 예측 모델(1046)을 통해 획득된 대상 가중치 행렬들을 적용하여 심층 신경망을 트레이닝 시킬 수 있다. 참고로, 도 10에서는 가중치 예측 모델들(예: 제1 가중치 예측 모델(1016), 제2 가중치 예측 모델(1026), 및 제n 가중치 예측 모델(1046))이 별개 모델인 것처럼 도시하였으나, 이로 한정하는 것은 아니다. 한 가중치 예측 모델이 여러 RCA 단위 블록들의 연산들에서 반복적으로 사용될 수도 있다.

[0080] 일 실시예에 따른 전자 장치는 제1 트레이닝 입력(1014) 및 제1 레이어의 가중치 행렬(1012)을 제1 가중치 예측 모델(1016)에 적용하여 제1 대상 가중치 행렬(1018)을 획득할 수 있다. 여기서, 제1 트레이닝 입력(1014)은 심층 신경망의 트레이닝 입력(1000)을 나타낼 수 있다. 전자 장치는 제1 트레이닝 입력(1014) 및 제1 대상 가중치 행렬(1018)을 제1 RCA 단위 블록(1010)에 대응하는 심층 신경망의 레이어에 적용하여 제2 레이어의 트레이닝 입력을 나타내는 제2 트레이닝 입력(1024)을 계산할 수 있다.

[0081] 전자 장치는 제2 트레이닝 입력(1024) 및 제2 레이어의 가중치 행렬(1022)을 제2 가중치 예측 모델(1026)에 적용하여 제2 대상 가중치 행렬(1028)을 획득할 수 있다. 전자 장치는 제2 트레이닝 입력(1024) 및 제2 대상 가중치 행렬(1028)을 제2 RCA 단위 블록(1020)에 대응하는 심층 신경망의 레이어에 적용하여 제3 레이어의 트레이닝 입력을 나타내는 제3 트레이닝 입력(1034)을 계산할 수 있다.

[0082] 전자 장치는 심층 신경망의 레이어가 n개인 경우에 기초하여, 제n 트레이닝 입력(1044) 및 제n 레이어의 가중치 행렬(1042)을 제n 가중치 예측 모델(1046)에 적용하여 제n 대상 가중치 행렬(1048)을 획득할 수 있다. 전자 장치는 제n 트레이닝 입력(1044) 및 제n 대상 가중치 행렬(1048)을 제n RCA 단위 블록(1040)에 대응하는 심층 신경망의 레이어에 적용하여 임시 출력(1050)을 계산할 수 있다. 전자 장치는, 트레이닝 타깃(1052) 및 임시 출력(1050)의 차이를 통해 손실(1054)을 계산할 수 있다. 트레이닝 타깃(1052)은 트레이닝 입력(1000)에 매핑된 트레이닝 타깃을 나타낼 수 있다. 이후, 전자 장치는 임시 출력(1050) 및 트레이닝 타깃(1052) 간의 손실(1054)이 최소화되도록 심층 신경망을 트레이닝 될 수 있다. 트레이닝 과정 동안 심층 신경망의 파라미터가 손실(1054)에 따라 업데이트될 수 있다.

[0083] 도 11a 내지 도 11c는 일 실시예에 따른 학습된 심층 신경망의 가중치 행렬을 RCA 회로에 적용하는 방법을 도시한 도면이다.

[0084] 도 11a는 일 실시예에 따른 전자 장치가 입력 값(1110a)을 학습된 심층 신경망(1105a)에 적용하여 출력 값(1140a)을 획득하는 것을 도시한 도면이다. 구체적으로, 학습된 심층 신경망(1105a)은 학습된 가중치 행렬(1120a)을 포함할 수 있다. 출력 값(1140a)은, 입력 값(1110a)이 학습된 심층 신경망(1105a)에 적용되었을 때 매핑되는 값으로서, 트레이닝 타깃에 가까운 값을 나타낼 수 있다. 또한, 전자 장치는 학습된 심층 신경망(1105a)의 추론 과정(1100a)을 소프트웨어(software) 상에서 수행할 경우, 왜곡되지 않은 출력 값을 획득할 수 있다.

[0085] 도 11b는 일 실시예에 따른 전자 장치가 입력 값(1110b) 및 학습된 심층 신경망(1105b)의 학습된 가중치 행렬(1120b)을 RCA 회로(1130b)에 적용하여 왜곡된 출력 값(1140b)을 획득하는 것을 도시한 도면이다. 여기서, 전자 장치는 행렬-벡터 곱 연산을 수행하기 위해 RCA 회로(1130b)에 학습된 가중치 행렬(1120b)을 적용할 수 있다. 다만, 학습된 가중치 행렬(1120b)은 RCA 회로(1130b)의 비이상성으로 인해, RCA 회로(1130b)에 적용될 때 왜곡된 가중치 행렬로 적용될 수 있다. 그 결과, 전자 장치는 입력 값(1110b)이 학습된 심층 신경망(1105b)에 적용되어 매핑되는 출력 값과 다른 왜곡된 출력 값(1140b)을 획득할 수 있다. 여기서 왜곡된 출력 값(1140b)은 입력 값(1110b)에 대해 기대되는 이상적인 출력 값(예: 도 11a의 출력 값(1140a))과 다른 출력 값을 나타낼 수 있다.

[0086] 도 11c는 일 실시예에 따른 전자 장치가 입력 값(1110c) 및 가중치 예측 모델(1115c)에 의해 학습된 심층 신경

망(1105c)의 학습된 가중치 행렬(1120c)을 RCA 회로(1130c)에 적용하여 출력 값(1140c)을 획득하는 것을 도시한 도면이다.

[0087] 일 실시예에 따른 전자 장치는, 행렬-벡터 곱 연산을 수행하기 위해 RCA 회로(1130c)에, 가중치 예측 모델(1115c)이 적용되어 학습된 심층 신경망(1105c)의 학습된 가중치 행렬(1120c)을 적용할 수 있다. 예를 들어, 학습된 가중치 행렬(1120c)은 RCA 회로(1130c)의 비이상성을 반영하여 학습된 가중치를 나타내므로, RCA 회로(1130c)에 적용될 때 정상적인 가중치 행렬(예: 도 11a의 학습된 가중치 행렬(1120a))로 적용될 수 있다. 그 결과, 전자 장치는 입력 값(1110c)이 학습된 심층 신경망(1105c)에 적용되어 매핑되는 출력 값(예: 도 11a의 출력 값(1140a))에 가까운 출력 값(1140c)을 획득할 수 있다.

[0088] 도 12는 일 실시예에 따른 RCA 회로 아날로그 소자의 비이상성 특성을 하드웨어 시뮬레이션에 의해 결과를 도시한 도면이다.

[0089] 일 실시예에 따른 그래프(1200)는 아날로그 소자(예: 도 1의 아날로그 소자(130))의 선 저항 및 I-V 비선형성의 영향을 나타낼 수 있다. 예를 들어, 전자 장치는 무작위 이진 가중치 행렬 및 무작위 이진 입력 전압 신호가 인가된 RCA 단위 블록의 크기가 64×64 인 RCA 회로의 시뮬레이션 결과를 획득할 수 있다. 무작위 이진 가중치 행렬은 심층 신경망의 가중치 행렬일 수 있다. 전자 장치는, 무작위 이진 가중치 행렬을 RCA 회로에 매핑하여 무작위 이진 가중치 행렬에 대응하는 컨덕턴스 값을 하드웨어 시뮬레이션에서 획득할 수 있다. 무작위 이진 입력 전압 신호는 심층 신경망의 트레이닝 입력 벡터를 RCA 회로에 전압으로 인가하기 위해 변환된 전압 신호일 수 있다. 시뮬레이션 결과는 RCA 회로에서 무작위 이진 가중치 행렬에 대응하는 컨덕턴스 값 및 무작위 이진 입력 전압 신호에 따라 인가된 전압 값의 행렬-벡터 연산의 결과에 대응하는 출력 전류를 나타낼 수 있다. 도 12에서 도시된 그래프(1200)의 가로축은 RCA 단위 블록의 열 지수(column index)를 나타낼 수 있고, 그래프(1200)의 세로축은 출력 전류(output current)를 나타낼 수 있다. 결과적으로 그래프(1200)는, 전압, 저항, 및 RCA 단위 블록의 크기 중 적어도 하나 이상에 의해 RCA 회로의 출력 전류가 영향이 있다는 결과를 도시한 그래프를 나타낼 수 있다.

[0090] 도 13은 일 실시예에 따른 IV-IR시나리오의 16가지 예측 결과를 비교한 그래프를 도시한 도면이다.

[0091] 일 실시예에 따른 그래프(1300)는 서로 다른 IV-IR(전류-전압)의 비이상성 및 전압 강하) 시나리오의 16가지 경우에 대한 예측 방법을 비교한 결과를 나타내는 그래프일 수 있다. 예를 들어, IV-IR 시나리오는 RCA 회로의 RCA 단위 블록의 크기, 입력 전압, 및 아날로그 소자가 나타내는 저항이 각각 다른 상황에서 예측 결과를 측정할 시나리오를 나타낼 수 있다. 그래프(1300)의 세로축은 RRMSE를 나타낼 수 있다. RRMSE는 회귀 모델(regression model)의 손실 함수로서 예측 값과 실제 값의 차이인 오차들의 제곱 평균의 루트 값을 나타낼 수 있다. RRMSE는 수학식 7에 의해 표현될 수 있다:

수학식 7

$$RRMSE = \sqrt{\frac{\sum_i \|y^{(i)} - y_{\text{ref}}^{(i)}\|_2^2}{\sum_i \|y_{\text{ref}}^{(i)}\|_2^2}}$$

[0092]

[0093] 여기서 $y^{(i)}$ 는 RCA 회로에서 측정된 출력 전류를 나타낼 수 있고, $y_{\text{ref}}^{(i)}$ 는 IV-IR 시나리오에 대한 정답 전류를 나타낼 수 있다.

[0094] 일 실시예에 따른 제1 비교 실시예(1302)는 GENIEx(Generalized Approach to Emulating Non-Ideality in Memristive Xbars using Neural Networks)을 심층 신경망의 트레이닝에 적용하여 트레이닝 된 심층 신경망을 RCA 회로 상에 적용하여 출력 전류를 획득하는 실시예를 나타낼 수 있다. 예를 들어, GENIEx는 트레이닝 입력 벡터와 가중치 행렬을 플래튼(flatten)하여 생성된 플랫(flat) 벡터를 연결하여 생성된 입력 데이터로 하여 1차원의 출력 벡터를 획득할 수 있다. 전자 장치는, 트레이닝 입력 벡터 및 가중치 행렬을 곱한 결과에 출력 벡터를 곱한 값을 심층 신경망의 트레이닝에서 임시 출력으로 하여 심층 신경망을 트레이닝시킬 수 있다. 제1 비교 실시예(1302)에 기초하여 RCA 회로 상에서의 IV-IR 시나리오 결과는, RCA 회로의 단위 블록의 크기가 커질수록 측정된 출력 전류가 정답 전류와 차이가 커지는 결과를 포함할 수 있다. 또한, 제1 비교 실시예(1302)에 기

초하여 RCA 회로 상에서의 IV-IR 시나리오 결과는, RCA 회로의 아날로그 소자의 저항 값이 커질수록 측정된 출력 전류가 정답 전류와 차이가 커지는 결과를 포함할 수 있다. 제1 비교 실시예(1302)는 GENIEx에서 가중치 행렬을 그대로 사용하지 않고 1차원의 플랫 벡터로 변환하여 획득한 출력 벡터를 심층 신경망의 트레이닝에 사용하기 때문에, GENIEx에 기초하여 트레이닝 된 심층 신경망을 RCA 회로에 적용한 IV-IR 시나리오 결과는 아날로그 소자의 비이상성에 취약한 결과를 나타낼 수 있다.

[0095] 일 실시예에 따른 제2 비교 실시예(1304)는 S-SCN(stochastic Scaling convolutional network)을 심층 신경망의 트레이닝에 적용하여 트레이닝 된 심층 신경망을 RCA 회로 상에 적용하여 출력 전류를 획득하는 실시예를 나타낼 수 있다. 예를 들어, S-SCN은 가중치 행렬을 입력 데이터로 하여 2차원의 출력 행렬을 획득할 수 있다. 전자 장치는, 획득된 출력 행렬 및 트레이닝 입력벡터에 곱한 값을 심층 신경망의 트레이닝에서 임시 출력으로 하여 심층 신경망을 트레이닝시킬 수 있다. 제2 비교 실시예(1304)에 기초하여 RCA 회로 상에서의 IV-IR 시나리오 결과는, 2차원의 출력 행렬을 구할 때 심층 신경망의 가중치 행렬의 공간 정보만을 유지하므로, RCA 회로의 아날로그 소자의 저항 값이 커질수록 아날로그 소자의 비이상성에 취약한 결과를 나타낼 수 있다. 다시 말해, S-SCN은 심층 신경망의 가중치 행렬의 공간 정보를 유지하지만 트레이닝 입력 벡터를 가중치 예측과정(예를 들어, 2차원의 출력 행렬을 획득하는 과정)에 사용하지 않기 때문에, RCA 회로의 입력 전압 값이 커질수록 아날로그 소자의 비이상성에 취약한 결과를 나타낼 수 있다.

[0096] 이와 달리, 일 실시예에 따른 제1 가중치 예측 모델(1306) 및 제2 가중치 예측 모델(1308)을 적용하여 트레이닝 된 심층 신경망을 RCA 회로에 적용하여 획득한 결과는, RCA 회로의 아날로그 소자의 저항 값이 커질수록 비교 실시예들과 달리 아날로그 소자의 비이상성에 대해 우수한 결과를 나타낼 수 있다. 예를 들어, 제1 가중치 예측 모델(1306)은, S-WPM(stochastic weight-centric prediction model)으로서, 트레이닝 입력 벡터의 원소들의 평균 데이터(예: 도 8b에서 평균 데이터(810b))에 기초하여 생성된 확장 입력 데이터 및 가중치 행렬로 트레이닝 된 가중치 예측 모델일 수 있다. 제2 가중치 예측 모델(1308)은, WPM(weight-centric prediction model)으로서, 트레이닝 입력 벡터로부터 생성된 확장 입력 데이터 및 가중치 행렬로 트레이닝 된 가중치 예측 모델일 수 있다. 제1 가중치 예측 모델(1306)을 적용하여 트레이닝 된 심층 신경망을 RCA 회로에 적용하여 획득한 결과는, 평균 데이터를 활용함으로써 입력 데이터의 의존성이 감소된 모델이므로, 트레이닝 입력 벡터를 그대로 사용하는 제2 가중치 예측 모델에 비해 우수한 성능 결과를 포함할 수 있다. 또한, 제1 가중치 예측 모델(1306) 및 제2 가중치 예측 모델(1308)을 적용하여 트레이닝 된 심층 신경망을 RCA 회로에 적용하여 획득한 결과는, 비교 실시예들과 달리 측정된 출력 전류와 정답 전류의 차이가 일관되게 우수한 결과를 나타낼 수 있다. 구체적으로, 제1 가중치 예측 모델(1306)을 적용하여 트레이닝 된 심층 신경망을 RCA 회로에 적용하여 획득한 결과는, 제1 비교 실시예(1302)에 비해 최대 90배 및 제2 비교 실시예(1304)에 비해 최대 6배까지 RCA 회로의 아날로그 소자의 비이상성 문제를 해결하는 결과를 포함할 수 있다.

[0097] 도 14는 일 실시예에 따른 심층 신경망의 트레이닝 시간을 비교한 그래프를 도시한 도면이다.

[0098] 일 실시예에 따른 그래프(1400)는 심층 신경망의 트레이닝 시간을 비교한 결과를 포함하는 그래프를 나타낼 수 있다. 그래프(1400)의 세로 축은 트레이닝 시간을 나타낼 수 있고, 그래프(1400)의 가로 축은 비교 실시예 및 본 발명의 가중치 예측 모델을 포함할 수 있다.

[0099] 예를 들어, 제1 비교 실시예(1402) 및 제2 가중치 예측 모델(1408)은, 심층 신경망의 트레이닝에 적용되기 위해서, 모든 입력 쌍(예를 들어, 트레이닝 입력 벡터 및 가중치 행렬)에 대해 재평가(예를 들어, 1번의 epoch마다 새로운 출력 벡터 또는 대상 가중치 행렬을 획득)가 필요하기 때문에, 입력 데이터에 대한 종속성으로 인해 트레이닝 시간이 느린 결과를 나타낼 수 있다. 이와 달리, 제1 가중치 예측 모델(1406)은 트레이닝 입력 벡터의 원소들의 평균 데이터(예: 도 8b에서 평균 데이터(810b))에 기초하여 생성된 확장 입력 데이터 및 가중치 행렬에 기초하여 생성된 대상 가중치 행렬을 획득함으로써, 심층 신경망의 트레이닝 과정 동안 대상 가중치 행렬을 한 번만 생성하기 때문에 트레이닝 시간이 빠른 결과를 나타낼 수 있다. 구체적으로, 제1 가중치 예측 모델(1406)을 심층 신경망의 트레이닝에 적용한 결과는, 컨볼루션 레이어의 행 및 열 방향에 걸쳐 다른 입력 데이터에 대해 여러 번 재사용되는 동일한 가중치(즉, 대상 가중치 행렬)에서 추가적인 속도 향상이 발생한 결과를 포함할 수 있다.

[0100] 도 15는 일 실시예에 따른 학습된 심층 신경망의 가중치 행렬을 RCA 회로에 적용했을 때 정확도(accuracy)를 측정한 결과를 도시한 그래프이다.

[0101] 일 실시예에 따른 그래프(1500)는 가중치 예측 모델을 사용함으로써 RCA 회로의 비이상성을 인식하는 트레이닝 이 된 심층 신경망 및 단순히 트레이닝 된 심층 신경망을, 각각 RCA 회로에 적용했을 때 결과를 나타낼 수

있다. 예를 들어, 소프트웨어에서의 정확도(1502)는 트레이닝 된 심층 신경망을 소프트웨어(software) 상에서의 획득한 추론 결과에 대한 정확도를 나타낼 수 있다. 제1 RCA에서의 정확도(1504)는 트레이닝 된 심층 신경망을 RCA 회로에 적용하여, RCA 회로 상에서의 추론 결과에 대한 정확도를 나타낼 수 있다. 마지막으로, 본 발명의 실시예인 제2 RCA에서의 정확도(1506)는 가중치 예측 모델을 사용함으로써 RCA 회로의 비이상성을 인식하여 트레이닝이 된 심층 신경망을 RCA 회로에 적용하여, RCA 회로 상에서의 추론 결과에 대한 정확도를 나타낼 수 있다. 그러므로, 전술한 결과들은 가중치 예측 모델을 심층 신경망의 트레이닝에 적용했을 때, 트레이닝 된 심층 신경망의 RCA 회로 상에서 추론 성능이 우수하다는 결과를 나타낼 수 있다.

[0102] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 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 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0103] 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, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0104] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 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.

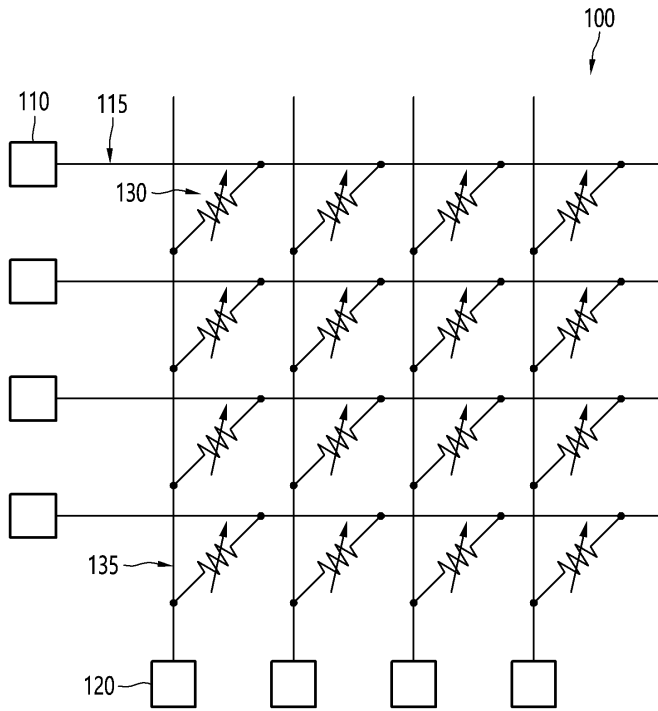
[0105] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0106] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 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.

[0107] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

도면1



도면2

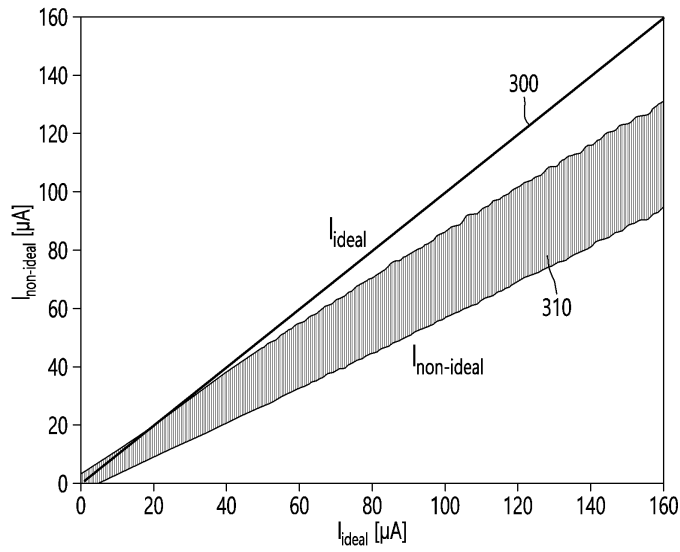
$$Y_k = \sum_{i=1}^n W_{ik} * X_i \sim 200$$

$$210 \left[\begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_n \end{array} \right]^T \left[\begin{array}{cccc} W_{11} & W_{12} & \cdots & W_{1m} \\ W_{21} & W_{22} & \cdots & W_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nm} \end{array} \right] 230$$

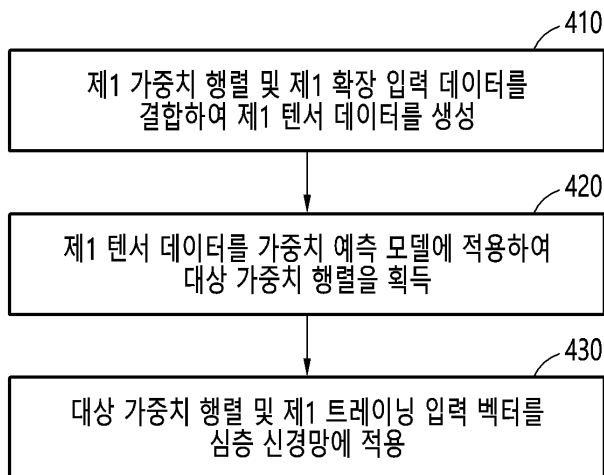
$$\parallel$$

$$\left[\begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_m \end{array} \right] 220$$

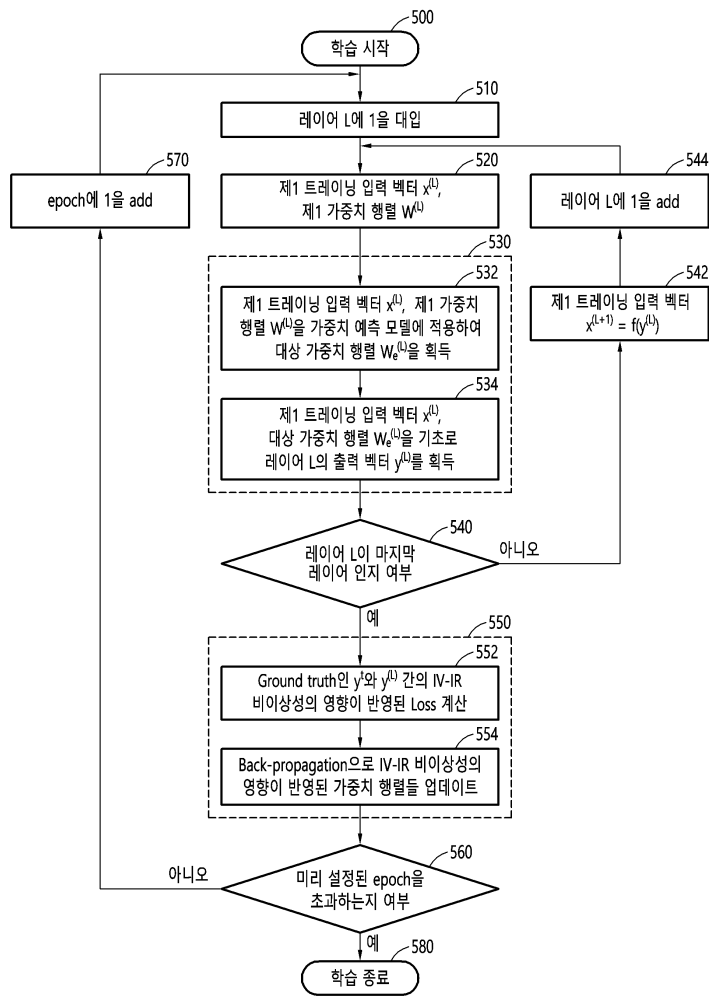
도면3



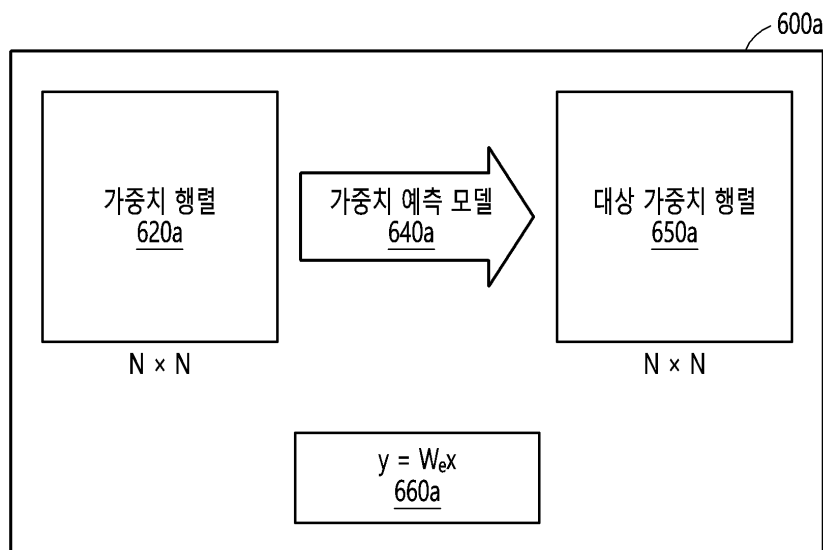
도면4



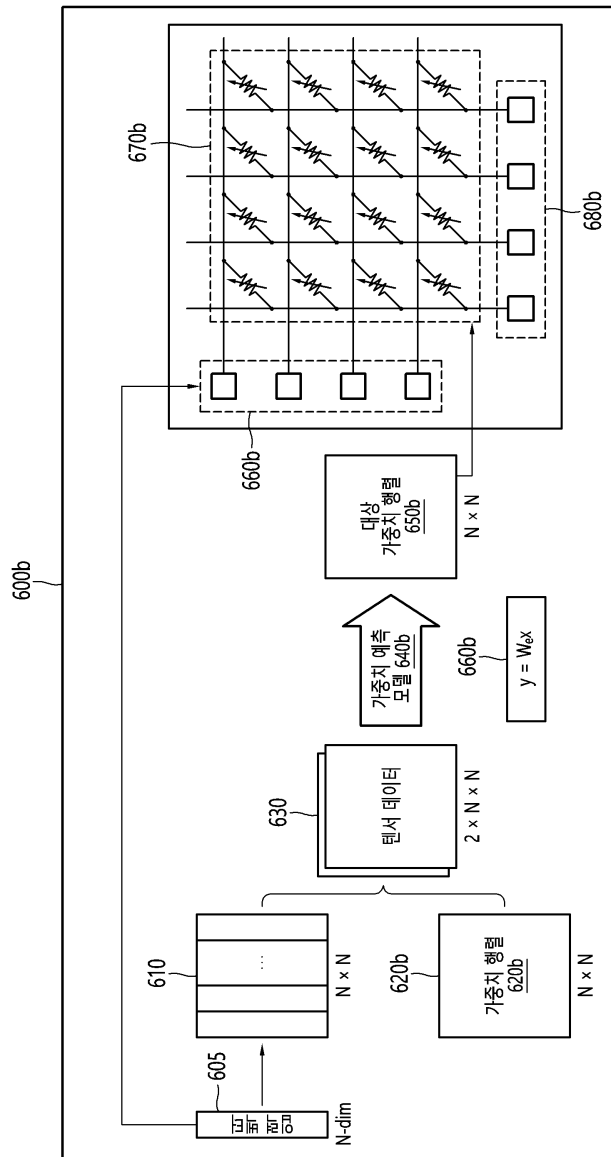
도면5



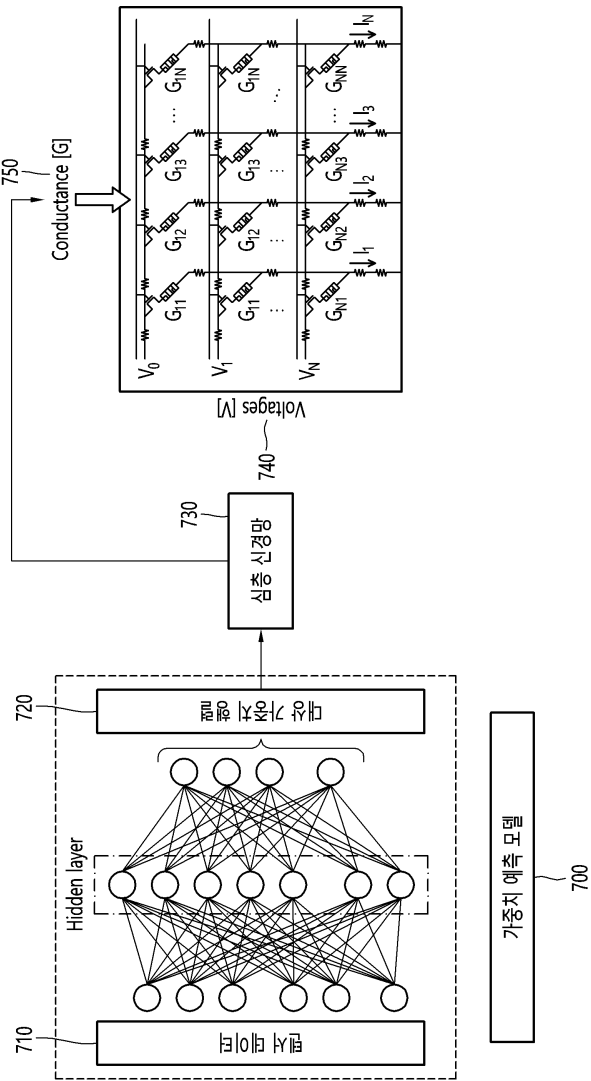
도면6a



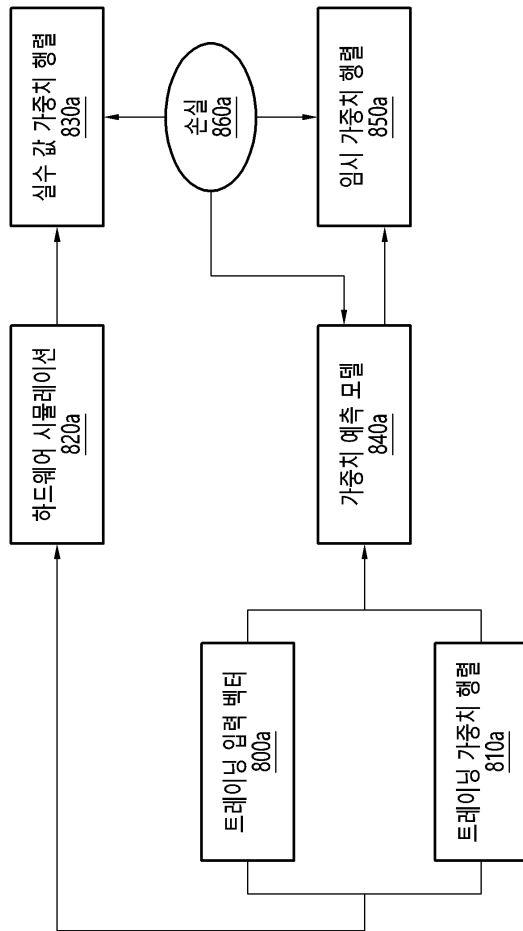
도면6b



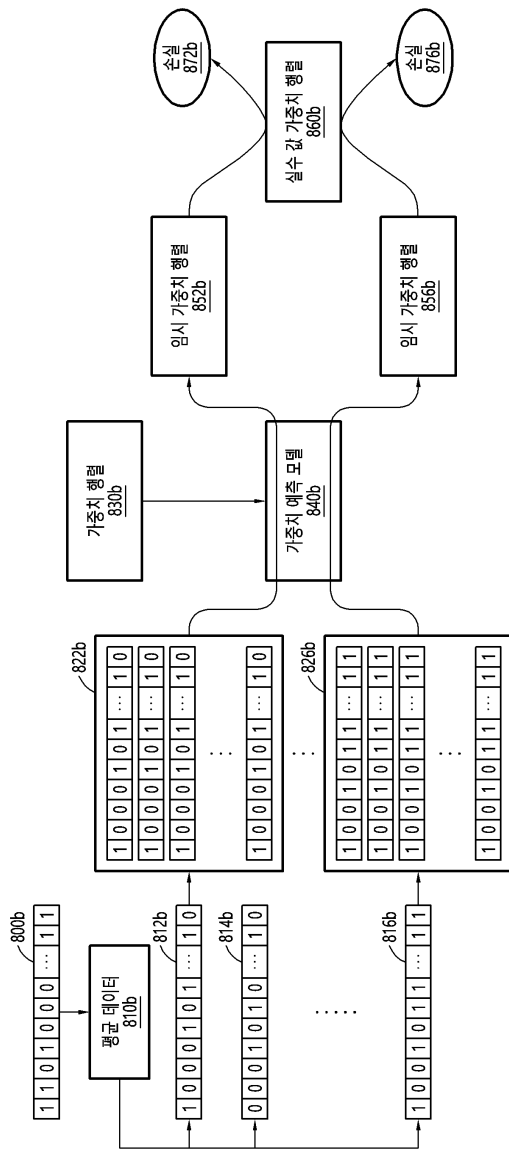
도면7



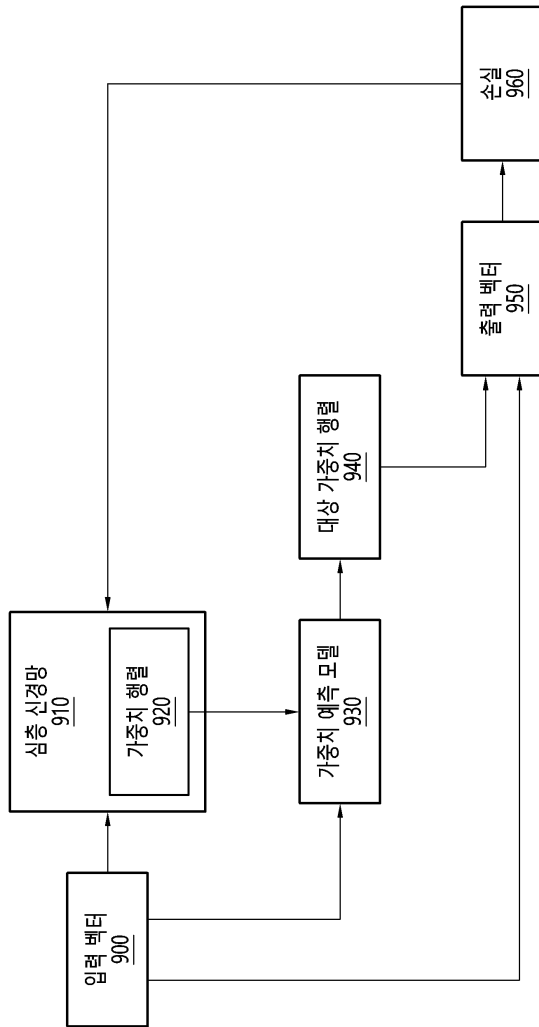
도면 8a



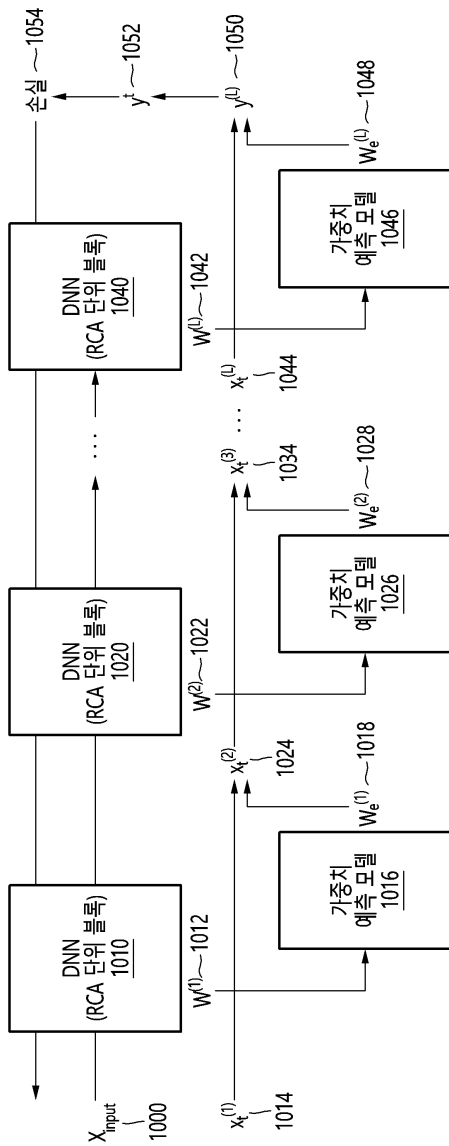
도면 8b



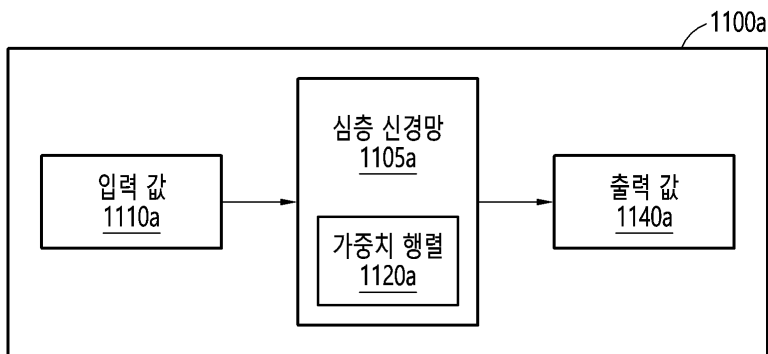
도면9



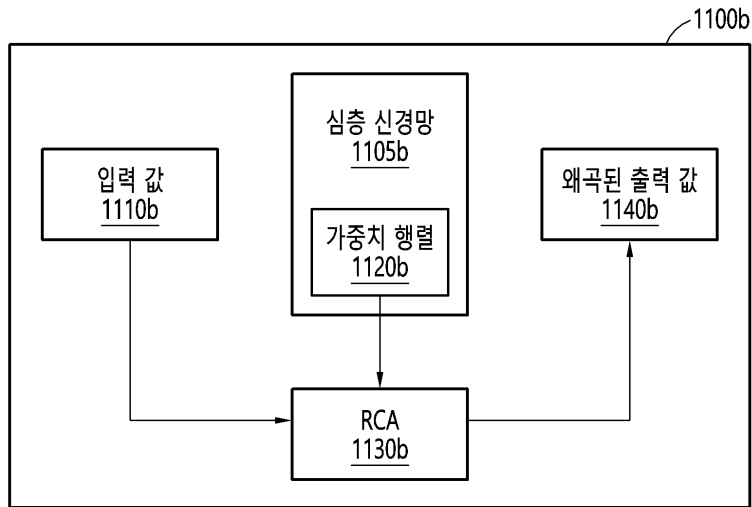
도면10



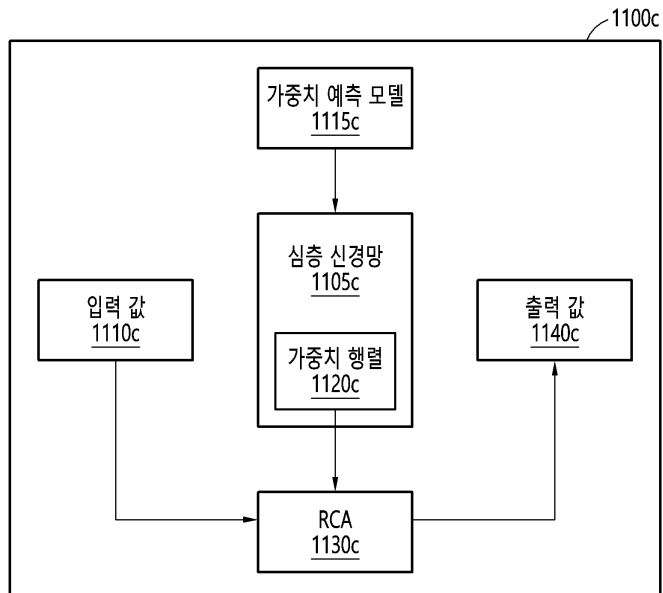
도면11a



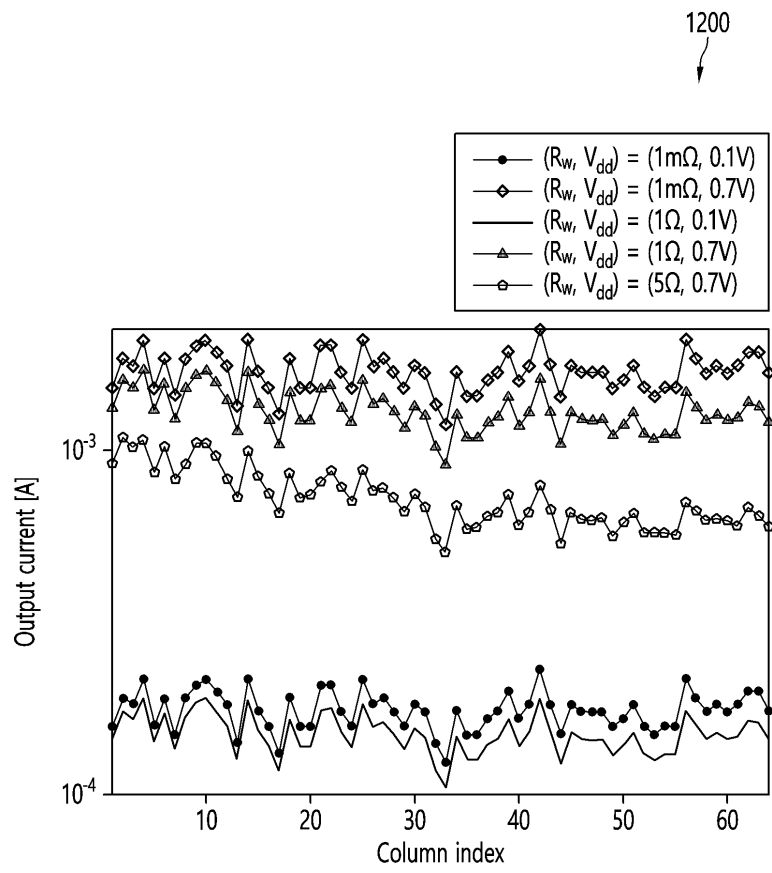
도면11b



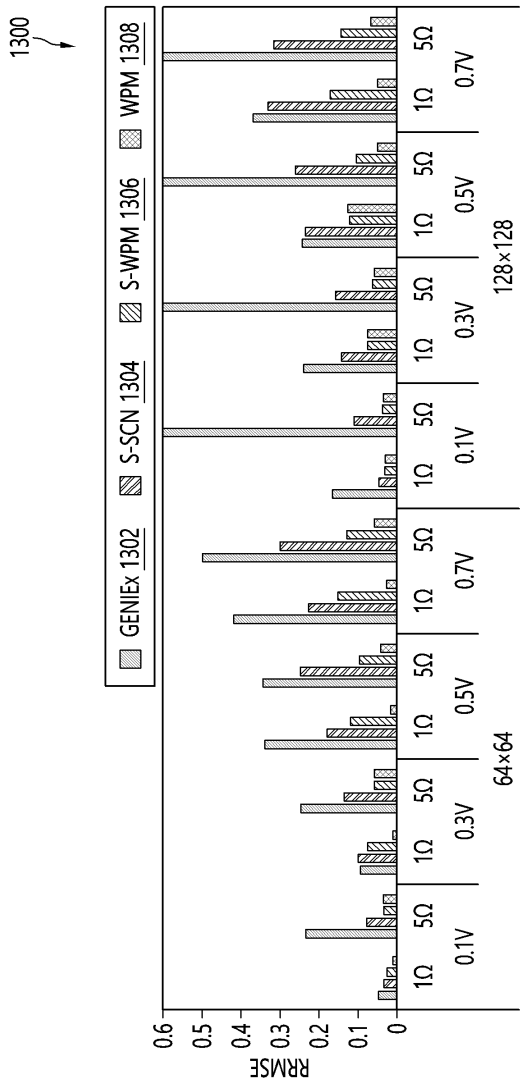
도면11c



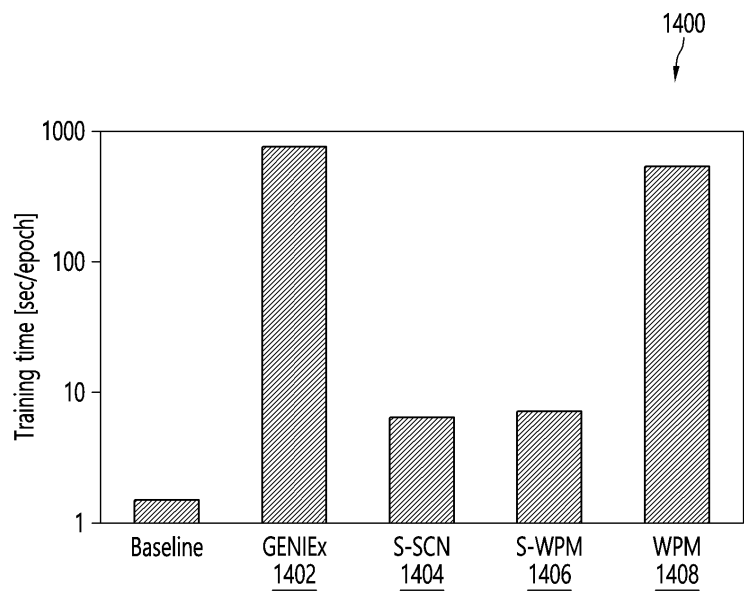
도면12



도면13



도면14



도면15

