



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0171618
(43) 공개일자 2023년12월21일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) G06F 17/16 (2006.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06F 17/16 (2013.01)
(21) 출원번호 10-2022-0071996
(22) 출원일자 2022년06월14일
심사청구일자 2022년06월14일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
이종은
울산광역시 울주군 언양읍 유니스트길 50 울산과
학기술원
이수길
울산광역시 울주군 언양읍 유니스트길 50 울산과
학기술원
(74) 대리인
특허법인지원

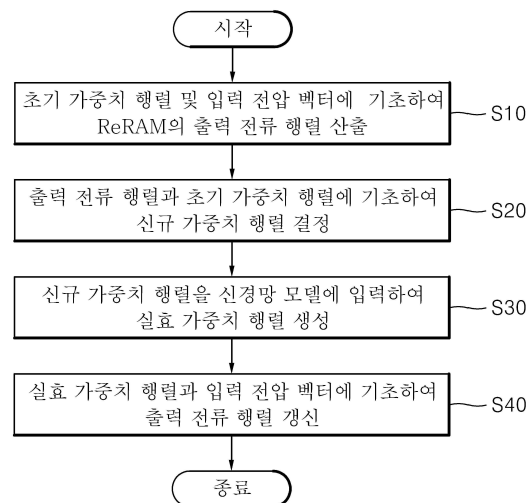
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 저항성 메모리 기반의 신경망 가속기에 적용되는 연산 오차 보정 방법

(57) 요약

본 발명은 저항성 메모리(ReRAM) 기반의 신경망 가속기에 적용되는 것으로서, 저항성 메모리의 IR 드랍(drop) 및 I-V 비선형성(nonlinearity)으로 인해 발생하는 행렬곱 계산 결과의 왜곡을 보정하는 연산 오차 보정 방법에 관한 것이다. 본 발명의 일 실시예에 따른 신경망 가속기의 연산 오차 보정 방법은 초기 가중치 행렬 및 입력 전압 벡터에 기초하여 ReRAM의 출력 전류 행렬을 산출하는 단계, 상기 출력 전류 행렬과 초기 가중치 행렬에 기초하여 신규 가중치 행렬을 결정하는 단계, 상기 신규 가중치 행렬을 신경망 모델에 입력하여 실효 가중치 행렬을 생성하는 단계 및 상기 실효 가중치 행렬과 상기 입력 전압 벡터에 기초하여 상기 출력 전류 행렬을 갱신하되, 갱신 전후의 상기 출력 전류 행렬 간의 상대 오차가 기준값 미만이 될 때까지 상기 출력 전류 행렬을 반복 갱신하는 단계를 포함하는 것을 특징으로 한다.

대 표 도 - 도3



(52) CPC특허분류
G06N 3/08 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711165109
과제번호	2020R1A2C201506612
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	핵심연구(개인)
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	대학ICT연구센터육성지원사업
연구과제명	스마트 모빌리티를 위한 인공지능 시스템반도체 핵심 기술 개발 및 인력 양성
기 여 율	40/100
과제수행기관명	인하대학교 산학협력단
연구기간	2021.07.01 ~ 2028.12.31

명세서

청구범위

청구항 1

행렬-벡터 곱 연산을 수행하는 저항성 메모리(ReRAM) 기반의 신경망 가속기에 적용되는 연산 오차 보정 방법에 있어서,

초기 가중치 행렬 및 입력 전압 벡터에 기초하여 상기 저항성 메모리의 출력 전류 행렬을 산출하는 단계;

상기 출력 전류 행렬과 상기 초기 가중치 행렬에 기초하여 신규 가중치 행렬을 결정하는 단계;

상기 신규 가중치 행렬을 신경망 모델에 입력하여 실효 가중치 행렬을 생성하는 단계; 및

상기 실효 가중치 행렬과 상기 입력 전압 벡터에 기초하여 상기 출력 전류 행렬을 갱신하되, 갱신 전후의 상기 출력 전류 행렬 간의 상대 오차가 기준값 미만이 될 때까지 상기 출력 전류 행렬을 반복 갱신하는 단계를 포함하는

신경망 가속기의 연산 오차 보정 방법.

청구항 2

제1항에 있어서,

상기 출력 전류 행렬을 산출하는 단계는

I-V 비선형성(nonlinearity)이 고려된 전류 도출 함수에 상기 초기 가중치 행렬 및 상기 입력 전압 벡터를 적용하여 상기 출력 전류 행렬을 산출하는 단계를 포함하는

신경망 가속기의 연산 오차 보정 방법.

청구항 3

제2항에 있어서,

상기 전류 도출 함수는 상기 저항성 메모리의 입출력 데이터를 커브 피팅(curve fitting)하여 결정되는

신경망 가속기의 연산 오차 보정 방법.

청구항 4

제2항에 있어서,

상기 전류 도출 함수는 상기 초기 가중치 행렬의 바이너리 값에 따라 하기 [수학식 1] 또는 [수학식 2]로 표현되는

[수학식 1]

$$i_{on} = a_{on}(e^{b_{on}v} - e^{-d_{on}v})$$

[수학식 2]

$$i_{off} = a_{off}(e^{b_{off}v} - e^{-d_{off}v})$$

(여기서 i_{on} 및 i_{off} 는 출력 전류, a_{on} , b_{on} , d_{on} , a_{off} , b_{off} , d_{off} 는 상기 저항성 메모리의 메타데이터에 의

해 결정되는 상수, v 는 입력 전압)
신경망 가속기의 연산 오차 보정 방법.

청구항 5

제1항에 있어서,
상기 신규 가중치 행렬을 결정하는 단계는
I-V 비선형성(nonlinearity)이 고려된 전류 도출 함수의 역함수에 상기 출력 전류 행렬과 상기 초기 가중치 행렬을 대입하여 전압 행렬을 산출하는 단계와,
상기 전압 행렬과 상기 출력 전류 행렬에 기초하여 컨덕턴스(conductance) 행렬을 산출하는 단계와,
상기 컨덕턴스 행렬에 비례하는 상기 신규 가중치 행렬을 결정하는 단계를 포함하는
신경망 가속기의 연산 오차 보정 방법.

청구항 6

제1항에 있어서,
상기 신경망 모델은 임의의 가중치 행렬을 입력 데이터로 하고, IR 드랍(drop)을 고려한 시뮬레이션 결과에 따라 생성된 가중치 행렬을 출력 데이터로 하는 훈련 데이터셋(training dataset)에 의해 지도학습(supervised learning)된
신경망 가속기의 연산 오차 보정 방법.

청구항 7

제1항에 있어서,
상기 출력 전류 행렬을 갱신하는 단계는
상기 출력 전류 행렬을, 상기 실효 가중치 행렬과 상기 입력 전압 벡터 간 행렬-벡터 곱 연산의 결과값으로 갱신하는 단계를 포함하는
신경망 가속기의 연산 오차 보정 방법.

청구항 8

제1항에 있어서,
상기 출력 전류 행렬을 갱신하는 단계는
상기 상대 오차가 기준값 이상이면 상기 갱신된 출력 전류 행렬과 상기 초기 가중치 행렬에 기초하여 상기 신규 가중치 행렬을 갱신하는 단계와,
상기 갱신된 신규 가중치 행렬을 상기 신경망 모델에 입력하여 상기 실효 가중치 행렬을 갱신하는 단계와,
상기 갱신된 실효 가중치 행렬과 상기 입력 전압 벡터에 기초하여 상기 출력 전류 행렬을 재갱신하는 단계를 포함하는
신경망 가속기의 연산 오차 보정 방법.

청구항 9

제1항에 있어서,

상기 갱신된 출력 전류 행렬 내 원소들을 합하여 상기 저항성 메모리의 출력 전류를 산출하는 단계를 더 포함하는

신경망 가속기의 연산 오차 보정 방법.

발명의 설명

기술 분야

[0001] 본 발명은 저항성 메모리(ReRAM) 기반의 신경망 가속기에 적용되는 것으로서, 저항성 메모리의 IR 드랍(drop) 및 I-V 비선형성(nonlinearity)으로 인해 발생하는 행렬곱 계산 결과의 왜곡을 보정하는 연산 오차 보정 방법에 관한 것이다.

배경 기술

[0003] 신경망 가속기(neural network acceleration)의 주요 동작은 행렬-벡터 곱(matrix-vector multiplication) 연산으로, 저항성 메모리 어레이(ReRAM Crossbar Array; RCA)는 행렬-벡터 곱 연산을 빠르고 효율적으로 수행할 수 있어 신경망 가속기를 위한 하드웨어로 이용되고 있다.

[0004] 저항성 메모리 어레이는 저항성 소자가 격자 형태로 배열된 구조를 갖는데, 여기서 곱하고자 하는 행렬의 값을 각 저항성 소자의 컨덕턴스(conductance) 값으로 매핑하고, 곱하고자 하는 벡터의 값을 전압으로 걸어주면, $I=V/R$ 에 따라 전류 값이 행렬곱의 결과로서 도출된다.

[0005] 그런데, 저항성 메모리 어레이에는 선저항(wire resistance), 기생 저항(parasitic resistance) 등의 요인으로 인해 전압원으로부터 멀어질수록 각 저항성 소자에 인가되는 전압이 예상치보다 낮아지는 IR 드랍 현상이 나타나며, 이로 인해 행렬곱의 결과로서 출력되는 전류 값에 왜곡이 발생하는 문제가 있다.

[0006] 또한, 저항성 소자의 컨덕턴스는 소자에 인가되는 전압에 따라 달라지며, 이러한 전압-종속 컨덕턴스(voltage-dependent conductance)는 저항성 메모리 어레이에 I-V 비선형성을 발생시켜 행렬곱의 결과로서 출력되는 전류 값에 왜곡이 발생하는 문제가 있다.

[0007] 이에 따라, 저항성 메모리 어레이에 발생하는 IR 드랍 및 I-V 비선형성을 고려하여 신경망 가속기의 출력 정확도를 높일 수 있는 방법이 요구된다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 저항성 메모리의 IR 드랍 및 I-V 비선형성으로 인해 발생하는 행렬곱 계산 결과의 왜곡을 보정하는 것을 목적으로 한다.

[0010] 본 발명의 목적들은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있고, 본 발명의 실시예에 의해 보다 분명하게 이해될 것이다. 또한, 본 발명의 목적 및 장점들은 특허 청구 범위에 나타낸 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

[0012] 전술한 목적을 달성하기 위한 본 발명의 일 실시예에 따른 신경망 가속기의 연산 오차 보정 방법은 행렬-벡터 곱 연산을 수행하는 저항성 메모리(ReRAM) 기반의 신경망 가속기에 적용되는 연산 오차 보정 방법에 있어서, 초

기 가중치 행렬 및 입력 전압 벡터에 기초하여 상기 저항성 메모리의 출력 전류 행렬을 산출하는 단계, 상기 출력 전류 행렬과 상기 초기 가중치 행렬에 기초하여 신규 가중치 행렬을 결정하는 단계, 상기 신규 가중치 행렬을 신경망 모델에 입력하여 실효 가중치 행렬을 생성하는 단계 및 상기 실효 가중치 행렬과 상기 입력 전압 벡터에 기초하여 상기 출력 전류 행렬을 갱신하되, 갱신 전후의 상기 출력 전류 행렬 간의 상대 오차가 기준값 미만일 때까지 상기 출력 전류 행렬을 반복 갱신하는 단계를 포함하는 것을 특징으로 한다.

[0013] 일 실시예에서, 상기 출력 전류 행렬을 산출하는 단계는 I-V 비선형성(nonlinearity)이 고려된 전류 도출 함수에 상기 초기 가중치 행렬 및 상기 입력 전압 벡터를 적용하여 상기 출력 전류 행렬을 산출하는 단계를 포함하는 것을 특징으로 한다.

[0014] 일 실시예에서, 상기 전류 도출 함수는 상기 저항성 메모리의 입출력 데이터를 커브 피팅(curve fitting)하여 결정되는 것을 특징으로 한다.

[0015] 일 실시예에서, 상기 전류 도출 함수는 상기 초기 가중치 행렬의 바이너리 값에 따라 하기 [수학식 1] 또는 [수학식 2]로 표현되는 것을 특징으로 한다.

[0016] [수학식 1]

$$i_{on} = a_{on}(e^{b_{on}v} - e^{-d_{on}v})$$

[0018] [수학식 2]

$$i_{off} = a_{off}(e^{b_{off}v} - e^{-d_{off}v})$$

[0020] (여기서 i_{on} 및 i_{off} 는 출력 전류, a_{on} , b_{on} , d_{on} , a_{off} , b_{off} , d_{off} 는 상기 저항성 메모리의 메타데이터에 의해 결정되는 상수, v 는 입력 전압)

[0021] 일 실시예에서, 상기 신규 가중치 행렬을 결정하는 단계는 I-V 비선형성(nonlinearity)이 고려된 전류 도출 함수의 역함수에 상기 출력 전류 행렬과 상기 초기 가중치 행렬을 대입하여 전압 행렬을 산출하는 단계와, 상기 전압 행렬과 상기 출력 전류 행렬에 기초하여 컨덕턴스(conductance) 행렬을 산출하는 단계와, 상기 컨덕턴스 행렬에 비례하는 상기 신규 가중치 행렬을 결정하는 단계를 포함하는 것을 특징으로 한다.

[0022] 일 실시예에서, 상기 신경망 모델은 임의의 가중치 행렬을 입력 데이터로 하고, IR 드랍(drop)을 고려한 시뮬레이션 결과에 따라 생성된 가중치 행렬을 출력 데이터로 하는 훈련 데이터셋(training dataset)에 의해 지도학습(supervised learning)된 것을 특징으로 한다.

[0023] 일 실시예에서, 상기 출력 전류 행렬을 갱신하는 단계는 상기 출력 전류 행렬을, 상기 실효 가중치 행렬과 상기 입력 전압 벡터 간 행렬-벡터 곱 연산의 결과값으로 갱신하는 단계를 포함하는 것을 특징으로 한다.

[0024] 일 실시예에서, 상기 출력 전류 행렬을 갱신하는 단계는 상기 상대 오차가 기준값 이상이면 상기 갱신된 출력 전류 행렬과 상기 초기 가중치 행렬에 기초하여 상기 신규 가중치 행렬을 갱신하는 단계와, 상기 갱신된 신규 가중치 행렬을 상기 신경망 모델에 입력하여 상기 실효 가중치 행렬을 갱신하는 단계와, 상기 갱신된 실효 가중치 행렬과 상기 입력 전압 벡터에 기초하여 상기 출력 전류 행렬을 재갱신하는 단계를 포함하는 것을 특징으로 한다.

[0025] 일 실시예에서, 상기 갱신된 출력 전류 행렬 내 원소들을 합하여 상기 저항성 메모리의 출력 전류를 산출하는 단계를 더 포함하는 것을 특징으로 한다.

발명의 효과

[0027] 본 발명은 저항성 메모리의 IR 드랍 및 I-V 비선형성으로 인해 발생하는 행렬곱 계산 결과의 왜곡을 보정함으로써, 저항성 메모리 기반 신경망 가속기의 추론 성능을 향상시킬 수 있다.

[0028] 상술한 효과와 더불어 본 발명의 구체적인 효과는 이하 발명을 실시하기 위한 구체적인 사항을 설명하면서 함께 기술한다.

도면의 간단한 설명

- [0030] 도 1은 저항성 메모리 기반 신경망 가속기의 구조를 도시한 도면.
- 도 2a 및 도 2는 MVM 연산에 왜곡이 발생하는 문제와 이를 보상하는 본 발명의 동작을 개략적으로 설명하기 위한 도면.
- 도 3은 본 발명의 일 실시예에 따른 연산 오차 보정 방법을 도시한 순서도.
- 도 4는 커브 피팅(curve fitting)을 통해 저항성 메모리의 전류 도출 함수를 결정하는 과정을 설명하기 위한 도면.
- 도 5는 출력 전류 행렬을 반복적으로 갱신함으로써 연산 오차를 보정하는 본 발명의 예시적 동작을 도시한 순서도.

발명을 실시하기 위한 구체적인 내용

- [0031] 전술한 목적, 특징 및 장점은 첨부된 도면을 참조하여 상세하게 후술되며, 이에 따라 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 본 발명의 기술적 사상을 용이하게 실시할 수 있을 것이다. 본 발명을 설명함에 있어서 본 발명과 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 상세한 설명을 생략한다. 이하, 첨부된 도면을 참조하여 본 발명에 따른 바람직한 실시예를 상세히 설명하기로 한다. 도면에서 동일한 참조부호는 동일 또는 유사한 구성요소를 가리키는 것으로 사용된다.
- [0032] 본 명세서에서 제1, 제2 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것으로, 특별히 반대되는 기재가 없는 한, 제1 구성요소는 제2 구성요소일 수도 있음은 물론이다.
- [0033] 또한, 본 명세서에서 "상부 (또는 하부)" 또는 구성요소의 "상 (또는 하)"에 임의의 구성이 배치된다는 것은, 임의의 구성이 상기 구성요소의 상면 (또는 하면)에 접하여 배치되는 것뿐만 아니라, 상기 구성요소와 상기 구성요소 상에 (또는 하에) 배치된 임의의 구성 사이에 다른 구성이 개재될 수 있음을 의미할 수 있다.
- [0034] 또한, 본 명세서에서 어떤 구성요소가 다른 구성요소에 "연결", "결합" 또는 "접속"된다고 기재된 경우, 상기 구성요소들은 서로 직접적으로 연결되거나 또는 접속될 수 있지만, 각 구성요소 사이에 다른 구성요소가 "개재"되거나, 각 구성요소가 다른 구성요소를 통해 "연결", "결합" 또는 "접속"될 수도 있는 것으로 이해되어야 할 것이다.
- [0035] 또한, 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서 상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다.
- [0036] 또한, 본 명세서에서, "A 및/또는 B" 라고 할 때, 이는 특별한 반대되는 기재가 없는 한, A, B 또는 A 및 B 를 의미하며, "C 내지 D" 라고 할 때, 이는 특별한 반대되는 기재가 없는 한, C 이상이고 D 이하인 것을 의미한다.
- [0037] 본 발명은 저항성 메모리(ReRAM) 기반의 신경망 가속기에 적용되는 것으로서, 저항성 메모리의 IR 드랍(drop) 및 I-V 비선형성(nonlinearity)으로 인해 발생하는 행렬곱 계산 결과의 왜곡을 보정하는 연산 오차 보정 방법에 관한 것이다. 이하, 도 1 내지 도 5를 참조하여 본 발명의 일 실시예에 따른 연산 오차 보정 방법을 구체적으로 설명하도록 한다.
- [0038] 도 1은 저항성 메모리 기반 신경망 가속기의 구조를 도시한 도면이다.
- [0039] 도 2a 및 도 2는 MVM 연산에 왜곡이 발생하는 문제와 이를 보상하는 본 발명의 동작을 개략적으로 설명하기 위한 도면이다.
- [0040] 도 3은 본 발명의 일 실시예에 따른 연산 오차 보정 방법을 도시한 순서도이다.
- [0041] 도 4는 커브 피팅(curve fitting)을 통해 저항성 메모리의 전류 도출 함수를 결정하는 과정을 설명하기 위한 도면이다.

- [0042] 도 5는 출력 전류 행렬을 반복적으로 갱신함으로써 연산 오차를 보정하는 본 발명의 예시적 동작을 도시한 순서도이다.
- [0043] 도 1을 참조하면, 저항성 메모리(10)는 크로스바 어레이의 구조의 저항 변화 메모리(ReRAM Crossbar Arrays, RCA)일 수 있고, 행렬-벡터 곱(Matrix-Vector Multiplication; MVM) 연산을 수행하는 수동형 RCA일 수 있다. 이러한 저항성 메모리(10)는 MVM 연산을 위한 신경망 가속기(neural network accelerator)에 적용될 수 있다.
- [0044] 구체적으로, 도 1에 도시된 바와 같이, 저항성 메모리(10)는 n차원의 입력 벡터가 인가되는 n개의 워드 라인(WL)과, $n \times m$ 크기의 가중치 행렬이 매핑되는 $n \times m$ 개의 저항성 소자와, MVM 연산에 따른 m차원의 출력 벡터를 도출하는 m개의 비트 라인(BL)을 포함할 수 있다.
- [0045] 저항성 메모리(10)에 인가되는 입력 벡터가 X, 가중치 행렬이 W, 출력 벡터가 Y라고 할 때, $Y=XW$ 가 성립하며, 이에 따라 저항성 메모리(10)는 MVM 연산의 결과를 출력하게 된다.
- [0046] 한편, 입력 벡터는 n개의 워드 라인(WL)에 인가되는 입력 전압(V)으로 매핑될 수 있고, 가중치 행렬 내 원소들은 각 저항성 소자의 컨덕턴스 값(G_{ij})에 매핑될 수 있다. 이 경우 옴의 법칙(Ohm's law)에 의해 저항성 메모리(10)는 아래 [수학식 1]에 따라 전류 값(I)을 출력할 수 있다.
- [0047] [수학식 1]
- [0048] $W \cdot V = I$
- [0049] 한편, 입력 전압($V=[v_1, v_2, \dots, v_n]$), 가중치 행렬(G_{ij}) 및 전류 값($I=[i_1, i_2, \dots, i_m]$)은 모두 벡터로 표현될 수 있으므로 [수학식 1]은 아래 [수학식 2]와 같이 나타낼 수 있다.
- [0050] [수학식 2]
- [0051]
$$\begin{bmatrix} G_{11} & \dots & G_{1m} \\ \vdots & \ddots & \vdots \\ G_{n1} & \dots & G_{nm} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = [i_1 \quad \dots \quad i_m]$$
- [0052] 이와 같이, 저항성 메모리(10)는 n개의 워드 라인(WL)에 입력되는 n차원의 입력 전압 벡터와, 각 저항성 소자의 컨덕턴스 값으로 결정되는 가중치 행렬 간의 MVM 연산을 수행하고, m개의 비트 라인(BL)을 통해 MVM 연산의 출력인 전류 벡터를 출력할 수 있다.
- [0053] 그런데, 저항성 메모리(10)는 그 구조상 선저항(wire resistance), 기생 저항(parasitic resistance) 등의 요인으로 인해 전압원으로부터 멀어질수록 각 저항성 소자에 인가되는 전압이 예상치보다 낮아지는 IR 드랍 현상이 나타나며, 이로 인해 MVM 연산의 결과로서 출력되는 전류 벡터에 왜곡이 발생하는 문제가 있다.
- [0054] 또한, 저항성 소자의 컨덕턴스는 소자에 인가되는 전압에 따라 달라지며, 이러한 전압-종속 컨덕턴스(voltage-dependent conductance)는 저항성 메모리(10)에 I-V 비선형성을 발생시켜 MVM 연산의 결과로서 출력되는 전류 벡터에 왜곡이 발생하는 문제가 있다.
- [0055] SPICE(Simulation Program with Integrated Circuit Emphasis)를 이용한 정확한 시뮬레이션을 통해 IR 드랍과 I-V 비선형성을 고려한 출력(전류 벡터)을 얻어낼 수는 있으나, 이 방법은 지나치게 많은 시간을 소요하므로 실시간 연산을 필요로 하는 신경망 가속기에 적용하는 것이 불가능하다.
- [0056] 도 2a를 참조하면, 앞서 설명한 요인에 따라 왜곡이 발생하는 경우, 입력 전압과 가중치 행렬의 MVM 연산의 결과 참값(I)과는 다른 왜곡된(distorted) 결과(I')가 도출될 수 있으며, 이는 아래 [수학식 3]과 같은 부등식으로 나타낼 수 있다.
- [0057] [수학식 3]
- [0058] $I' = g(W, V) \neq W \cdot V$
- [0059] (여기서 g는 왜곡된 연산 결과를 도출하는 임의의 함수)
- [0060] SPICE를 이용하는 경우 [수학식 3]에 기재된 함수 g를 추정할 수 있으나, 앞서 설명한 바와 같이 이러한 방식은 많은 시간을 소요하므로, 본 발명에서는 함수 g를 추정하는 대신, 도 2b에 도시된 바와 같이 가중치 행렬을 이

용하여, 입력 전압과의 MVM 연산의 결과가 아래 [수학식 4]를 만족하는 실효 가중치 행렬을 예측할 수 있다.

[0061] [수학식 4]

$$I = W_e \cdot V$$

[0063] 이하에서는 실효 가중치 행렬을 예측하고 이를 통해 저항성 메모리(10)의 출력 전류를 결정하는 과정을 구체적으로 설명하도록 한다.

[0064] 도 3을 참조하면, 본 발명의 일 실시예에 따른 연산 오차 보정 방법은 초기 가중치 행렬(W) 및 입력 전압 벡터(V)에 기초하여 저항성 메모리(10)의 출력 전류 행렬(I_d)을 산출하는 단계(S10), 출력 전류 행렬(I_d)과 초기 가중치 행렬(W)에 기초하여 신규 가중치 행렬(W')을 결정하는 단계(S20), 신규 가중치 행렬(W')을 신경망(Neural Network; NN) 모델에 입력하여 실효 가중치 행렬(W_e)을 생성하는 단계(S30) 및 실효 가중치 행렬(W_e)과 입력 전압 벡터(V)에 기초하여 출력 전류 행렬(I_d)을 갱신하는 단계(S40)를 포함할 수 있다.

[0065] 다만, 도 3에 도시된 연산 오차 보정 방법은 일 실시예에 따른 것이고, 발명을 이루는 각 단계들이 도 3에 도시된 실시예에 한정되는 것은 아니며, 필요에 따라 일부 단계가 부가, 변경 또는 삭제될 수 있다.

[0066] 도 3에 도시된 각 단계들은 CPU(central processing unit), GPU(graphics processing unit) 등과 같은 프로세서에 의해 수행될 수 있으며, 프로세서는 후술하는 발명의 동작을 수행하기 위하여 ASICs(application specific integrated circuits), DSPs(digital signal processors), DSPDs(digital signal processing devices), PLDs(programmable logic devices), FPGAs(field programmable gate arrays), 제어기(controller), 마이크로 컨트롤러(micro-controllers) 중 적어도 하나의 물리적인 요소를 더 포함할 수 있다.

[0067] 이하, 도 3에 도시된 각 단계들에 대해 구체적으로 설명하도록 한다.

[0068] 프로세서는 MVM 연산의 대상이 되는 초기 가중치 행렬(W)과 입력 전압 벡터(V)에 기초하여 저항성 메모리(10)의 출력 전류 행렬(I_d)을 산출할 수 있다(S10). 이 때, 프로세서는 도 1에서 설명한 저항성 메모리(10)에 입력 전압 벡터(V)와 초기 가중치 행렬(W)을 실제로 적용하는 대신에, 시뮬레이션을 통해 출력 전류 행렬(I_d)을 산출할 수 있다.

[0069] 시뮬레이션에서, 프로세서는 I-V 비선형성이 고려된 전류 도출 함수에 초기 가중치 행렬(W)과 입력 전압 벡터(V)를 적용하여 출력 전류 행렬(I_d)을 산출할 수 있다. 본 발명에서 전류 도출 함수는 저항성 메모리(10)의 입출력 데이터에 의해 귀납적으로 도출되므로, 전류 도출 함수에는 I-V 비선형성이 반영될 수 있다.

[0070] 도 4를 참조하여 구체적으로 설명하면, 프로세서는 일정 범위의 전압(입력 데이터)을 저항성 메모리(10)에 입력하고, 저항성 메모리(10)에서 출력되는 전류(출력 데이터)를 검출할 수 있다. 프로세서는 이와 같은 입출력 데이터를 커브 피팅(curve fitting)하여 저항성 메모리(10)의 메타데이터(metadata)와 I-V 비선형성에 의해 결정되는 전류 도출 함수를 구할 수 있다.

[0071] 전류 도출 함수는 지수(exponential) 함수, 하이퍼볼릭 사인(sinh) 함수 등으로 피팅될 수 있고, 예컨대 아래 [수학식 5] 및 [수학식 6]과 같이 지수 함수로 표현될 수 있다. 전류 도출 함수에서 저항성 메모리(10)의 메타데이터에 의해 결정되는 상수들은 커브 피팅에 의해 그 값이 결정될 수 있다.

[0072] [수학식 5]

$$i_{on} = a_{on}(e^{b_{on}v} - e^{-d_{on}v})$$

[0074] [수학식 6]

$$i_{off} = a_{off}(e^{b_{off}v} - e^{-d_{off}v})$$

[0076] (여기서 i_{on} 및 i_{off} 는 출력 전류, a_{on} , b_{on} , d_{on} , a_{off} , b_{off} , d_{off} 는 저항성 메모리(10)의 메타데이터에 의해 결정되는 상수, v는 입력 전압)

[0077] 앞서 설명한 초기 가중치 행렬(W)의 각 원소들은 바이너리(binary) 값을 가질 수 있고, 프로세서는 초기 가중치 행렬(W)의 바이너리 값에 따라 출력 전류 행렬(I_d) 산출에 [수학식 5] 또는 [수학식 6]을 적용할 수 있다. 예컨

대, 프로세서는 초기 가중치 행렬(W)의 원소 값이 1이면 [수학식 5]를 통해 출력 전류를 산출할 수 있고, 원소 값이 0이면 [수학식 6]을 통해 출력 전류를 산출할 수 있다.

[0078] 프로세서는 입력 전압 벡터(V)를 구성하는 각 원소(입력 전압)별로 출력 전류(i : i_{on} 또는 i_{off})를 산출할 수 있고, 산출된 출력 전류(i)의 집합인 출력 전류 행렬(I_d)을 생성할 수 있다.

[0079] 단계(S10)에 따라 출력 전류 행렬(I_d)이 산출되면, 프로세서는 출력 전류 행렬(I_d)과 초기 가중치 행렬(W)에 기초하여 신규 가중치 행렬(W')을 결정할 수 있다(S20). 여기서 신규 가중치 행렬(W')은 I-V 비선형성을 고려하여 보정된 행렬로서, 신규 가중치 행렬(W')의 결정에는 전술한 전류 도출 함수가 이용될 수 있다.

[0080] 구체적으로, 프로세서는 전류 도출 함수를 이용하여 출력 전류 행렬(I_d)에 대응하는 저항성 메모리(10)의 컨덕턴스(conductance)를 산출하고, 이에 기초하여 신규 가중치 행렬(W')을 결정할 수 있다.

[0081] 컨덕턴스 행렬(G_d)의 산출을 위해 먼저 프로세서는 I-V 비선형성이 고려된 전류 도출 함수의 역함수에, 출력 전류 행렬(I_d)과 초기 가중치 행렬(W)을 대입하여 전압 행렬(V_d)을 산출할 수 있다. 예컨대, 전술한 전류 도출 함수를 h 라고 한다면 프로세서는 아래 [수학식 7]에 따라 전압 행렬(V_d)을 산출할 수 있다.

[0082] [수학식 7]

$$V_d = h^{-1}(I_d, W)$$

[0084] 이어서, 프로세서는 전압 행렬(V_d)과 출력 전류 행렬(I_d)에 기초하여 컨덕턴스 행렬(G_d)을 산출할 수 있고, 예컨대 아래 [수학식 8]에 따라 컨덕턴스 행렬(G_d)을 산출할 수 있다.

[0085] [수학식 8]

$$G_d = I_d / V_d$$

[0087] (여기서 / 는 요소별 나눗셈(element-wise division)을 의미)

[0088] 앞서 도 1을 참조하여 설명한 바와 같이, 가중치 행렬은 저항성 소자의 컨덕턴스에 의해 결정될 수 있다. 이에 따라, 프로세서는 컨덕턴스 행렬(G_d)에 비례하는 신규 가중치 행렬(W')을 결정할 수 있고, 예컨대 아래 [수학식 9]에 따라 신규 가중치 행렬(W')을 결정할 수 있다.

[0089] [수학식 9]

$$W' = c \cdot G_d$$

[0091] (여기서 c 는 단위 변환을 위한 비례 상수)

[0092] 단계(S20)에 따라 신규 가중치 행렬(W')이 결정되면, 프로세서는 신규 가중치 행렬(W')을 신경망 모델에 입력하여 실효 가중치 행렬(W_e)을 생성할 수 있다(S30).

[0093] 본 발명에서 신경망 모델은 IR 드랍을 고려하기 위한 수단으로서, 신규 가중치 행렬(W')을 입력받고, IR 드랍의 영향이 고려된 실효 가중치 행렬(W_e)을 출력할 수 있다. 이를 위해, 신경망 모델은 미리 지도학습(supervised learning)될 수 있다.

[0094] 구체적으로, 신경망 모델은 임의의 가중치 행렬을 입력 데이터(input data)로 하고, IR 드랍을 고려한 시뮬레이션(예컨대, SPICE) 결과에 따라 생성된 실측(Ground Truth; GT) 가중치 행렬을 출력 데이터(output data, 또는 라벨 데이터(label data))로 하는 훈련 데이터셋(training dataset)에 의해 지도학습될 수 있다. 이에 따라, 신경망 모델은 IR 드랍과 이에 따른 가중치 행렬의 변화 간의 상관관계를 학습할 수 있다.

[0095] 학습이 완료된 신경망 모델은 앞서 I-V 비선형성이 고려된 신규 가중치 행렬(W')을 입력받고, IR 드랍의 영향이 추가적으로 고려된 실효 가중치 행렬(W_e)을 출력할 수 있다.

[0096] 한편, 신경망 모델에 입력되는 신규 가중치 행렬(W')은 행과 열로 구성되므로, 행렬 데이터 구조에 대한 학습 성능을 향상시키기 위해 신경망 모델은 RCN(Row-Column Network)을 포함할 수 있다. 또한, IR 드랍은 저항성 소

자의 공간적 배치에 영향을 받으므로, 공간적 상관 관계에 대한 학습 성능을 향상시키기 위해 신경망 모델은 SCN(Scaling Convolutional Network)을 포함할 수도 있다. 이 외에도, 신경망 모델은 당해 기술분야에서 이용되는 다양한 신경망과, 그에 따른 아키텍처를 포함할 수 있다.

[0097] 단계(S30)에 따라 실효 가중치 행렬(W_e)이 생성되면, 프로세서는 실효 가중치 행렬(W_e)과 입력 전압 벡터(V)에 기초하여 출력 전류 행렬(I_d)을 갱신할 수 있다(S40). 다시 말해, 단계(S10)에서 초기 가중치 행렬(W)에 따라 산출된 출력 전류 행렬(I_d)은, 실효 가중치 행렬(W_e)에 따라 산출된 출력 전류 행렬(I_d)로 갱신될 수 있다.

[0098] 프로세서는 출력 전류 행렬(I_d)을, 실효 가중치 행렬(W_e)과 입력 전압 벡터(V) 간 MVM 연산의 결과값으로 갱신할 수 있고, 예컨대 출력 전류 행렬(I_d)을 아래 [수학식 10]의 결과값으로 갱신할 수 있다.

[0099] [수학식 10]

$$I_{d,new} = W_e \circ (V \otimes \vec{1})$$

[0101] (여기서 $\circ$ 는 요소별 곱셈(element-wise multiplication), $\otimes$ 는 외적(outer-product), $\vec{1}$ 는 원소 값이 모두 1인 열(row) 벡터)

[0102] 한편, 출력 전류 행렬(I_d)의 갱신 동작은 출력 전류 행렬(I_d)이 일정 수준으로 수렴할 때까지 반복될 수 있다. 구체적으로, 갱신 동작은 갱신 전후의 출력 전류 행렬(I_d) 간의 상대 오차가 기준값 미만일 때까지 반복될 수 있고, 예컨대 갱신 동작은 하기 [수학식 11]이 만족될 때까지 반복될 수 있다.

[0103] [수학식 11]

$$|(I_{d,new} - I_d)/I_d| < \eta$$

[0105] (여기서 η 는 미리 설정된 기준값)

[0106] 이하에서는 갱신 전후의 출력 전류 행렬(I_d) 간의 상대 오차가 기준값 이상일 때, 프로세서에 의해 수행되는 추가 갱신 동작을 단일 사이클로 설명하나, 추가 갱신 동작은 갱신 전후의 출력 전류 행렬(I_d) 간의 상대 오차가 기준값 미만일 때까지 반복될 수 있다.

[0107] 프로세서는 출력 전류 행렬(I_d) 간의 상대 오차가 기준값 이상이면, 이전 사이클에서 갱신된 출력 전류 행렬(I_d)과 초기 가중치 행렬(W)에 기초하여 신규 가중치 행렬(W')을 갱신할 수 있다.

[0108] 구체적으로, 프로세서는 갱신된 출력 전류 행렬(I_d)을 [수학식 7]에 대입하여, 전압 행렬(V_d)을 갱신할 수 있고, 각각 갱신된 출력 전류 행렬(I_d) 및 전압 행렬(V_d)을 [수학식 8]에 대입하여 컨덕턴스 행렬(G_d)을 갱신할 수 있다. 이어서, 프로세서는 갱신된 컨덕턴스 행렬(G_d)을 [수학식 9]에 대입하여 신규 가중치 행렬(W')을 갱신할 수 있다.

[0109] 이어서, 프로세서는 갱신된 신규 가중치 행렬(W')을 신경망 모델에 입력하여 실효 가중치 행렬(W_e)을 갱신할 수 있다. 신경망 모델의 동작에 대해서는 전술한 바 있으므로 여기서는 자세한 설명을 생략하도록 한다.

[0110] 이어서, 프로세서는 갱신된 실효 가중치 행렬(W_e)과 입력 전압 벡터(V)에 기초하여 출력 전류 행렬(I_d)을 재갱신할 수 있다. 구체적으로, 프로세서는 갱신된 실효 가중치 행렬(W_e)과 입력 전압 벡터(V)를 [수학식 10]에 대입하여 출력 전류 행렬(I_d)을 재갱신할 수 있다.

[0111] 앞서 설명한 갱신 동작이 반복되어 갱신 전후의 출력 전류 행렬(I_d) 간의 상대 오차가 기준값 미만이면, 프로세서는 갱신 동작을 중단하고 마지막으로 갱신된 출력 전류 행렬(I_d) 내 원소들을 합하여 저항성 메모리(10)의 출력 전류(i)를 산출할 수 있다.

[0112] 도 5를 참조하여, 이상에서 설명한 본 발명의 동작을 예시적으로 종합하도록 한다.

[0113] 프로세서는 초기 가중치 행렬(W) 및 입력 전압 벡터(V)를 입력받을 수 있다(S11). 첫번째 사이클에서($n <= 1$)

프로세서는 I-V 비선형성이 고려된 전류 도출 함수에 초기 가중치 행렬(W)과 입력 전압 벡터(V)를 적용함으로써 출력 전류 행렬(I_d)을 초기화(initialize)할 수 있다(S12).

[0114] 이어서, 프로세서는 초기화된 출력 전류 행렬(I_d)과 초기 가중치 행렬(W)을, 전류 도출 함수의 역함수에 대입하여 전압 행렬(V_d)을 산출할 수 있고(S21), 출력 전류 행렬(I_d)과 단계(S21)에서 산출된 전압 행렬(V_d) 간의 비에 기초하여 신규 가중치 행렬(W')을 결정할 수 있다(S22).

[0115] 이어서, 프로세서는 신경망 모델에 신규 가중치 행렬(W')을 입력함으로써 실효 가중치 행렬(W_e)을 생성할 수 있고(S30), 실효 가중치 행렬(W_e)과 입력 전압 벡터(V)를 MVM 연산하여 출력 전류 행렬(I_d)을 갱신할 수 있다(S40).

[0116] 프로세서는 갱신 전후의 상기 출력 전류 행렬(I_d) 간의 상대 오차와 기준값을 비교하여 출력 전류 행렬(I_d)이 수렴하는지를 판단할 수 있고(S50), 상대 오차가 기준값 미만일 때까지 출력 전류 행렬(I_d)의 갱신 동작을 반복($n < n+1$)할 수 있다(S60).

[0117] 갱신 동작의 반복이 중단되면, 프로세서는 출력 전류 행렬(I_d) 내 원소들을 합하여 저항성 메모리의 출력 전류(i)를 산출할 수 있다(S70).

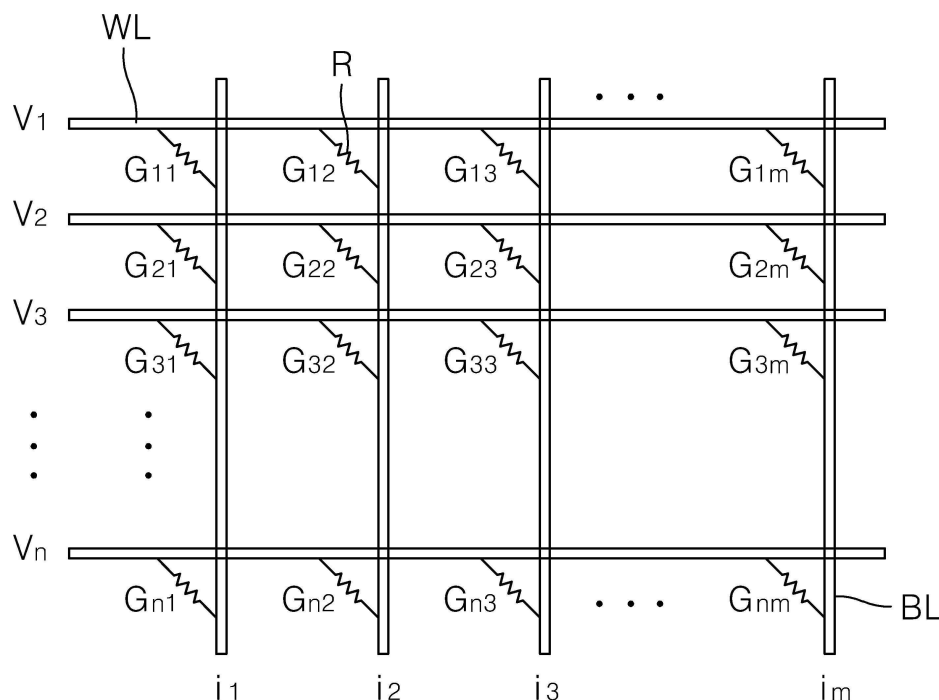
[0118] 전술한 바와 같이, 본 발명은 저항성 메모리(10)의 IR 드랍 및 I-V 비선형성으로 인해 발생하는 행렬곱 계산 결과의 왜곡을 보정함으로써, 저항성 메모리(10) 기반 신경망 가속기의 추론 성능을 향상시킬 수 있다.

[0119] 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.

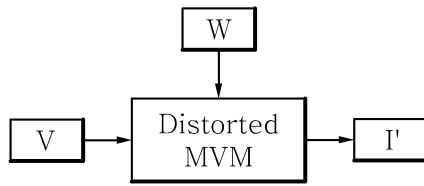
도면

도면1

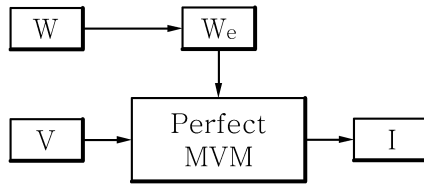
10



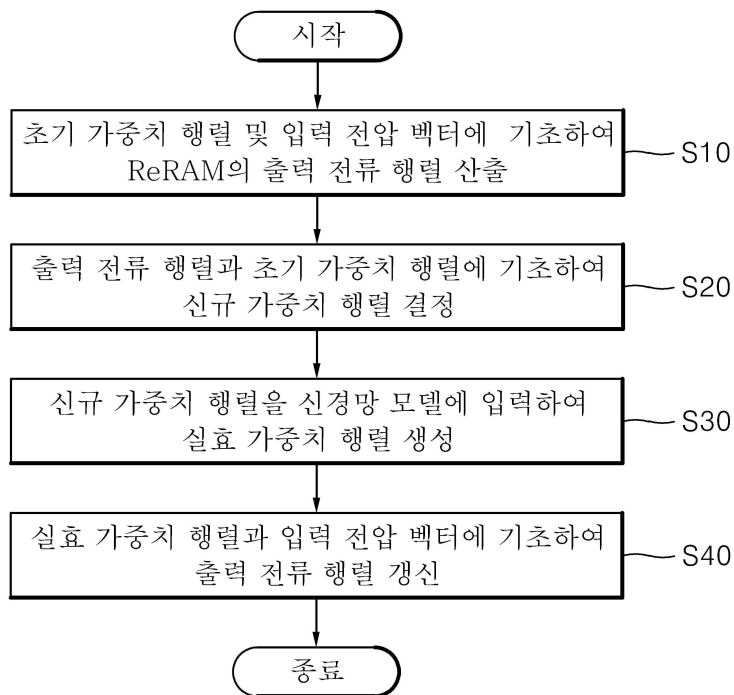
도면2a



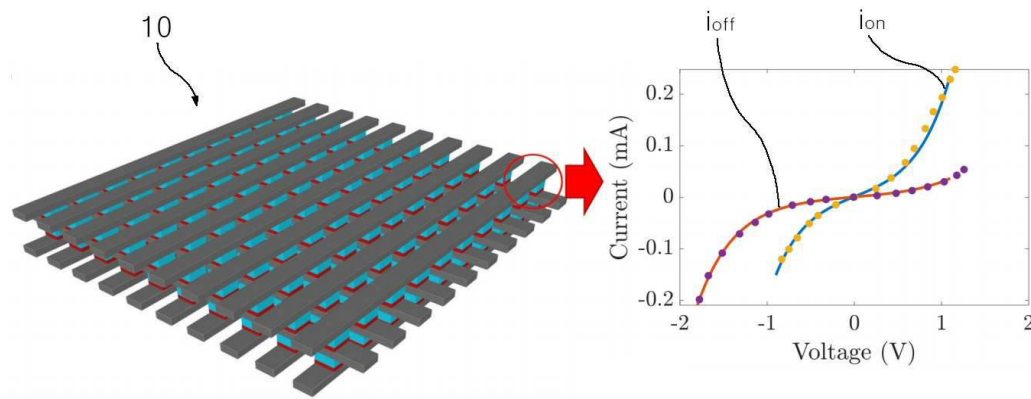
도면2b



도면3



도면4



도면5

