



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0077159
(43) 공개일자 2021년06월25일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/049 (2013.01)
(21) 출원번호 10-2019-0168488
(22) 출원일자 2019년12월17일
심사청구일자 2019년12월17일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
정성욱
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C513(신촌동)
장효정
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C206(신촌동)
김기룡
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C206(신촌동)
(74) 대리인
민영준

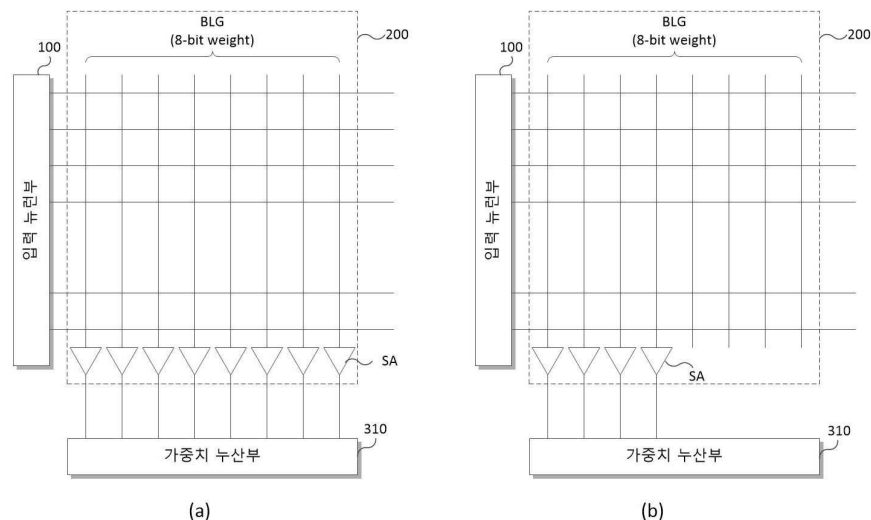
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 스파이킹 신경망 및 이의 운용 방법

(57) 요약

본 발명은 다수의 워드 라인과 다수의 비트 라인에 의해 정의되는 다수의 메모리 셀을 포함하여 가중치를 저장하는 가중치 시냅스부, 입력 데이터에 응답하여 다수의 입력 스파이크를 생성하여 다수의 워드 라인 중 대응하는 워드 라인을 활성화하는 입력 뉴런부 및 활성화된 워드 라인과 다수의 비트 라인이 교차하는 영역에 배치된 다수의 메모리 셀 각각에 비트값이 저장된 n비트의 가중치 중 하위 k비트를 제외하여 상위 n-k비트로 비트 스케일링된 가중치를 인가받아 누산하고, 이전 비트 스케일링된 가중치가 누산된 막전위값을 가산하여 막전위값을 획득하며, 획득된 막전위값이 비트 스케일링된 가중치에 대응하여 비트 스케일링된 발화 문턱값 이상이면 출력 스파이크를 발화하고, 발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 출력 뉴런부를 포함하는 스파이킹 신경망 및 이의 운용 방법을 제공할 수 있다.

대표도



이 발명을 지원한 국가연구개발사업

과제고유번호	2017R1A2B2006679
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	중견연구자지원사업
연구과제명	Domain Wall Motion 시냅스 기반의 On-Chip 지도-자율 통합학습 뉴로모픽 SoC
개발(3/3)(2017.3.1~2020.2.29)	
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2019.03.01 ~ 2020.02.29

명세서

청구범위

청구항 1

다수의 워드 라인과 다수의 비트 라인에 의해 정의되는 다수의 메모리 셀을 포함하여 가중치를 저장하는 가중치 시냅스부;

입력 데이터에 응답하여 다수의 입력 스파이크를 생성하여 상기 다수의 워드 라인 중 대응하는 워드 라인을 활성화하는 입력 뉴런부; 및

상기 활성화된 워드 라인과 상기 다수의 비트 라인이 교차하는 영역에 배치된 다수의 메모리 셀 각각에 비트값이 저장된 n 비트의 가중치 중 하위 k 비트를 제외하여 상위 $n-k$ 비트로 비트 스케일링된 가중치를 인가받아 누산하고, 이전 비트 스케일링된 가중치가 누산된 막전위값을 가산하여 막전위값을 획득하며, 획득된 막전위값이 비트 스케일링된 가중치에 대응하여 비트 스케일링된 발화 문턱값 이상이면 출력 스파이크를 발화하고, 발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 출력 뉴런부를 포함하는 스파이킹 신경망.

청구항 2

제1 항에 있어서, 상기 출력 뉴런부는

상기 가중치 시냅스부의 다수의 비트 라인이 n 개의 비트 라인 단위로 그룹화된 다수의 비트 라인 그룹 중 비트 라인 그룹 각각에서 상위 $n-k$ 비트에 대응하는 비트 라인을 통해 인가되는 가중치를 상기 비트 스케일링된 가중치로 획득하는 스파이킹 신경망.

청구항 3

제2 항에 있어서, 상기 출력 뉴런부는

상기 비트 스케일링된 가중치의 비트 스케일링 비율 이하로 상기 발화 문턱값을 비트 스케일링하는 스파이킹 신경망.

청구항 4

제2 항에 있어서, 상기 출력 뉴런부는

상기 다수의 비트 라인 그룹 각각에 대응하는 다수의 출력 뉴런을 포함하고,

상기 다수의 출력 뉴런 각각은

대응하는 비트 라인 그룹의 상위 $n-k$ 비트에 대응하는 비트 라인으로부터 비트 스케일링된 가중치를 획득하여 누산하고, 이전 획득된 막전위값을 가산하여 막전위 값을 획득하는 가중치 누산부;

획득된 막전위 값이 상기 비트 스케일링된 발화 문턱값 이상이면 출력 스파이크를 발화하는 발화 판별부; 및

기 지정된 기간 동안 발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 발화 문턱값 조절부를 포함하는 스파이킹 신경망.

청구항 5

제4 항에 있어서, 상기 출력 뉴런은

이전 획득된 막전위값에 기 지정된 누설 팩터를 가중하여 누설 막전위값을 획득하는 누설부를 더 포함하고,

상기 가중치 누산부는 누산된 비트 스케일링된 가중치에 상기 누설 막전위값을 가산하여 상기 막전위값을 획득하는 스파이킹 신경망.

청구항 6

제4 항에 있어서, 상기 출력 뉴런은

기지정된 억제 기간 또는 내화 기간동안 출력 스파이크를 발화하지 않도록, 상기 막전위값을 기지정된 레벨로 초기화하는 리셋 섹터부를 더 포함하는 스파이킹 신경망.

청구항 7

제4 항에 있어서, 상기 발화 문턱값 조절부는

기지정된 기간 동안 상기 출력 스파이크의 개수를 카운트하고, 카운트된 출력 스파이크의 개수가 기지정된 항상성 업 개수 이상이면, 상기 발화 문턱값을 기지정된 문턱값 변화량만큼 증가시키고, 카운트된 출력 스파이크의 개수가 기지정된 항상성 다운 개수 이하이면, 상기 발화 문턱값을 기지정된 문턱값 변화량만큼 감소시키는 스파이킹 신경망.

청구항 8

제4 항에 있어서, 상기 가중치 시냅스부는

상기 다수의 비트 라인 그룹 각각에서 상위 $n-k$ 비트에 대응하는 비트 라인 각각과 대응하는 출력 뉴런 사이에 연결되어, 대응하는 비트 라인을 통해 인가되는 비트 신호를 대응하는 출력 뉴런으로 전달하는 다수의 센스 앰프를 더 포함하는 스파이킹 신경망.

청구항 9

제2 항에 있어서, 상기 스파이킹 신경망은

상기 가중치 시냅스부에 저장된 가중치를 학습시키는 가중치 학습부를 더 포함하고,

상기 가중치 학습부는 상기 다수의 입력 스파이크가 생성된 시간 정보를 포함하는 입력 스파이크 히스토리와 상기 출력 스파이크가 발화된 시간 정보를 포함하는 출력 스파이크 히스토리를 저장하여, 입력 스파이크가 생성된 시간과 출력 스파이크가 발화된 시간 차에 기반하여 상기 가중치를 업데이트하는 스파이킹 신경망.

청구항 10

제9 항에 있어서, 상기 가중치 학습부는

비트 스케일링되지 않은 n 비트의 가중치를 업데이트하는 스파이킹 신경망.

청구항 11

입력 데이터에 응답하여 다수의 입력 스파이크를 생성하여, 다수의 워드 라인과 다수의 비트 라인에 의해 정의되는 다수의 메모리 셀을 포함하여 가중치를 저장하는 가중치 시냅스부에서 상기 다수의 워드 라인 중 대응하는 워드 라인을 활성화하는 단계;

상기 활성화된 워드 라인과 상기 다수의 비트 라인이 교차하는 영역에 배치된 다수의 메모리 셀 각각에 비트값이 저장된 n 비트의 가중치 중 하위 k 비트를 제외하여 상위 $n-k$ 비트로 비트 스케일링된 가중치를 인가받아 누산하는 단계;

누산된 비트 스케일링된 가중치에 이전 비트 스케일링된 가중치가 누산된 막전위값을 가산하여 막전위값을 획득하는 단계;

획득된 막전위값이 비트 스케일링된 가중치에 대응하여 비트 스케일링된 발화 문턱값 이상이면, 출력 스파이크를 발화하는 단계; 및

발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 단계를 포함하는 스파이킹 신경망의 운용 방법.

청구항 12

제11 항에 있어서, 상기 누산하는 단계는

상기 가중치 시냅스부의 다수의 비트 라인이 n 개의 비트 라인 단위로 그룹화된 다수의 비트 라인 그룹 중 대응

하는 비트 라인 그룹에서 상위 $n-k$ 비트에 대응하는 비트 라인을 통해 인가되는 가중치를 상기 비트 스케일링된 가중치로 획득하는 스파이킹 신경망의 운용 방법.

청구항 13

제12 항에 있어서, 상기 발화 문턱값을 조절하는 단계는

상기 비트 스케일링된 가중치의 비트 스케일링 비율 이하로 상기 발화 문턱값을 비트 스케일링하고, 비트 스케일링된 발화 문턱값을 조절하는 스파이킹 신경망의 운용 방법.

청구항 14

제13 항에 있어서, 상기 막전위값을 획득하는 단계는

이전 획득된 막전위값에 기지정된 누설 팩터를 가중하여 누설 막전위값을 획득하는 단계; 및

누산된 비트 스케일링된 가중치에 상기 누설 막전위값을 가산하여 상기 막전위값을 획득하는 단계를 포함하는 스파이킹 신경망의 운용 방법.

청구항 15

제13 항에 있어서, 상기 스파이킹 신경망의 운용 방법은

기지정된 억제 기간 또는 내화 기간동안 출력 스바이크를 발화하지 않도록, 상기 막전위값을 기지정된 레벨로 초기화하는 단계를 더 포함하는 스파이킹 신경망의 운용 방법.

청구항 16

제13 항에 있어서, 상기 발화 문턱값을 조절하는 단계는

상기 출력 스파이크의 개수를 카운트하는 단계;

카운트된 출력 스파이크의 개수가 기지정된 항상성 업 개수 이상이면, 상기 발화 문턱값을 기지정된 문턱값 변화량만큼 증가시키는 단계; 및

카운트된 출력 스파이크의 개수가 기지정된 항상성 다운 개수 이하이면, 상기 발화 문턱값을 기지정된 문턱값 변화량만큼 감소시키는 단계를 포함하는 스파이킹 신경망의 운용 방법.

청구항 17

제12 항에 있어서, 상기 스파이킹 신경망의 운용 방법은

상기 가중치 시냅스부에 저장된 가중치를 학습시키는 단계를 더 포함하고,

상기 학습시키는 단계는

상기 다수의 입력 스파이크가 생성된 시간 정보를 포함하는 입력 스파이크 히스토리와 상기 출력 스파이크가 발화된 시간 정보를 포함하는 출력 스파이크 히스토리를 저장하는 단계;

저장된 입력 스파이크 히스토리와 출력 스파이크 히스토리를 이용하여 입력 스파이크가 생성된 시간과 출력 스파이크가 발화된 시간 차를 계산하는 단계; 및

계산된 시간 차에 기반하여 상기 가중치를 업데이트하는 단계를 포함하는 스파이킹 신경망의 운용 방법.

청구항 18

제17 항에 있어서, 상기 가중치를 업데이트하는 단계는

비트 스케일링되지 않은 n 비트의 가중치를 업데이트하는 스파이킹 신경망의 운용 방법.

발명의 설명

기술 분야

본 발명은 스파이킹 신경망 및 이의 운용 방법에 관한 것으로, 집적도와 전력 효율성을 높일 수 있는 스파이킹

[0001]

신경망 및 이의 운용 방법에 관한 것이다.

배경 기술

- [0002] 뉴로모픽(Neuromorphic) 기술은 인간의 신경구조를 하드웨어적으로 모방하기 위한 기술로서, 기존 컴퓨팅 아키텍처가 인지처리 기능을 수행함에 있어 인간에 비해 효율성이 매우 낮고 전력 소모가 크다는 한계를 극복하기 위해 제안되었다.
- [0003] 뉴로모픽 기술에는 대표적으로 스파이킹 신경망(Spiking Neural Network: 이하 SNN)이 있다. SNN은 인간의 두뇌가 뉴런(Neuron)-시냅스부(Synapse) 구조를 가지고 있고, 뉴런과 뉴런을 잇는 시냅스가 스파이크 형태의 전기 신호로 정보를 전달한다는 특징을 모방하여 고안된 신경망이다. 이러한 SNN은 이진 값을 갖는 스파이크 신호가 전송되는 타이밍, 즉 시간 차에 기초하여 정보를 처리한다.
- [0004] SNN은 소프트웨어적으로 구현되어 컨볼루션 신경망(Convolution Neural Network: CNN), 순환 신경망(Recurrent Neural Network: RNN) 등으로 대표되는 심층 신경망(Deep Neural Network)에 비해 곱셈 연산을 요구하지 않아 간단한 하드웨어 구조로 구현될 수 있어 고성능의 컴퓨터 시스템을 요구하지 않고, 전력 소모가 매우 적다는 장점이 있다.
- [0005] 그러나 SNN을 이용하여 인간의 두뇌를 모방하기 위해서는 적어도 수십만에서 많게는 수조개의 뉴런 및 시냅스를 집적해야 한다. 따라서 집적도와 전력 효율성은 SNN의 구현에 있어 매우 중요한 이슈이다.
- [0006] SNN은 일반적으로 CMOS(complementary metal-oxide semiconductor) 기반의 메모리 소자 형태로 제작되었으나 최근에는 RRAM(Resistive Random Access Memory) 또는 PRAM(Phase-change Memory)과 같은 새로운 종류의 메모리 소자의 특성을 이용하고자 하는 시도가 계속되고 있다. 이와 같은 새로운 메모리 소자를 이용하는 경우, 각 소자의 물질적 특성에 따라 뉴런의 행동 특성을 모방하기 때문에 집적도를 높이고 전력 효율성을 향상시킬 수 있으나, 각 소자의 특성을 정밀하게 조절하는 것이 어렵고, 소자간 특성 편차가 크기 때문에 신뢰도가 낮다는 한계가 있다.
- [0007] 이에 CMOS 기반 디지털 회로에서 집적도와 전력 효율성을 높이기 위한 방법이 요구되고 있다.

선행기술문헌

특허문헌

- [0008] (특허문헌 0001) 한국 공개 특허 제10-2018-0062934호 (2018.06.11 공개)

발명의 내용

해결하려는 과제

- [0009] 본 발명의 목적은 집적도와 전력 효율성을 높일 수 있는 스파이킹 신경망 및 이의 운용 방법을 제공하는데 있다.
- [0010] 본 발명의 다른 목적은 집적도와 전력 효율성을 높이면서 출력되는 스파이킹 타이밍의 정확도를 유지할 수 있는 스파이킹 신경망 및 이의 운용 방법을 제공하는데 있다.

과제의 해결 수단

- [0011] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 스파이킹 신경망은 다수의 워드 라인과 다수의 비트 라인에 의해 정의되는 다수의 메모리 셀을 포함하여 가중치를 저장하는 가중치 시냅스부; 입력 데이터에 응답하여 다수의 입력 스파이크를 생성하여 상기 다수의 워드 라인 중 대응하는 워드 라인을 활성화하는 입력 뉴런부; 및 상기 활성화된 워드 라인과 상기 다수의 비트 라인이 교차하는 영역에 배치된 다수의 메모리 셀 각각에 비트값이 저장된 n비트의 가중치 중 하위 k비트를 제외하여 상위 n-k비트로 비트 스케일링된 가중치를 인가받아 누산하고, 이전 비트 스케일링된 가중치가 누산된 막전위값을 가산하여 막전위값을 획득하며, 획득된 막전위값이 비트 스케일링된 가중치에 대응하여 비트 스케일링된 발화 문턱값 이상이면 출력 스파이크를 발화하고, 발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 출력 뉴런부를 포함한다.

- [0012] 상기 출력 뉴런부는 상기 가중치 시냅스부의 다수의 비트 라인이 n 개의 비트 라인 단위로 그룹화된 다수의 비트 라인 그룹 중 비트 라인 그룹 각각에서 상위 $n-k$ 비트에 대응하는 비트 라인을 통해 인가되는 가중치를 상기 비트 스케일링된 가중치로 획득할 수 있다.
- [0013] 상기 출력 뉴런부는 상기 비트 스케일링된 가중치의 비트 스케일링 비율 이하로 상기 발화 문턱값을 비트 스케일링할 수 있다.
- [0014] 상기 출력 뉴런부는 상기 다수의 비트 라인 그룹 각각에 대응하는 다수의 출력 뉴런을 포함하고, 상기 다수의 출력 뉴런 각각은 대응하는 비트 라인 그룹의 상위 $n-k$ 비트에 대응하는 비트 라인으로부터 비트 스케일링된 가중치를 획득하여 누산하고, 이전 획득된 막전위값을 가산하여 막전위 값을 획득하는 가중치 누산부; 획득된 막전위 값이 상기 비트 스케일링된 발화 문턱값 이상이면 출력 스파이크를 발화하는 발화 판별부; 및 기지정된 기간 동안 발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 발화 문턱값 조절부를 포함할 수 있다.
- [0015] 상기 출력 뉴런은 이전 획득된 막전위값에 기지정된 누설 팩터를 가중하여 누설 막전위값을 획득하는 누설부를 더 포함하고, 상기 가중치 누산부는 누산된 비트 스케일링된 가중치에 상기 누설 막전위값을 가산하여 상기 막전위값을 획득할 수 있다.
- [0016] 상기 출력 뉴런은 기지정된 억제 기간 또는 내화 기간동안 출력 스파이크를 발화하지 않도록, 상기 막전위값을 기지정된 레벨로 초기화하는 리셋 섹터부를 더 포함할 수 있다.
- [0017] 상기 발화 문턱값 조절부는 기지정된 기간 동안 상기 출력 스파이크의 개수를 카운트하고, 카운트된 출력 스파이크의 개수가 기지정된 항상성 업 개수 이상이면, 상기 발화 문턱값을 기지정된 문턱값 변화량만큼 증가시키고, 카운트된 출력 스파이크의 개수가 기지정된 항상성 다운 개수 이하이면, 상기 발화 문턱값을 기지정된 문턱값 변화량만큼 감소시킬 수 있다.
- [0018] 상기 가중치 시냅스부는 상기 다수의 비트 라인 그룹 각각에서 상위 $n-k$ 비트에 대응하는 비트 라인 각각과 대응하는 출력 뉴런 사이에 연결되어, 대응하는 비트 라인을 통해 인가되는 비트 신호를 대응하는 출력 뉴런으로 전달하는 다수의 센스 앰프를 더 포함할 수 있다.
- [0019] 상기 스파이킹 신경망은 상기 가중치 시냅스부에 저장된 가중치를 학습시키는 가중치 학습부를 더 포함하고, 상기 가중치 학습부는 상기 다수의 입력 스파이크가 생성된 시간 정보를 포함하는 입력 스파이크 히스토리와 상기 출력 스파이크가 발화된 시간 정보를 포함하는 출력 스파이크 히스토리를 저장하여, 입력 스파이크가 생성된 시간과 출력 스파이크가 발화된 시간 차에 기반하여 상기 가중치를 업데이트할 수 있다.
- [0020] 상기 가중치 학습부는 비트 스케일링되지 않은 n 비트의 가중치를 업데이트할 수 있다.
- [0021] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 스파이킹 신경망의 운용 방법은 입력 데이터에 응답하여 다수의 입력 스파이크를 생성하여, 다수의 워드 라인과 다수의 비트 라인에 의해 정의되는 다수의 메모리 셀을 포함하여 가중치를 저장하는 가중치 시냅스부에서 상기 다수의 워드 라인 중 대응하는 워드 라인을 활성화하는 단계; 상기 활성화된 워드 라인과 상기 다수의 비트 라인이 교차하는 영역에 배치된 다수의 메모리 셀 각각에 비트값이 저장된 n 비트의 가중치 중 하위 k 비트를 제외하여 상위 $n-k$ 비트로 비트 스케일링된 가중치를 인가받아 누산하는 단계; 누산된 비트 스케일링된 가중치에 이전 비트 스케일링된 가중치가 누산된 막전위값을 가산하여 막전위값을 획득하는 단계; 획득된 막전위값이 비트 스케일링된 가중치에 대응하여 비트 스케일링된 발화 문턱값 이상이면, 출력 스파이크를 발화하는 단계; 및 발화된 출력 스파이크의 개수를 카운트하여 비트 스케일링된 발화 문턱값을 조절하는 단계를 포함한다.

발명의 효과

- [0022] 따라서, 본 발명의 실시예에 따른 스파이킹 신경망 및 이의 운용 방법은 생체 뉴런의 항상성에 기초하여 스파이킹 신경망에서 출력되는 스파이크의 개수에 따라 발화 문턱값을 가변하고, 가변되는 발화 문턱값에 따라 가중치의 비트 스케일을 가변함으로써, 스파이킹 타이밍의 정확도를 유지하면서도 집적도와 전력 효율성을 크게 높일 수 있는 스파이킹 신경망 및 이의 운용 방법을 제공하는데 있다.

도면의 간단한 설명

- [0023] 도 1은 스파이킹 신경망의 개략적 구조를 나타낸다.

도 2는 도 1의 출력 뉴런의 구성의 일예를 나타낸다.

도 3 및 도 4는 도 1의 가중치 시냅스와 출력 뉴런의 동작을 설명하기 위한 도면이다.

도 5는 발화 문턱값 조절부에 의해 가변되는 발화 문턱값의 일예를 나타낸다.

도 6은 본 실시예에 따른 SNN이 가중치를 비트 스케일링하는 개념을 나타낸다.

도 7은 가중치 비트 스케일링에 따른 발화 문턱값의 변화를 나타낸다.

도 8은 가중치 비트 스케일링에 의한 학습 결과를 비교하여 나타낸다.

도 9는 가중치 비트 스케일링 비트 수에 따른 학습 정확도를 나타낸다.

도 10은 가중치 비트 스케일링 비트 수에 따른 전력 소비량 변화를 나타낸다.

도 11은 본 발명의 일 실시예에 따른 스파이킹 신경망의 운용 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

- [0024] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.
- [0025] 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 부재를 나타낸다.
- [0026] 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.
- [0027] 도 1은 스파이킹 신경망의 개략적 구조를 나타내고, 도 2는 도 1의 출력 뉴런의 구성의 일예를 나타내며, 도 3 및 도 4는 도 1의 가중치 시냅스와 출력 뉴런의 동작을 설명하기 위한 도면이다. 그리고 도 5는 발화 문턱값 조절부에 의해 가변되는 발화 문턱값의 일예를 나타낸다.
- [0028] 도 1을 참조하면, SNN은 입력 뉴런부(100)와 가중치 시냅스부(200), 출력 뉴런부(300) 및 가중치 학습부(400)를 포함한다.
- [0029] 입력 뉴런부(100)는 입력 데이터를 기지정된 방식에 따라 다수의 입력 스파이크로 변환하고, 변환된 다수의 입력 스파이크를 가중치 시냅스부(200)로 전달한다. 인간의 두뇌는 각종 감각 기관에서 감지된 정보를 전기적 신호의 형태로 전달받고, 뉴런과 뉴런을 잇는 시냅스가 스파이크 형태의 전기 신호로 뉴런 사이에 정보가 전달되도록 한다. 이에 인간의 두뇌를 모방하도록 구성되는 SNN의 입력 뉴런부(100) 또한 입력 데이터를 기지정된 방식에 따라 입력 스파이크로 변환하여 가중치 시냅스부(200)로 전달한다.
- [0030] 입력 뉴런부(100)는 다수의 입력 뉴런으로 구성되어, 각각 대응하는 입력 데이터의 값에 따라 시간차를 두고 발생하는 입력 스파이크로 변환할 수 있다. 일례로 입력 뉴런부(100)의 다수의 입력 뉴런은 이미지에서 대응하는 각 픽셀의 픽셀값을 입력 데이터로 인가받고, 인가된 픽셀값에 따라 시간차를 두고 발생하는 입력 스파이크를 출력할 수 있다.
- [0031] 입력 뉴런부(100)의 입력 뉴런이 입력 데이터를 입력 스파이크로 변환하는 방법은 다양하게 연구되어 있으며, 일례로 포아송 분포(Poisson distribution)에 기반하여 입력 데이터를 입력 스파이크로 변환할 수 있다. 즉 입력 데이터에 대해 포아송 분포에 기반한 확률에 근거하여 다수의 스파이크를 발생하거나 발생하지 않도록 하여 입력 스파이크를 생성한다. 이때 입력 뉴런부(100)의 입력 뉴런은 상기한 바와 같이, 입력 데이터를 한번의 입력 스파이크들로 변환하는 것이 아니라, 입력 데이터에 따른 확률에 기초하여 다수의 입력 스파이크를 시간차를 두고 여러 번 생성하여 출력할 수 있다.
- [0032] 즉 도 3에 도시된 바와 같이, 각각의 입력 데이터에 대해 다수의 입력 스파이크를 서로 다른 시간에 여러 번 생성할 수 있다.

- [0033] 여기서 입력 뉴런부(100)는 이진 데이터의 형식으로 입력 스파이크를 생성한다. 일례로 입력 스파이크가 발생되면 1의 값을 출력하고, 입력 스파이크가 발생되지 않으면 0의 값을 출력하도록 설정될 수 있다.
- [0034] 가중치 시냅스부(200)는 입력 뉴런부(100)에서 인가되는 입력 스파이크에 가중치를 가중하여 출력 뉴런부(300)로 전달한다. 가중치 시냅스부(200)에는 다수의 가중치가 저장되며, 저장된 다수의 가중치는 가중치 학습부(400)에 의해 업데이트된다.
- [0035] 가중치 시냅스부(200)는 SNN에서 두뇌의 시냅스를 모방한 구성으로, 다수의 가중치를 저장하는 메모리 어레이와 유사한 구조를 가질 수 있다. 일례로 가중치 시냅스부(200)는 SRAM(Static Random access memory) 등으로 구현될 수 있다.
- [0036] 가중치 시냅스부(200)는 입력 뉴런부(100) 중 대응하는 입력 뉴런에서 입력 스파이크가 인가되면 활성화되는 다수의 워드 라인(WL)과 다수의 워드 라인(WL)에 교차하는 다수의 비트 라인(BL) 및 다수의 워드 라인(WL)과 다수의 비트 라인(BL)에 의해 정의되는 다수의 메모리 셀(MC)을 포함하여 구성될 수 있다. 그리고 다수의 메모리 셀(MC)은 각각 가중치에서 대응하는 비트값을 저장한다.
- [0037] 여기서 워드 라인(WL)은 입력 뉴런의 개수에 대응하는 개수로 배치될 수 있으며, 비트 라인(BL)은 가중치의 비트 수에 대응하는 개수 단위의 비트 라인 그룹(BLG)으로 배치되며, 비트 라인 그룹(BLG)은 후술하는 출력 뉴런부(300)의 출력 뉴런의 개수에 대응하는 개수로 배치될 수 있다. 즉 출력 뉴런부(300)가 병렬로 출력할 수 있는 출력 스파이크의 수에 대응하는 개수의 비트 라인 그룹(BLG)이 배치된다. 예를 들어 가중치가 n비트의 이진 값으로 정의되고 출력 뉴런부(300)의 출력 뉴런의 개수가 m개이면, n개의 비트 라인(BL)이 비트 라인 그룹(BLG)을 구성하고, m개의 비트 라인 그룹(BLG)이 배치되어야 하므로, 가중치 시냅스부(200)는 $n * m$ 개의 비트 라인(BL)을 포함할 수 있다. 그리고 각각의 워드 라인(WL)과 비트 라인 그룹(BLG)이 교차하는 위치의 메모리 셀에는 상기한 바와 같이, 가중치의 대응하는 비트값이 저장된다. 만일 입력 뉴런부(100)가 N개의 입력 뉴런을 구비하여 각각 입력 데이터에 응답하여 입력 스파이크($x_1 \sim x_N$)를 출력하는 경우, 가중치 시냅스부(200)는 $N * n * m$ 개의 메모리 셀(MC)을 포함할 수 있다.
- [0038] 가중치 시냅스부(200)는 특정 입력 뉴런에서 입력 스파이크가 발생되면, 대응하는 워드 라인(WL)이 활성화되고, 활성화된 워드 라인(WL)과 다수의 비트 라인(BL)이 교차하는 위치의 메모리 셀(MC)에 저장된 비트값을 출력 뉴런부(300)로 출력한다. 즉 인가된 입력 스파이크에 대응하는 가중치를 출력 뉴런부(300)로 출력하도록 구성될 수 있다.
- [0039] 가중치 시냅스부(200)는 입력 뉴런부(100)로부터 시간차를 두고 발생되어 인가되는 다수의 입력 스파이크 각각에 응답하여 대응하는 가중치를 반복적으로 출력 뉴런부(300)로 출력한다.
- [0040] 출력 뉴런부(300)는 가중치 시냅스부(200)에서 출력되는 가중치의 누적 합에 기초하여 다수의 출력 스파이크를 발화한다. 출력 뉴런부(300)는 다수의 출력 뉴런을 포함할 수 있으며, 다수의 출력 뉴런 각각은 LIF(Leaky-Integrate and Fire) 뉴런 모델에 기반하여 구성될 수 있다.
- [0041] SNN에서는 주로 인간 신경망의 동작 특성을 모방한 LIF 뉴런 모델이 이용되며, LIF 뉴런 모델은 입력되는 입력 스파이크에 따른 가중치를 누적(Integrate)하고, 누적된 가중치에 의해 획득되는 막전위값이 기지정된 기준 레벨 이상이면 출력 스파이크를 발화(Fire)하고, 막전위값은 시간이 흐를수록 누설(Leaky)되어 점차 약해지도록 구성된 모델이다.
- [0042] 이에 LIF 뉴런 모델에 기반하여 구성되는 출력 뉴런부(300)의 다수의 출력 뉴런 각각 또한 가중치 시냅스부(200)로부터 인가되는 가중치를 누적합하여 막전위값을 획득하고, 획득된 막전위값이 기지정된 발화 문턱값 이상이면, 출력 스파이크를 발화하도록 구성된다. 그리고 출력 뉴런은 시간의 흐름에 따라 막전위값을 점차 감소시키도록 구성될 수 있다.
- [0043] 도 2는 출력 뉴런부(300)의 다수의 출력 뉴런 중 하나의 구성을 상세하게 나타낸 도면으로, 출력 뉴런은 가중치 누산부(310), 누설부(320), 리셋 선택부(330), 발화 판별부(340), 스파이크 카운터(350), 업 다운 판별부(360) 및 발화 문턱값 발생부(370)를 포함할 수 있다.
- [0044] 가중치 누산부(310)는 도 3에 도시된 바와 같이, 입력 뉴런부(100)의 다수의 입력 뉴런에서 발생하는 다수의 입력 스파이크 각각에 의해 활성화되는 워드 라인(WL)에 응답하여 가중치 시냅스부(200)로부터 인가되는 모든 가중치(w_i)와 이전 획득된 막전위값($V(t-1)$)을 가산한다. 도 3에 도시된 바와 같이, 가중치 시냅스부(200)는 다

수의 입력 스파이크($x_1 \sim x_N$)가 인가되면, 입력된 입력 스파이크($x_1 \sim x_N$) 각각에 대응하는 가중치($w_1 \sim w_N$)를 획득하고, 가중치가 가중된 가중 입력 스파이크($x_i w_i$)를 모두 가산하며, 가산된 가중 입력 스파이크($x_i w_i$)에 이전 획득된 막전위값($V(t-1)$)을 가산한다.

[0045] 상기한 바와 같이, 이진값의 형태의 입력 스파이크($x_1 \sim x_N$)가 대응하는 위드 라인(WL)을 활성화하고, 가중치 시냅스부(200)는 활성화된 위드 라인(WL)에 대응하는 가중치를 출력 뉴런으로 전달하므로, 가중 입력 스파이크($x_i w_i$)의 합($\sum x_i w_i$)은 실제로 가중치의 합으로 볼 수 있다.

[0046] 한편 가중치 누산부(310)는 입력 스파이크의 합($\sum x_i w_i$)에 이전 획득된 막전위값($V(t-1)$)을 그대로 더하는 것이 아니라 누설부(320)가 이전 획득된 막전위값($V(t-1)$)에 누설 팩터(λ)(여기서 $0 < \lambda < 1$ 인 실수)를 가중한 누설 막전위값($\lambda V(t-1)$)을 더하여 막전위값($V(t)$)을 획득할 수 있다.

[0047] 누설부(320)가 이전 막전위값($V(t-1)$)에 누설 팩터(λ)를 가중하는 것은 두뇌의 신경 세포가 시간이 지날수록 기억하고 있는 정보를 점차로 소실하는 특성을 반영한 것이며, 가중치 누산부(310)가 입력 스파이크의 합($\sum x_i w_i$)에 이전 누설 막전위값($\lambda V(t-1)$)을 더하는 것은 반복적인 자극을 통해 정보가 기억되는 특성을 반영한 것이다. 즉 누설부(320)와 가중치 누산부(310)는 각각 LIF 뉴런 모델에서 누설(Leaky)과 누적(Integrate)의 기능을 담당한다.

[0048] 그리고 리셋 선택부(330)는 가중치 누산부(310)에서 획득된 막전위값($V(t)$) 또는 기지정된 레벨을 갖는 리셋 전압(V_{rst}) 중 하나를 선택하여 발화 판별부(340)로 전달한다. 여기서 리셋 전압(V_{rst})은 막전위값($V(t)$)을 초기화하기 위한 전압으로 일예로 0V로 설정될 수 있다. 일예로 리셋 선택부(330)는 출력 스파이크를 발화한 출력 뉴런이 기지정된 내화 기간(refract period) 동안 발화하지 않도록 하는 내화 신호 또는 출력 스파이크가 발화되지 않은 출력 뉴런이 기지정된 억제 기간(inhibition period) 동안 발화하지 않도록 하는 억제 신호에 응답하여 막전위값($V(t)$)을 초기화하는 리셋 전압(V_{rst})을 선택할 수 있다.

[0049] 발화 판별부(340)는 리셋 선택부(330)에서 인가되는 막전위값($V(t)$) 또는 리셋 전압(V_{rst})과 발화 문턱값 발생부(370)에서 인가되는 발화 문턱값(V_{th})을 비교하고, 비교 결과에 따라 출력 스파이크를 발화한다.

[0050] 발화 판별부(340)는 리셋 선택부(330)에서 리셋 전압(V_{rst})이 인가되면, 출력 스파이크를 출력하지 않는다. 그러나 가중치 누산부(310)에서 출력되는 막전위값($V(t)$)은 가중 입력 스파이크($x_i w_i$)의 합($\sum x_i w_i$)에 이전 누설 막전위값($\lambda V(t-1)$)을 더한 신호로서 발화 문턱값(V_{th})보다 높거나 낮을 수 있다.

[0051] 발화 판별부(340)는 도 4에 도시된 바와 같이 막전위값($V(t)$)이 발화 문턱값(V_{th})보다 전압 레벨이 낮으면, 출력 스파이크를 발화하지 않고, 막전위값($V(t)$)이 발화 문턱값(V_{th})보다 전압 레벨이 높으면, 출력 스파이크를 발화한다.

[0052] 스파이크 카운터(350)와 업 다운 판별부(360) 및 발화 문턱값 발생부(370)는 출력 입력 스파이크의 개수에 따라 발화 문턱값(V_{th})의 전압 레벨을 조절하는 발화 문턱값 조절부를 구성하며, 발화 문턱값 조절부는 생체 뉴런의 항상성(Homeostasis)과 같은 생물학적 타당성(biologically plausible)을 구현하기 위해 추가되는 구성이다. 항상성은 SNN의 학습 시에 출력 뉴런의 민감도(excitability)를 기지정된 레벨로 유지시켜, 비지도 학습에서 과적합(overfitting) 등의 문제를 방지하여 안정성을 보장하고 정확도를 높이기 위해 적용된다.

[0053] 스파이크 카운터(350)는 기지정된 항상성 기간(homeostasis epoch)동안 발화 판별부(340)에서 출력되는 출력 스파이크의 개수를 카운트하여 카운팅값(cnt)을 출력한다. 그리고 스파이크 카운터(350)는 항상성 기간마다 리셋 신호(rst)에 응답하여 카운팅값(cnt)을 초기화한다.

[0054] 스파이크 카운터(350)가 카운팅값(cnt)을 출력하면, 업 다운 판별부(360)는 카운팅값(cnt)이 항상성 업 문턱값(H_{th_up})보다 크거나 항상성 다운 문턱값(H_{th_dn})보다 작은지 판별하고, 판별 결과 따라 발화 문턱값(V_{th})을 증가시키기 위한 항상성 업 신호(H_{up})를 출력하거나, 발화 문턱값(V_{th})을 감소시키기 위한 항상성 다운 신호(H_{dn})를 출력한다.

[0055] 업 다운 판별부(360)는 도 2에 도시된 바와 같이, 항상성 업 판별부(361)와 항상성 다운 판별부(362)를 포함할

수 있다. 항상성 업 판별부(361)는 카운팅값(cnt)을 기지정된 항상성 업 문턱값(H_{th_up})과 비교하여 카운팅값(cnt)이 항상성 업 문턱값(H_{th_up})보다 크면, 발화 문턱값(V_{th})을 증가시키기 위한 항상성 업 신호(H_{up})를 출력한다. 그리고 항상성 다운 판별부(362)는 카운팅값(cnt)을 기지정된 항상성 다운 문턱값(H_{th_dn})과 비교하여 카운팅값(cnt)이 항상성 다운 문턱값(H_{th_dn})보다 작으면, 발화 문턱값(V_{th})을 감소시키기 위한 항상성 다운 신호(H_{dn})를 출력한다. 여기서 항상성 업 문턱값(H_{th_up})과 항상성 다운 문턱값(H_{th_dn})은 미리 설정된 항상성 문턱값(H_T)을 기준으로 기지정된 레벨 큰 값과 기지정된 레벨 작은 값일 수 있다.

[0056] 발화 문턱값 발생부(370)는 업 다운 판별부(360)에서 인가되는 항상성 업 신호(H_{up}) 또는 항상성 다운 신호(H_{dn})에 응답하여 문턱값 변화량(ΔV_{th})을 결정하고, 결정된 문턱값 변화량(ΔV_{th})을 이전 발화 문턱값(V_{th})에 더하여 발화 문턱값($V_{th} = V_{th} + \Delta V_{th}$)을 가변하여 발화 판별부(340)로 출력한다.

[0057] 일례로 발화 문턱값 발생부(370)는 항상성 업 신호(H_{up})에 응답하여 양의 기지정된 크기의 문턱값 변화량(ΔV_{th})만큼 발화 문턱값(V_{th})을 증가시킬 수 있으며, 항상성 다운 신호(H_{dn})에 응답하여 음의 기지정된 크기의 문턱값 변화량(ΔV_{th})만큼 발화 문턱값(V_{th})을 감소시킬 수 있다. 즉 발화 문턱값 조절부는 도 5의 (a)에 도시된 바와 같이, 항상성 기간동안 발화 판별부(340)에서 발화되는 출력 스파이크의 개수에 따라 발화 문턱값(V_{th})을 증가시키거나 감소시킴으로써, 동일한 항상성 기간동안 발화되는 출력 스파이크의 개수가 가급적 일정하게 되도록 한다.

[0058] 이때, 항상성 업 신호(H_{up})와 항상성 다운 신호(H_{dn})에 따른 문턱값 변화량(ΔV_{th})은 동일한 크기를 가질 수 있으나, 서로 상이하게 지정될 수도 있다.

[0059] 상기한 바와 같이, 발화 문턱값 조절부는 항상성 기간 동안 카운트되는 출력 스파이크의 개수인 카운팅값(cnt)을 항상성 문턱값(H_T)을 기준으로 설정된 항상성 업 문턱값(H_{th_up}) 및 항상성 다운 문턱값(H_{th_dn})과 비교하여, 항상성 업 신호(H_{up}) 또는 항상성 다운 신호(H_{dn})를 출력하고, 항상성 업 신호(H_{up}) 또는 항상성 다운 신호(H_{dn})에 응답하여 발화 문턱값(V_{th})을 다양하게 조절할 수 있다. 그리고 발화 문턱값(V_{th})을 다양하게 조절할 수 있으므로 가중치의 비트 스케일링에 의한 오차가 발생하는 것을 최대한 억제할 수 있을 뿐만 아니라 과적합이 발생하는 것을 방지할 수 있어 SNN이 다양한 입력 데이터에 대해 정확하게 학습을 수행할 수 있도록 한다.

[0060] 여기서 항상성 기간은 학습 시에 동일 입력 데이터가 반복 입력되는 횟수를 기준으로 설정될 수 있으며, 도 5의 (a)에서는 일례로 입력 데이터가 500회 반복 입력되고 2개의 입력 데이터가 입력되는 기간을 항상성 기간으로 설정한 경우를 도시하였다. 항상성 기간은 다양하게 설정될 수 있으며, 일례로 50개의 입력 데이터 각각에 대해 500회 반복 입력되는 2500 사이클 기간으로 설정될 수 있다.

[0061] 도 5의 (b)에서는 SNN이 학습되는 학습 기간 동안 발화 문턱값(V_{th})의 변화를 나타낸다. 여기서는 SNN이 디지털 CMOS 회로를 기반으로 구현되는 것으로 가정하였으며, 이에 발화 문턱값(V_{th}) 또한 전압 레벨이 아닌 디지털 값으로 표현하였다. 도 5의 (b)에 도시된 바와 같이, 발화 문턱값 조절부에 의해 발화 문턱값(V_{th})은 매우 큰 폭의 다양한 값을 가질 수 있다. 여기서 발화 문턱값 조절부가 발화 문턱값(V_{th})을 출력 스파이크가 발화되거나 발화되지 않을 때마다 또는 항상성 기간동안 발화된 출력 스파이크의 개수에 비례하여 발화 문턱값(V_{th})을 증가 또는 감소시키지 않고, 발화 문턱값(V_{th})을 일정한 문턱값 변화량(ΔV_{th})만큼 증가시키거나 감소시키는 것은 발화 문턱값 조절부의 구현이 용이하도록 하기 위함이다.

[0062] 상기에서는 스파이크 카운터(350)가 카운트된 출력 스파이크의 개수에 대응하는 카운팅값(cnt)을 출력하고, 업 다운 판별부(360)의 항상성 업 판별부(361)와 항상성 다운 판별부(362)가 각각 카운팅값(cnt)을 항상성 업 문턱값(H_{th_up}) 및 항상성 다운 문턱값(H_{th_dn})과 비교하여 항상성 업 신호(H_{up}) 또는 항상성 다운 신호(H_{dn})를 출력하는 것으로 설명하였으나, 스파이크 카운터(350)가 카운트한 출력 스파이크의 개수를 기지정된 항상성 업 개수 및 항상성 다운 개수와 비교하여 항상성 업 신호(H_{up}) 또는 항상성 다운 신호(H_{dn})를 출력하도록 간략하게 구성될 수도 있다.

[0063] 가중치 학습부(400)는 스파이크 타이밍 의존 가소성(Spike-timing-dependent plasticity: 이하 STDP) 기법에

따라 가중치 시냅스부(200)의 다수의 가중치를 학습시킨다. STDP 기법에 따라 가중치를 학습시키는 가중치 학습부(400)는 출력 스파이크가 발화된 시점을 기준으로 이전 기지정된 기간 동안 입력 뉴런부(100)에서 입력 스파이크가 발생된 시간 정보를 기반으로 가중치를 증가시키는 한편, 입력 뉴런부(100)으로부터 입력 뉴런부(100)에서 입력 스파이크가 발생된 시점을 기준으로 이전 기지정된 기간 동안 출력 스파이크가 발화된 시간 정보를 기반으로 가중치를 감소시킬 수 있다.

[0064] 가중치 학습부(400)는 우선 출력 뉴런부(300)에서 출력 스파이크가 출력되면, 출력 스파이크가 발화되기 이전 기지정된 학습 윈도우 크기(예를 들면 60)에 대응하여 과거 기지정된 기간 동안 가중치 시냅스부(200)로 인가된 입력 스파이크의 정보를 나타내는 입력 스파이크 히스토리를 분석한다. 그리고 출력 스파이크가 발화된 시점을 기준으로 입력 스파이크 히스토리에 저장된 이전 입력 스파이크가 발생된 시간을 비교하여 시간차에 따라 가중치의 변화량을 계산하여 가중치를 증가시킨다.

[0065] 한편, 가중치 학습부(400)는 입력 뉴런부(100)에서 입력 스파이크가 발생되어 가중치 시냅스부(200)로 인가되면, 입력 스파이크가 발생되기 이전 기지정된 학습 윈도우 크기에 대응하여 과거 기지정된 기간 동안 발화된 출력 스파이크의 정보를 나타내는 출력 스파이크 히스토리를 분석한다. 그리고 입력 스파이크가 발생된 시점을 기준으로 출력 스파이크가 발화된 시간 사이의 시간차에 따라 가중치의 변화량을 계산하여 가중치를 감소시킨다.

[0066] 가중치 학습부(400)는 출력 스파이크가 발화된 시점으로부터 이전 입력 스파이크가 발생된 시점까지의 시간차가 짧으면, 입력된 입력 스파이크와 출력 스파이크 사이의 상관 관계가 높아지도록 가중치를 증가시킨다. 반대로, 입력 스파이크가 발생된 시점으로부터 이전 출력 스파이크가 발화된 시점까지의 시간차가 짧으면 입력된 입력 스파이크와 출력 스파이크 사이의 상관 관계가 낮아지도록 가중치를 감소시킨다. 이는 출력 스파이크가 이전 입력된 입력 스파이크에 대한 반응으로 발화되어야 하는 반면, 입력되는 입력 스파이크는 이전 발화된 출력 스파이크에 가능한 무관하게 발생되어야 하기 때문이다.

[0067] 즉 가중치 학습부(400)는 인가되는 입력 스파이크와 출력 스파이크의 발생 시간 차에 기반하여 가중치를 증가 또는 감소시킨다.

[0068] 가중치 학습부(400)가 STDP 기법에 기반하여 가중치 시냅스부(200)를 학습시키는 방법은 공지된 기술이므로 여기서는 상세하게 설명하지 않는다.

[0069] 한편, 상기한 바와 같이 출력 뉴런부(300)의 발화 판별부(340)는 막전위값($V(t)$)이 발화 문턱값(V_{th})보다 높으면 출력 스파이크를 발화하고, 출력 뉴런의 발화 문턱값(V_{th})이 다양하게 가변된다. 따라서 실제로는 막전위값($V(t)$)의 상위 비트의 값이 우세하게 되는 반면, 하위 비트의 값은 가변되는 발화 문턱값(V_{th})에 의해 오차가 무시될 수 있다.

[0070] 일반적으로 SNN에서는 더 많은 뉴런을 구현할수록 더 높은 정확성을 가지지만, 뉴런의 전력 소모 비중이 크기 때문에 뉴런의 집적도를 높이면서 전력 소비를 줄일 수 있는 가장 간단한 방법은 가중치의 비트 수를 줄이는 것이다. 가중치의 비트 수를 줄이게 되면, 출력 뉴런부(300)의 다수의 출력 뉴런 각각이 더 적은 비트에 대한 연산을 수행하게 됨에 따라 회로 구조가 단순해져 SNN의 집적도를 향상시킬 수 있다. 또한 가중치의 비트 수를 줄이게 되면, 가중치 시냅스부(200)는 상기 다수의 비트 라인 각각에 연결되어, 대응하는 비트 라인을 통해 전달되는 가중치의 비트값을 출력 뉴런부(300)로 전달하는 센스 앰프(미도시)의 개수를 줄여 SNN의 집적도를 더욱 높일 수 있다.

[0071] 그러나 가중치의 비트 수를 줄여 낮은 정밀도의 가중치를 이용하는 경우, 학습의 안정성이 낮아지는 문제가 있다.

[0072] 이에 본 실시예에서는 가중 입력 스파이크($x_i w_i$)의 합($\sum x_i w_i$)을 계산하는 과정에서 가중치에 대해 비트 스케일링(bit scaling)함으로써, 출력 스파이크는 정상적으로 발화하되 SNN의 집적도를 높이고 전력 소비를 줄일 수 있도록 한다.

[0073] 상기에서는 출력 뉴런부(300)가 다수의 비트 라인 그룹(BLG)에 대응하여 다수의 출력 뉴런을 포함하는 것으로 설명하였으나, 경우에 따라서는 적어도 하나의 출력 뉴런을 구비하고, 스파이크가 발생된 뉴런의 가중치만을 획득하도록 구성될 수도 있다.

[0074] 도 6은 본 실시예에 따른 SNN이 가중치를 비트 스케일링하는 개념을 나타내고, 도 7은 가중치 비트 스케일링에

따른 발화 문턱값의 변화를 나타낸다.

- [0075] 도 6에서 (a)는 가중치를 비트 스케일링 하지 않는 경우의 구성을 나타내고, (b)는 가중치를 비트 스케일링 하지 않는 경우의 구성을 나타낸다. 도 6의 (a)와 (b)를 비교하면, (a)에서는 가중치 시냅스부(200)에서 센스 앰프(SA)가 다수의 비트 라인 그룹(BLG) 각각에 대해 모든 비트 라인에 대응하여 8개씩 구비되어 가중치 누산부(310)로 8비트의 가중치를 인가하도록 구성되었다. 그에 반해, 본 실시예에 따라 가중치를 비트 스케일링하는 경우에는 (b)에 도시된 바와 같이, 비트 라인 그룹(BLG)의 비트 라인 중 가중치의 하위 비트에 대응하는 기지정된 개수(여기서는 4개)의 비트 라인에 대응하는 센스 앰프(SA)가 구비되지 않는다. 따라서 가중치 누산부(310)는 가중치의 상위 비트만을 인가받는다. 즉 (a)에서는 8비트의 가중치가 가중치 누산부(310)로 인가되는 반면, (b)에서는 4비트의 가중치가 가중치 누산부(310)로 인가된다.
- [0076] 따라서 (a)와 (b)의 가중치 누산부(310)에서 계산되는 막전위값($V(t)$)의 값이 서로 상이하게 출력된다. 그러나 상기한 바와 같이, 막전위값($V(t)$)은 출력 스파이크를 발화하기 위한 값이며, 출력 스파이크의 발화는 막전위값($V(t)$)과 발화 문턱값(V_{th})의 비교를 통해 발생된다. 그리고 발화 문턱값(V_{th})은 발화 문턱값 조절부의 의해 가변된다.
- [0077] 따라서 가중치를 비트 스케일링하여 줄어든 비트 수에 의한 오차보다 발화 문턱값(V_{th})의 변화가 더 크면, 가중치를 비트 스케일링하여 발생하는 오차는 무시될 수 있으며, 출력 스파이크는 동일한 시점에 발화될 수 있다. 상기한 바와 같이, 가중치를 8비트에서 4비트로 4비트만큼 비트 스케일링하고, 이에 대응하여 도 7의 (a)에 도시된 16비트 레벨을 갖는 발화 문턱값(V_{th})을 (b)와 같이 12비트로 4비트만큼 비트 스케일링하면, 가중치를 비트 스케일링하여 발생하는 오차는 무시될 수 있다.
- [0078] 가중치의 비트 스케일링 비율과 발화 문턱값(V_{th})의 비트 스케일링 비율은 반드시 동일하지 않아도 무방하지만, 발화 문턱값(V_{th})의 비트 스케일링 비율을 가중치의 비트 스케일링 비율을 초과하여 증가시키는 경우, 출력 스파이크가 발화되는 시점에 변화가 발생할 수 있다. 즉 출력 스파이크에 오차가 발생될 수 있다. 따라서 발화 문턱값(V_{th})의 비트 스케일링 비율을 가중치의 비트 스케일링 비율 이내로 설정되는 것이 바람직하다.
- [0079] 이와 같이 비트 스케일링을 수행하는 경우, 출력 뉴런부(300)의 다수의 출력 뉴런 각각은 8비트의 가중치에 대한 누적 연산을 수행하지 않고, 4비트의 가중치에 대한 누적 연산을 수행하여 막전위값을 획득하고, 획득된 막전위값에 따라 출력 스파이크를 발화할 수 있다. 따라서 출력 뉴런부(300)의 다수의 출력 뉴런 각각이 비트 스케일링되어 줄어든 가중치의 비트수에 따라 회로 구조가 단순해질 수 있어 SNN의 집적도를 향상시킬 수 있다. 일례로 가중치 누산부(310) 또한 8비트의 가중치들을 누산하기 위한 16비트 가산기 대신 12비트의 누산기로 대체할 수 있다. 그리고 도 2에 도시된 출력 뉴런의 나머지 구성 요소들 또한 비트 스케일링된 발화 문턱값(V_{th})과 비트 스케일링된 가중치에 따라 더 적은 비트수를 처리하도록 구현될 수 있어 SNN의 집적도를 향상시킬 수 있다.
- [0080] 또한 (b)에 도시된 바와 같이, 가중치 시냅스부(200)의 센스 앰프(SA) 개수를 줄일 수 있다. 만일 8비트의 가중치에 대해 4비트로 비트 스케일링하는 경우, 센스 앰프의 개수를 1/2로 줄일 수 있게 된다. 따라서 SNN의 집적도를 더욱 향상시킬 수 있다.
- [0081] 한편, 전력 측면에서 살펴보면, 가중치 시냅스부(200)는 비트 라인 그룹(BLG)의 비트 라인이 미리 프리차지(precharge)되어야 하므로, 프리차지 전력(E_{PRE})이 요구된다. 그리고 하나의 입력 스파이크가 인가되면 워드 라인(WL)을 활성화해야 하며, 활성화된 워드 라인(WL)에 대응하는 가중치가 각 비트 라인에 연결된 센스 앰프(SA)로 인가되므로, 센스 앰프(SA)가 대응하는 비트 신호를 증폭하기 위한 워드라인 활성화 전력(E_{WL})이 요구된다. 또한 가중치 누산부(310)가 인가된 가중치를 누산하기 위한 전력(E_{ADD})이 요구된다.
- [0082] (a)의 경우에는 비트 라인 그룹(BLG)의 n 개의 비트 라인이 프리차지되어야 하고, n 개의 센스 앰프가 비트 신호를 증폭해야 하며, 가산기가 n 비트의 가중치를 16비트로 누산해야 하므로, M 개의 출력 뉴런에 대한 전력 소비는 $(E_{PRE} + E_{WL}) * n * M + (E_{ADD_16} * M)$ 으로 계산된다. 여기서 E_{ADD_16} 은 16비트 누산기의 전력 소비를 나타낸다.
- [0083] 반면 (b)의 경우에는 비트 라인 그룹(BLG)의 $n/2$ 개의 비트 라인이 프리차지되어야 하고, $n/2$ 개의 센스 앰프가 비트 신호를 증폭해야 하며, 가산기가 $n/2$ 비트의 가중치를 12비트로 누산해야 하므로, M 개의 출력 뉴런에 대한 전력 소비는 $(E_{PRE} + E_{WL}) * n/2 * M + (E_{ADD_12} * M)$ 으로 계산된다. 여기서 E_{ADD_12} 은 12비트 누산기의 전력 소비

를 나타낸다. 즉 전력 소비가 대략 50% 수준으로 크게 줄어들게 됨을 알 수 있다.

- [0084] 다만 도 6의 (b)에 도시된 바와 같이 비트 스케일링을 수행하는 경우에도 비트 라인 그룹(BLG)의 비트 라인의 수는 그대로 유지되어야 한다. 이는 가중치 학습부(400)가 비록 입력 스파이크와 출력 스파이크 사이의 시간 차에 따라 가중치를 업데이트하지만, 이때 업데이트되는 가중치의 변화는 매우 작은 값으로 나타난다. 그러나 가중치 학습부(400)가 업데이트하는 가중치가 비트 스케일링되면, 가중치의 변화폭이 증가됨으로 인해 학습의 안정성이 낮아질 뿐만 아니라 정상적인 학습이 수행되지 않을 수 있다. 이에 비트 라인 그룹(BLG)의 비트 라인의 수를 그대로 유지하여, 가중치 시냅스부(200)에 저장된 가중치는 비트 스케일링하지 않음으로써, 출력 뉴런에서의 가중치 비트 스케일링에도 불구하고 가중치 학습부(400)가 기존과 동일한 정밀도로 가중치를 업데이트할 수 있도록 하여 학습의 안정성을 유지할 수 있도록 한다.
- [0085] 도 8은 가중치 비트 스케일링에 의한 학습 결과를 비교하여 나타낸다.
- [0086] 도 8에서 (a)는 가중치를 비트 스케일링하지 않고 학습을 수행하여 가중치를 재구성하여 나타난 결과를 나타내고, (b)는 가중치를 비트 스케일링하여 학습을 수행하여 가중치를 재구성하여 나타난 결과를 나타낸다.
- [0087] (a)와 (b)를 비교하면 가중치 비트 스케일링 여부에 거의 영향을 받지 않고 정상적으로 학습이 수행됨을 알 수 있다.
- [0088] 도 9는 가중치 비트 스케일링 비트 수에 따른 학습 정확도를 나타내고, 도 10은 가중치 비트 스케일링 비트 수에 따른 전력 소비량 변화를 나타낸다.
- [0089] 도 9를 참조하면, 가중치의 초기값이 랜덤하게 설정되고 인코딩된 입력 스파이크가 확률에 기초하여 발생되므로, 학습이 수행되는 입력 데이터에 따라 정확도에 일부 편차가 존재하지만, 4비트까지의 비트 스케일링에 대해서는 정확도가 낮아지지 않는다는 것을 알 수 있다. 즉 가중치에 대해 비트 스케일링을 수행하더라도 출력 스파이크가 발화되는 타이밍의 오차는 비트 스케일링을 수행하지 않는 경우와 거의 동일하게 나타난다.
- [0090] 그러나 도 10에 도시된 바와 같이, 가중치에 대해 비트 스케일링을 수행하는 경우, 비트 스케일링되는 비트 수의 증가에 비례하여 SNN이 가중치에 대한 가산 연산을 수행할 때 소모되는 에너지 소비가 크게 줄어들게 됨을 알 수 있다. 특히 도 10에서 비트 스케일링을 수행하지 않는 경우와 4비트의 비트 스케일링을 수행한 경우를 비교하면 에너지 소비가 1/2 수준으로 줄어들었음을 알 수 있다.
- [0091] 도 11은 본 발명의 일 실시예에 따른 스파이킹 신경망의 운용 방법을 나타낸다.
- [0092] 도 11을 참조하여, 본 실시예에 따른 스파이킹 신경망의 운용 방법을 설명하면, 우선 입력 데이터에 대응하는 다수의 입력 스파이크를 생성한다(S10). 생성된 입력 스파이크는 다수의 가중치가 저장된 가중치 시냅스의 다수의 워드 라인 중 대응하는 워드 라인을 활성화한다.
- [0093] 그리고 활성화된 워드 라인과 교차하는 다수의 비트 라인에 의해 정의되는 메모리 셀에 저장된 가중치가 비트 라인을 통해 인가되어 획득된다. 이때, 가중치는 n비트의 이진 데이터 형태로 저장되지만, 기지정된 하위 k비트를 제외한 상위 (n-k) 비트로 비트 스케일링된 가중치를 인가받아 획득한다(S20).
- [0094] 그리고 인가되는 다수의 입력 스파이크 각각에 대응하는 다수의 비트 스케일링된 가중치를 누산한다(S30). 또한 이전 획득된 막전위값($V(t-1)$)에 기지정된 누설 팩터(λ)를 가중한 누설 막전위값($\lambda V(t-1)$)을 누산된 비트 스케일링 가중치에 더하여 막전위값($V(t)$)을 획득한다(S40).
- [0095] 막전위값($V(t)$)이 획득되면, 획득된 막전위값($V(t)$)을 비트 스케일링된 발화 문턱값(V_{th})과 비교한다(S50). 막전위값($V(t)$)이 발화 문턱값(V_{th}) 미만이면, 다시 입력 스파이크에 따른 비트 스케일링 가중치를 획득하고 누산하여 막전위값($V(t+1)$)을 획득한다.
- [0096] 그러나 막전위값($V(t)$)이 발화 문턱값(V_{th}) 이상이면, 출력 스파이크를 발화하고, 막전위값($V(t)$)을 초기화 한다(S60).
- [0097] 출력 스파이크가 발화되면, 기지정된 항상성 기간 동안 발화된 출력 스파이크의 개수를 카운트하고 카운트된 출력 스파이크의 개수에 따라 비트 스케일링된 발화 문턱값(V_{th})을 조절한다(S70). 이때, 비트 스케일링된 발화 문턱값(V_{th})은 카운트된 출력 스파이크의 개수에 무관하게 증가 또는 감소될 수 있으며, 카운트된 출력 스파이크의 개수에 비례하여 증가 또는 감소되도록 설정될 수도 있다.

[0098] 본 발명에 따른 방법은 컴퓨터에서 실행시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0099] 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.

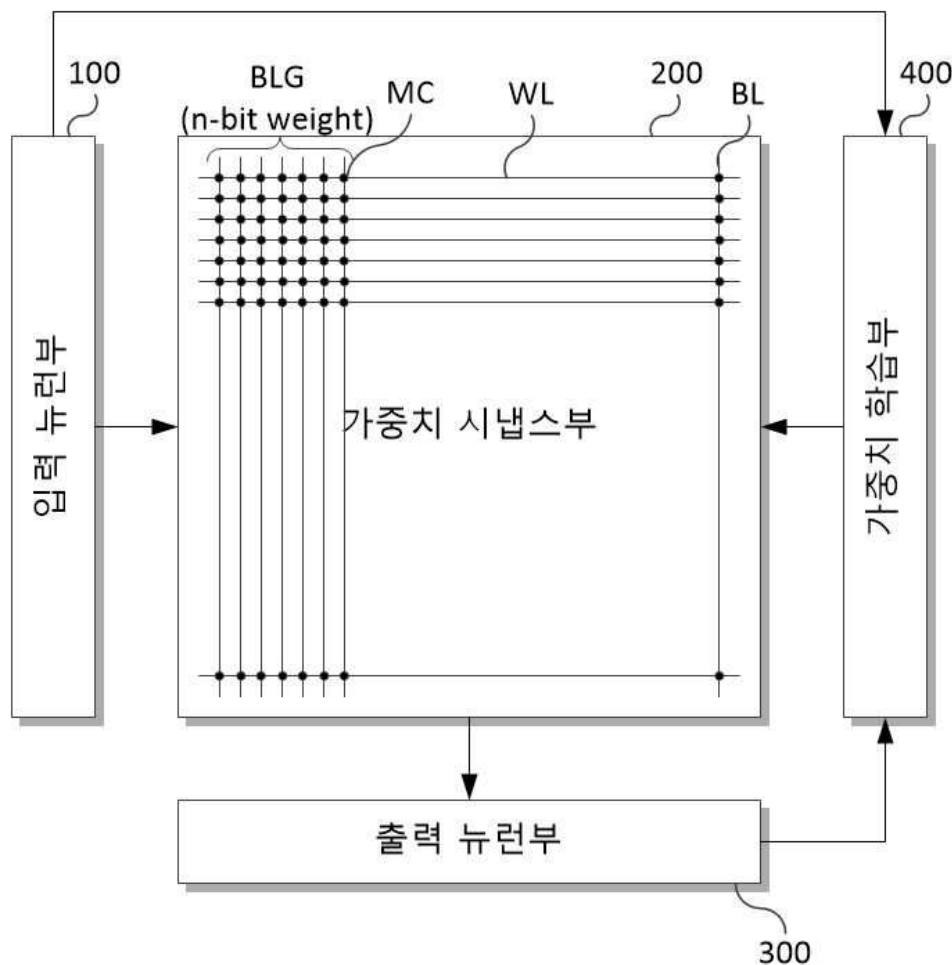
[0100] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

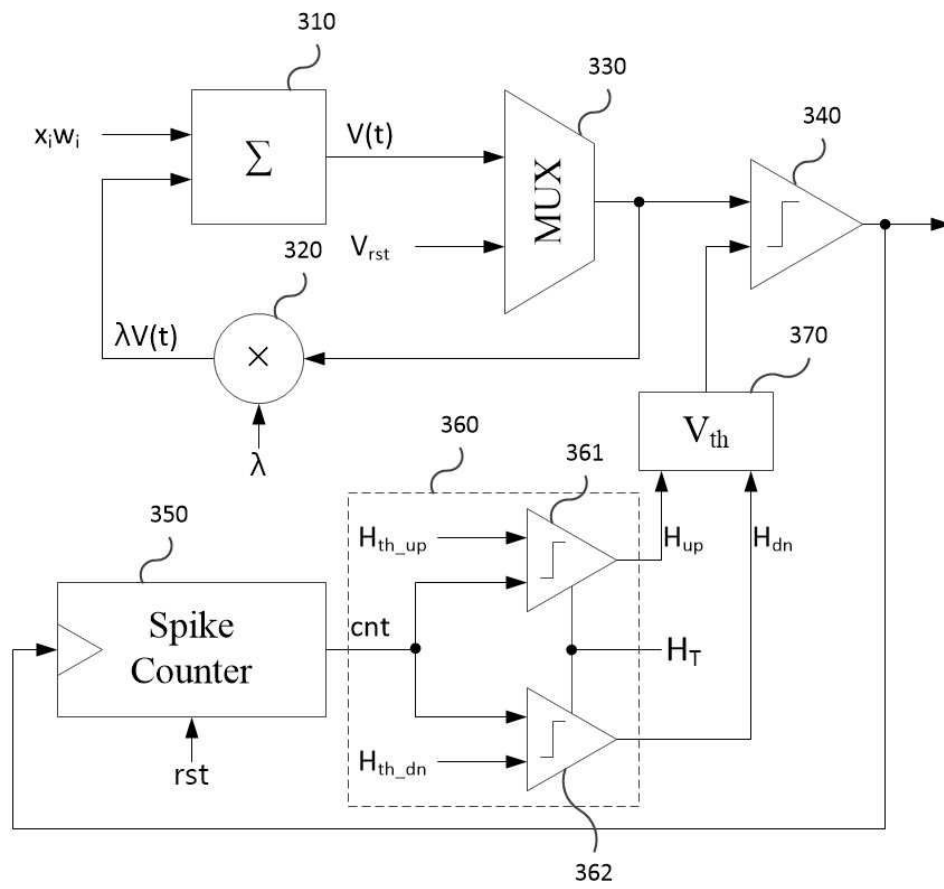
[0101] 100: 입력 뉴런부 200: 가중치 시냅스부
300: 출력 뉴런부 400: 가중치 학습부
SA: 센스 앰프 310: 가중치 누산부
320: 누설부 330: 리셋 선택부
340: 발화 판별부 350: 스파이크 카운터
360: 업 다운 판별부 370: 발화 문턱값 발생부

도면

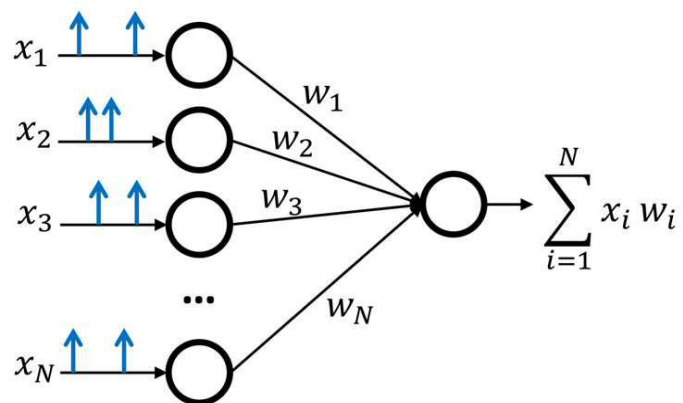
도면1



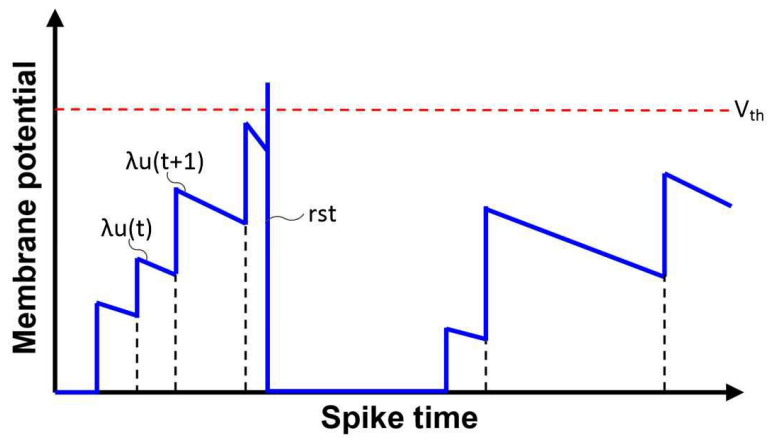
도면2



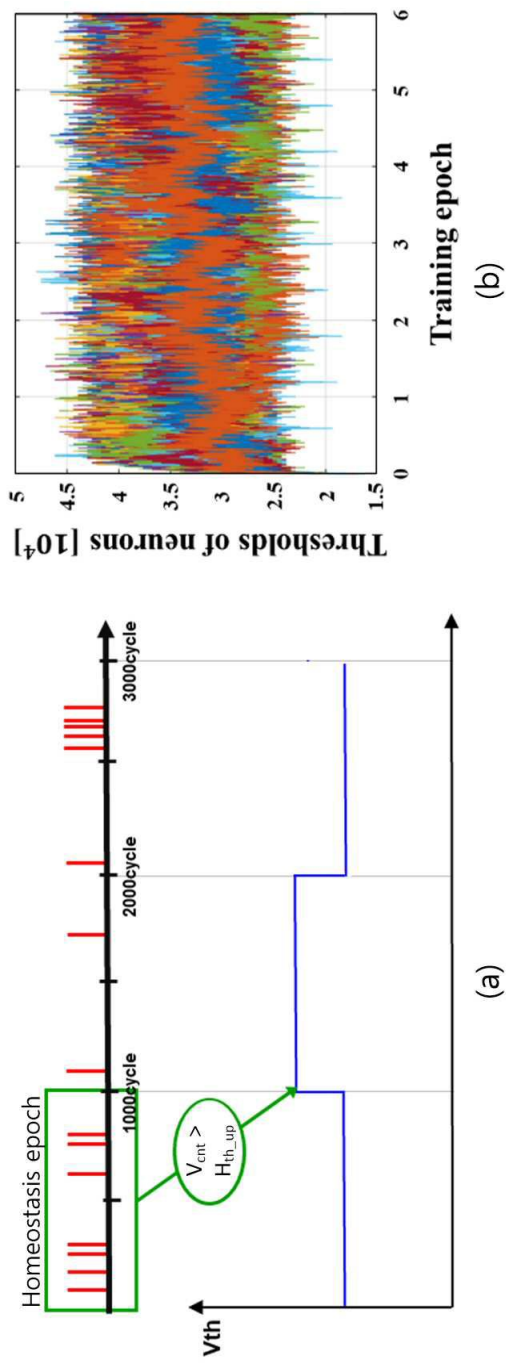
도면3



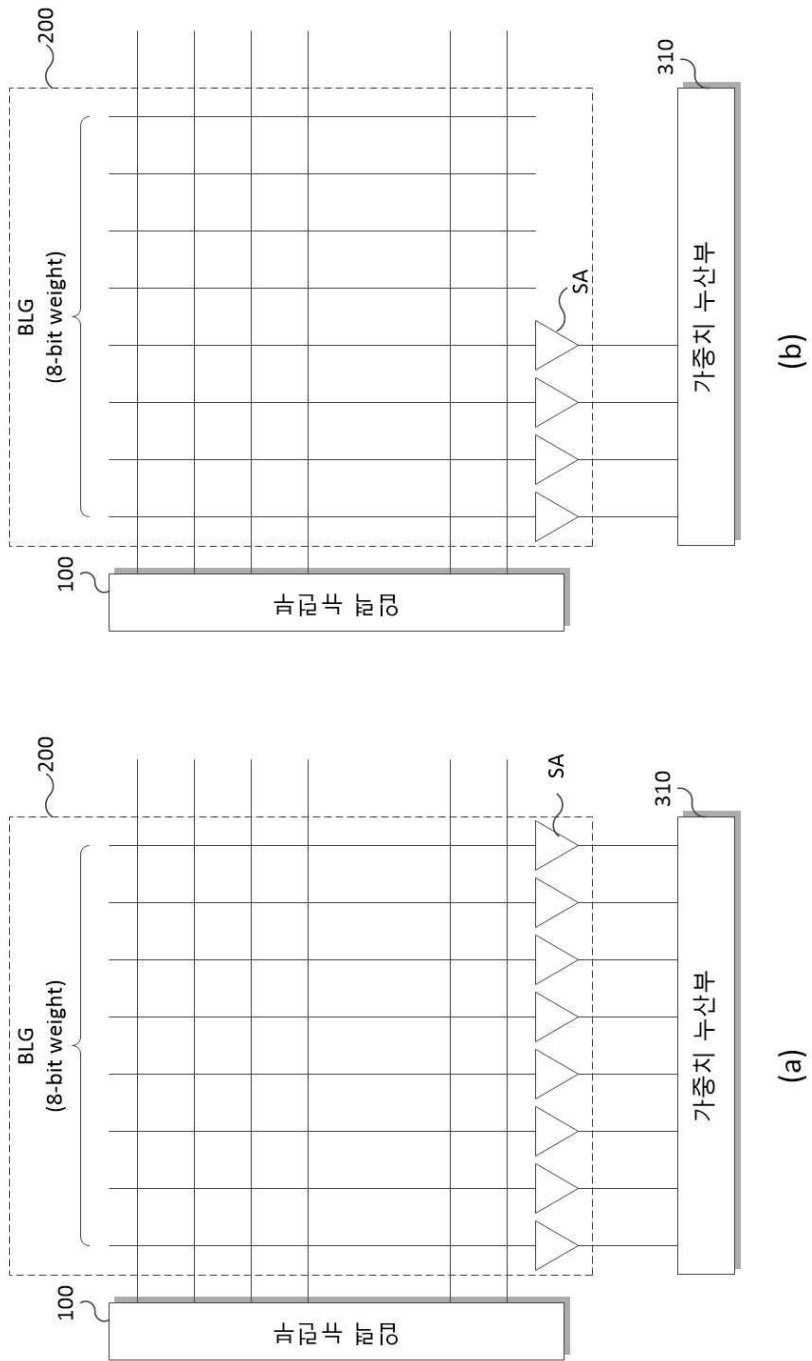
도면4



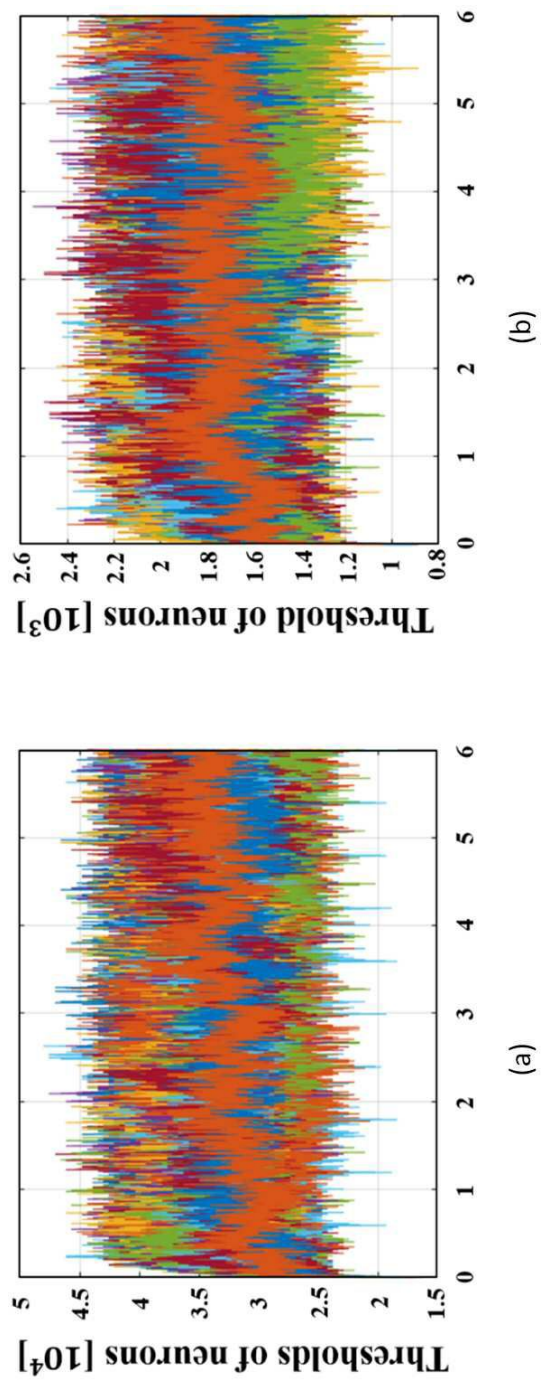
도면5



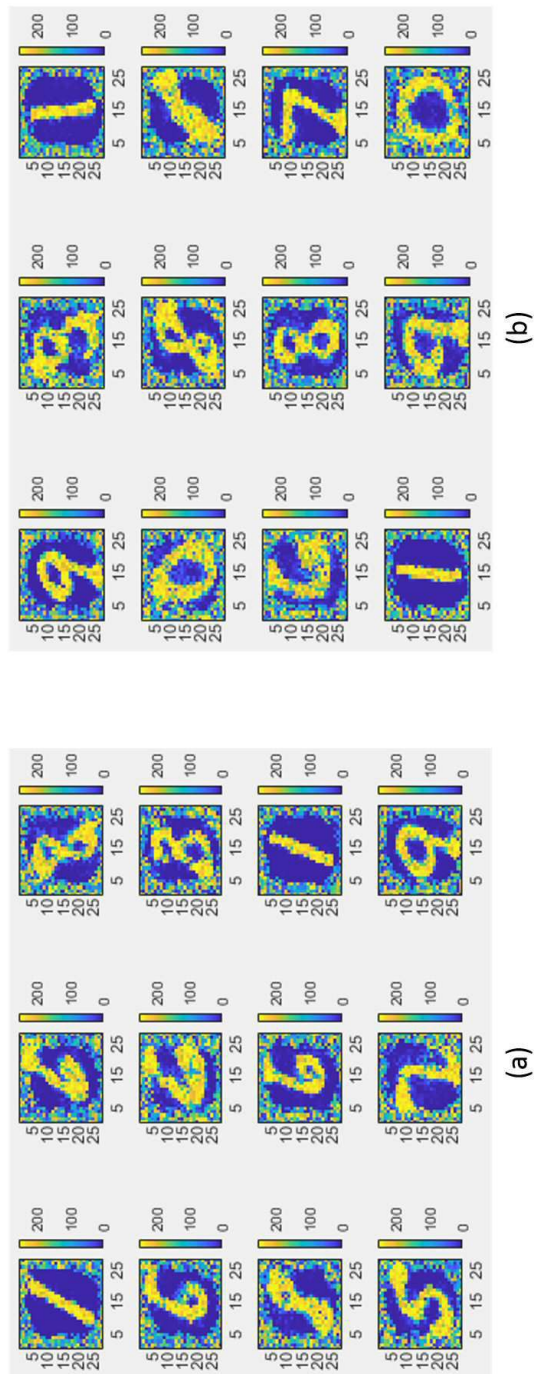
도면6



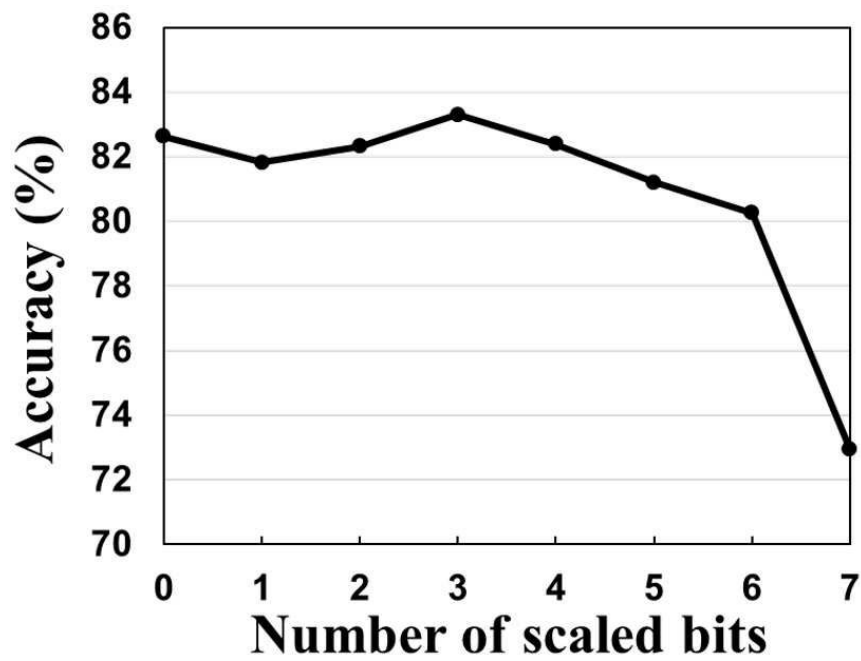
도면7



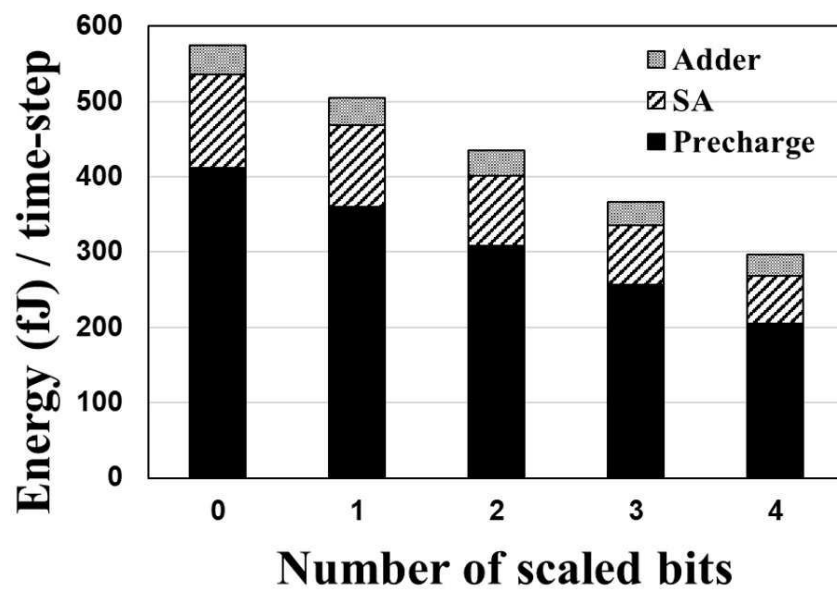
도면8



도면9



도면10



도면11

