(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0032840
(43) 공개일자 2025년03월07일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/049 (2023.01)
(21) 출원번호 10-2024-0034849
(22) 출원일자 2024년03월13일
심사청구일자 2024년03월13일
(30) 우선권주장
1020230115430 2023년08월31일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
박종선
서울특별시 서초구 신반포로 32
유동우
서울특별시 동대문구 안암로18가길 28(제기동)
(74) 대리인
특허법인주연케이알피

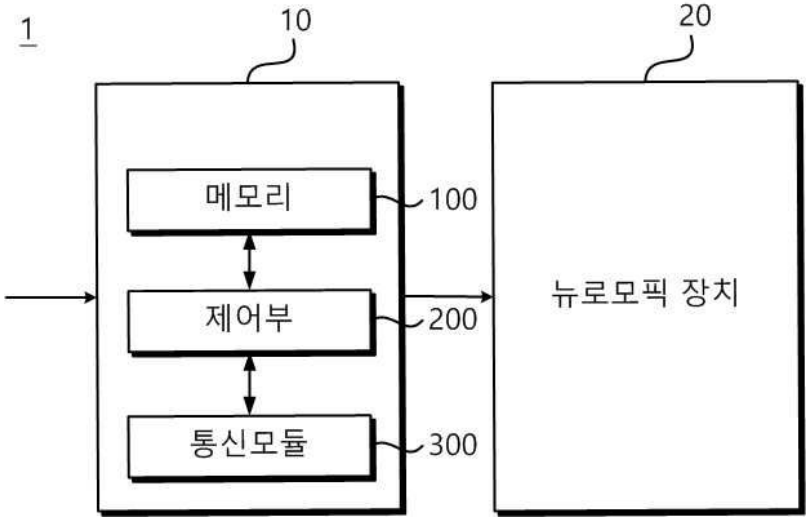
전체 청구항 수 : 총 11 항

(54) 발명의 명칭 **빈도 코딩 기반의 스파이킹 뉴럴 네트워크의 추론 가속 장치 및 그 동작 방법**

(57) 요 약

본 발명은 스파이킹 뉴럴 네트워크에 기반한 추론 가속 장치의 동작 방법에 관한 것으로, 입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택하는 단계, 상기 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단하는 단계 및 상기 활성 뉴런의 추론 연산을 수행하고, 상기 비활성 뉴런의 추론 연산을 중지하는 동적 프루닝(pruning)을 수행하는 단계를 포함하고, 상기 활성 뉴런은 막전위값이 상기 기준값 이상인 뉴런이고, 상기 비활성 뉴런은 막전위값이 상기 기준값 미만인 뉴런일 수 있다.

대 표 도 - 도3



(52) CPC특허분류
G06N 3/082 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711182941
과제번호 2020R1A2C3014820
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)
연구과제명 온-디바이스 개인화를 위한 에너지 효율적인 딥러닝 가속기 설계
과제수행기관명 고려대학교 산학협력단
연구기간 2023.03.01 ~ 2024.02.29
이 발명을 지원한 국가연구개발사업
과제고유번호 1711193814
과제번호 2021-0-00903-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보보호핵심원천기술개발(R&D)
연구과제명 고신뢰 온-디바이스 딥러닝 가속기 설계를 위한 물리채널 기반 취약점 검증 및 대응 기술 개발
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711193790
과제번호 2022-0-00266-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 PIM인공지능반도체핵심기술개발(설계)
연구과제명 초저전력 Low-Bit Precision 혼성모드 SRAM PIM 개발
과제수행기관명 서울대학교 산학협력단
연구기간 2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

스파이킹 뉴럴 네트워크에 기반한 추론 가속 장치의 동작 방법에 있어서,

입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택하는 단계;

상기 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단하는 단계; 및

상기 활성 뉴런의 추론 연산을 수행하고, 상기 비활성 뉴런의 추론 연산을 중지하는 동적 프루닝(pruning)을 수행하는 단계를 포함하고,

상기 활성 뉴런은 막전위값이 상기 기준값 이상인 뉴런이고,

상기 비활성 뉴런은 막전위값이 상기 기준값 미만인 뉴런인, 추론 가속 장치의 동작 방법.

청구항 2

제1항에 있어서,

상기 적어도 하나 이상의 선택 알고리즘은 제1 선택 알고리즘 및 제2 선택 알고리즘을 포함하고,

상기 제1 선택 알고리즘은 이분법(bisection method) 기반의 선택 알고리즘을 포함하고,

상기 제2 선택 알고리즘은 그리디 베스트-퍼스트 탐색(Greedy Best-First Search) 알고리즘을 포함하는, 추론 가속 장치의 동작 방법.

청구항 3

제2항에 있어서,

상기 제1 선택 알고리즘은,

손실 함수에 기초하여 상기 적어도 하나 이상의 층 각각에서 손실비가 가장 작은 상기 기준값을 선택하는, 추론 가속 장치의 동작 방법.

청구항 4

제3항에 있어서,

상기 뉴런의 막전위값에 대한 기준값을 선택하는 단계는,

상기 적어도 하나 이상의 층 각각의 첫 층부터 순차적으로 상기 제1 선택 알고리즘에 대한 적용 범위를 설정하는 단계;

제1 기준값을 선택하는 단계;

상기 제1 기준값에서 소정의 값만큼 감소시킨 제2 기준값을 선택하는 단계;

상기 제2 기준값의 손실에 기초하여 상기 제1 기준값을 선택하는 단계 및 상기 제2 기준값을 찾는 단계를 반복하는 단계; 및

상기 손실비가 가장 작은 상기 제2 기준값을 선택하는 단계를 포함하는, 추론 가속 장치의 동작 방법.

청구항 5

제4항에 있어서,

상기 제1 기준값은 손실값(loss)이 $(1+\beta)loss^{init}$인 막전위값이고,

상기 제2 기준값은 손실값이 $(1+\gamma)loss^{init}$인 막전위값인, 추론 가속 장치의 동작 방법.

여기에서, $loss^{init}$는 초기 손실값, β는 임의의 출력 손실비, γ는 임의의 출력 손실비, $0 \leq \gamma \leq \beta$임

청구항 6

제2항에 있어서,

상기 제2 선택 알고리즘은,

손실 함수에 기초하여 연산량 감소비가 가장 큰 상기 기준값을 선택하고,

상기 연산량 감소비는 연산량의 감소량 대비 손실값 증가량의 비율인, 추론 가속 장치의 동작 방법.

청구항 7

제6항에 있어서,

상기 뉴런의 막전위값에 대한 기준값을 선택하는 단계는,

상기 적어도 하나 이상의 층에서 상기 제2 선택 알고리즘에 대한 제1 기준값 집합을 설정하는 단계;

상기 제1 기준값 집합 중에서 적어도 하나 이상의 기준값을 소정의 값만큼 증가시키는 증가 단계;

상기 소정의 값만큼 증가된 제2 기준값 집합에 대한 상기 연산량 감소비를 연산하는 연산 단계;

상기 제2 기준값 집합의 각각의 기준값의 연산량에 기초하여 상기 증가 단계 및 상기 연산 단계를 반복하는 단계; 및

상기 연산량 감소비가 가장 큰 상기 기준값을 선택하는 단계를 포함하는, 추론 가속 장치의 동작 방법.

청구항 8

스파이킹 뉴럴 네트워크의 추론 가속 장치에 있어서,

적어도 하나의 프로그램이 기록된 메모리; 및

상기 적어도 하나의 프로그램을 실행함으로써 상기 스파이킹 뉴럴 네트워크를 프루닝(pruning)하는 동작을 처리하는 제어부를 포함하고,

상기 제어부는,

입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택하는 단계;

상기 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단하는 단계; 및

상기 활성 뉴런의 추론 연산을 수행하고, 상기 비활성 뉴런의 추론 연산을 중지하는 동적 프루닝(pruning)을 수행하는 단계를 포함하고,

상기 활성 뉴런은 막전위값이 상기 기준값 이상인 뉴런이고,

상기 비활성 뉴런은 막전위값이 상기 기준값 미만인 뉴런인, 스파이킹 뉴럴 네트워크의 추론 가속 장치.

청구항 9

제8항에 있어서,

상기 적어도 하나 이상의 선택 알고리즘은 제1 선택 알고리즘 및 제2 선택 알고리즘을 포함하고,

상기 제1 선택 알고리즘은 이분법(bisection method) 기반의 선택 알고리즘을 포함하고,

상기 제2 선택 알고리즘은 그리디 베스트-퍼스트 탐색(Greedy Best-First Search) 알고리즘을 포함하는, 스파이킹 뉴럴 네트워크의 추론 가속 장치.

청구항 10

제9항에 있어서,

상기 제1 선택 알고리즘은,

손실 함수에 기초하여 상기 적어도 하나 이상의 층 각각에서 손실비가 가장 작은 상기 기준값을 선택하는, 스파이킹 뉴럴 네트워크의 추론 가속 장치.

청구항 11

제9항에 있어서,

상기 제2 선택 알고리즘은,

손실 함수에 기초하여 연산량 감소비가 가장 큰 상기 기준값을 선택하는, 스파이킹 뉴럴 네트워크의 추론 가속 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 빈도 코딩(rate coding) 기반의 스파이킹 뉴럴 네트워크를 이용하는 추론 가속 장치 및 그 동작 방법에 관한 것이다.

배 경 기 술

[0002] 최근 IoT(Internet of Things), 엣지(edge) 컴퓨팅에 적합한 신경망으로 스파이킹 뉴럴 네트워크(이하 SNN, spiking neural network)가 많은 관심을 받고 있다. SNN은 생물학적 뉴런의 동작을 모방한 아키텍처이며, 이 아키텍처는 이벤트를 기반으로 작동하여 저전력을 소비한다. 더불어, 생체신호 모방을 통해 낮은 복잡도의 장치에서 지능을 구현하고자 하는 뉴로모픽 기술이 연구되고 있다.

[0003] 하지만 SNN은 주로 뉴런, 시냅스의 구조와 신호전달 체계에 관한 모델이 연구되고 있어 아직까지 반도체에 즉시 적용 가능한 단순하고 효과적인 모델이 없는 실정이다. 즉, 저전력 시스템에서 더 많은 뉴런을 지원하기 위한 SNN의 아키텍처가 필요하다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 한국 공개 특허 제 10-2023-0126388 호

발명의 내용

해결하려는 과제

[0005] 본 발명의 목적은 동적 프루닝 기법을 이용 빈도 코딩 기반의 스파이킹 뉴럴 네트워크의 추론 가속 장치 및 그 동작 방법을 제공함에 있다.

과제의 해결 수단

[0006] 본 발명에 따르면, 스파이킹 뉴럴 네트워크의 추론 가속 장치의 동작 방법은 입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택하는 단계, 상기 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단하는 단계 및 상기 활성 뉴런의 추론 연산을 수행하고, 상기 비활성 뉴런의 추론 연산을 중지하는 동적 프루닝(pruning)을 수행하는 단계를 포함할 수 있다. 여기에서, 상기 활성 뉴런은 막전위값이 상기 기준값 이상인 뉴런일 수 있고, 상기 비활성 뉴런은 막전위값이 상기 기준값 미만인 뉴런일 수 있다.

[0007] 그리고 상기 적어도 하나 이상의 선택 알고리즘은 제1 선택 알고리즘 및 제2 선택 알고리즘을 포함할 수 있고,

상기 제1 선택 알고리즘은 이분법(bisection method) 기반의 선택 알고리즘을 포함할 수 있고, 상기 제2 선택 알고리즘은 그리디 베스트-퍼스트 탐색(Greedy Best-First Search) 알고리즘을 포함할 수 있다.

[0008] 그리고 상기 제1 선택 알고리즘은 손실 함수에 기초하여 상기 적어도 하나 이상의 층 각각에서 손실비가 가장 작은 상기 기준값을 선택할 수 있다.

[0009] 그리고 상기 뉴런의 막전위값에 대한 기준값을 선택하는 단계는, 상기 적어도 하나 이상의 층 각각의 첫 층부터 순차적으로 상기 제1 선택 알고리즘에 대한 적용 범위를 설정하는 단계, 제1 기준값을 선택하는 단계, 상기 제1 기준값에서 소정의 값만큼 감소시킨 제2 기준값을 선택하는 단계, 상기 제2 기준값의 손실에 기초하여 상기 제1 기준값을 선택하는 단계 및 상기 제2 기준값을 찾는 단계를 반복하는 단계 및 상기 손실비가 가장 작은 상기 제2 기준값을 선택하는 단계를 포함할 수 있다.

[0010] 그리고 상기 제1 기준값은 손실값(loss)이 $(1+\beta)loss^{init}$인 막전위값일 수 있고, 상기 제2 기준값은 손실값이 $(1+\gamma)loss^{init}$인 막전위값일 수 있다. 여기에서, $loss^{init}$는 초기 손실값, β는 임의의 출력 손실비, γ는 임의의 출력 손실비, $0 \leq \gamma \leq \beta$이다.

[0011] 그리고 상기 제2 선택 알고리즘은, 손실 함수에 기초하여 연산량 감소비가 가장 큰 상기 기준값을 선택할 수 있고, 상기 연산량 감소비는 연산량의 감소량 대비 손실값 증가량의 비율일 수 있다.

[0012] 그리고 상기 뉴런의 막전위값에 대한 기준값을 선택하는 단계는, 상기 적어도 하나 이상의 층에서 상기 제2 선택 알고리즘에 대한 제1 기준값 집합을 설정하는 단계, 상기 제1 기준값 집합 중에서 적어도 하나 이상의 기준값을 소정의 값만큼 증가시키는 증가 단계, 상기 소정의 값만큼 증가된 제2 기준값 집합에 대한 상기 연산량 감소비를 연산하는 연산 단계, 상기 제2 기준값 집합의 각각의 기준값의 연산량에 기초하여 상기 증가 단계 및 상기 연산 단계를 반복하는 단계 및 상기 연산량 감소비가 가장 큰 상기 기준값을 선택하는 단계를 포함할 수 있다.

[0013] 또한, 본 발명에 따르면, 스파이킹 뉴럴 네트워크의 추론 가속 장치는 적어도 하나의 프로그램이 기록된 메모리 및 상기 적어도 하나의 프로그램을 실행함으로써 상기 스파이킹 뉴럴 네트워크를 프루닝(pruning)하는 동작을 처리하는 제어부를 포함할 수 있다. 상기 제어부는, 입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택하는 단계, 상기 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단하는 단계; 및 상기 활성 뉴런의 추론 연산을 수행하고, 상기 비활성 뉴런의 추론 연산을 중지하는 동적 프루닝(pruning)을 수행하는 단계를 포함할 수 있고, 상기 활성 뉴런은 막전위값이 상기 기준값 이상인 뉴런일 수 있고, 상기 비활성 뉴런은 막전위값이 상기 기준값 미만인 뉴런일 수 있다.

발명의 효과

[0014] 본 발명의 실시예에 따른 스파이킹 뉴럴 네트워크의 추론 가속 장치 및 그 동작 방법은 추론에 소모되는 에너지 및 시간을 감소시킬 수 있다.

[0015] 이에 따라 본 발명의 실시예에 따른 스파이킹 뉴럴 네트워크의 추론 가속 장치 및 그 동작 방법은 저전력으로 동작하는 장치에서 구동될 수 있다.

도면의 간단한 설명

[0016] 도 1a 및 도 1b는 빈도 코딩 기반의 스파이킹 뉴럴 네트워크에 대한 예시이다.

도 2a 및 도 2b는 스파이킹 뉴럴 네트워크의 복수의 뉴런에 대한 막전위값에 대한 예시이다.

도 3은 본 발명의 실시예에 따른 스파이킹 뉴럴 네트워크의 추론 가속 장치를 포함하는 추론 가속 시스템의 개략적인 구성도이다.

도 4는 본 발명의 실시예에 따른 추론 가속 장치의 동작 방법에 대한 순서도이다.

도 5a 및 도 5b는 본 발명의 실시예에 따른 제1 선택 알고리즘에 대한 예시이다.

도 6은 본 발명의 실시예에 따른 제2 선택 알고리즘에 대한 예시이다.

도 7은 본 발명의 실시예에 따른 추론 가속 장치의 동작 방법과 종래 기술을 비교한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0017] 본 발명의 실시예들은 컴퓨터 장치를 통해 구현될 수 있다. 이때, 컴퓨터 장치에는 본 발명의 일실시예에 따른 컴퓨터 프로그램이 설치 및 구동될 수 있고, 컴퓨터 장치는 구동된 컴퓨터 프로그램의 제어에 따라 본 발명의 실시예들에 따른 질의 처리 방법을 수행할 수 있다. 상술한 컴퓨터 프로그램은 컴퓨터 장치와 결합되어 질의 처리 방법을 컴퓨터에 실행시키기 위해 컴퓨터 판독 가능한 기록매체에 저장될 수 있다. 실시예에 따라 본 발명인 스파이킹 뉴럴 네트워크의 추론 가속 장치(10)는 둘 이상의 컴퓨터 장치들간의 연계를 통해 구현될 수도 있다.

[0018] 상기 컴퓨터 장치는 메모리, 프로세서, 통신 인터페이스 그리고 입출력 인터페이스를 포함할 수 있다. 상기 메모리는 컴퓨터에서 판독 가능한 기록매체로서, RAM(random access memory), ROM(read only memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록장치(permanent mass storage device)를 포함할 수 있다. 여기서 ROM과 디스크 드라이브와 같은 비소멸성 대용량기록장치는 상기 메모리와는 구분되는 별도의 영구 저장 장치로서 상기 컴퓨터 장치에 포함될 수도 있다.

[0019] 또한, 상기 메모리에는 운영체제와 적어도 하나의 프로그램 코드가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 상기 메모리와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 상기 메모리로 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 인터페이스를 통해 상기 메모리에 로딩될 수도 있다. 예를 들어, 소프트웨어 구성요소들은 네트워크를 통해 수신되는 파일들에 의해 설치되는 컴퓨터 프로그램에 기반하여 상기 컴퓨터 장치의 상기 메모리에 로딩될 수 있다.

[0020] 상기 프로세서는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 상기 메모리 또는 상기 통신 인터페이스에 의해 상기 프로세서로 제공될 수 있다. 예를 들어 상기 프로세서는 상기 메모리와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.

[0021] 상기 통신 인터페이스는 네트워크를 통해 상기 컴퓨터 장치가 다른 장치(일례로, 앞서 설명한 저장장치들)와 서로 통신하기 위한 기능을 제공할 수 있다. 일례로, 상기 컴퓨터 장치의 상기 프로세서가 상기 메모리와 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청이나 명령, 데이터, 파일 등이 상기 통신 인터페이스의 제어에 따라 상기 네트워크를 통해 다른 장치들로 전달될 수 있다. 역으로, 다른 장치로부터의 신호나 명령, 데이터, 파일 등이 상기 네트워크를 거쳐 상기 컴퓨터 장치의 상기 통신 인터페이스를 통해 상기 컴퓨터 장치로 수신될 수 있다. 상기 통신 인터페이스를 통해 수신된 신호나 명령, 데이터 등은 상기 프로세서나 상기 메모리로 전달될 수 있고, 파일 등은 상기 컴퓨터 장치가 더 포함할 수 있는 저장 매체(상술한 영구 저장 장치)로 저장될 수 있다.

[0022] 상기 입출력 인터페이스는 상기 입출력 장치와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 입력 장치는 마이크, 키보드 또는 마우스 등의 장치를, 그리고 출력 장치는 디스플레이, 스피커와 같은 장치를 포함할 수 있다. 다른 예로 상기 입출력 인터페이스는 터치스크린과 같이 입력과 출력을 위한 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수도 있다. 상기 입출력 장치는 상기 컴퓨터 장치와 하나의 장치로 구성될 수도 있다.

[0023] 또한, 다른 실시예들에서 상기 컴퓨터 장치 상술한 구성요소들보다 더 적은 혹은 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, 상기 컴퓨터 장치는 상술한 상기 입출력 장치 중 적어도 일부를 포함하도록 구현되거나 또는 트랜시버(transceiver), 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수도 있다.

[0024] 이하에서는 본 발명의 기술사상을 본 발명의 기술분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있을 정도로 상세히 설명하기 위하여, 본 발명의 실시 예들이 첨부된 도면을 참조하여 설명될 것이다.

[0026] 도 1a 및 도 1b는 빈도 코딩 기반의 스파이킹 뉴럴 네트워크(이하 SNN)에 대한 예시이다.

[0027] 심층 신경망(Deep Neural Network)의 일종인 SNN은 인공 신경망을 구현하는 하나의 방식으로서, 복수의 층(layer)을 포함할 수 있다. 예를 들어, SNN은 입력 데이터가 인가되는 입력층(Input Layer), 학습을 바탕으로 입력 데이터에 기반한 예측을 통해 도출된 결과 값을 출력하는 출력층(Output Layer) 및 입력층과 출력층 사이

에 다중의 은닉층(Hidden Layer)을 포함한다.

[0028] SNN의 층들 각각에 포함된 뉴런(채널)들은 서로 연결되어 데이터를 처리할 수 있다. 예를 들어, 하나의 뉴런은 다른 뉴런들로부터 데이터인 스파이크를 수신하여 연산할 수 있고, 연산 결과를 또 다른 뉴런들로 출력할 수 있다.

[0029] 뉴런들 각각은 이전 층에 포함된 뉴런들로부터 수신된 스파이크, 가중치 및 바이어스에 기초하여 자신의 활성을 결정할 수 있다. 여기에서 가중치는 각 뉴런에서의 출력 스파이크를 연산하기 위해 이용되는 파라미터로서, 뉴런들 간의 연결관계에 할당되는 값일 수 있다.

[0030] 도 1a를 참조하면, SNN은 이진 스파이크(binary spike) 신호로 정보를 전달할 수 있다. SNN은 타임 스텝(time step)에 따라 막전위값을 축적해 문턱값을 초과하는 경우에 스파이크가 발생할 수 있다. SNN은 스파이크가 발생하는 경우에만 연산이 일어나므로, 저전력 하드웨어 구현이 가능하다. 입력 데이터(이미지 등)은 먼저 시간에 대한 1 bit 스파이크로 변환되어 네트워크에 입력될 수 있다. 즉, 스파이크가 인코딩되어 입력층에 입력될 수 있다. 이후, 입력층의 뉴런이 업데이트될 수 있고, 뉴런이 업데이트되면 스파이크가 발생하여 다음 층인 은닉층으로 전달될 수 있다.

[0031] 네트워크 은닉층에 있는 뉴런들에 입력 스파이크가 들어오는 경우에, 그 스파이크가 들어온 시냅스의 가중치를 뉴런의 막전위에 더할 수 있다. 입력 스파이크가 막전위값에 충분히 쌓여 뉴런의 막전위값이 문턱값을 넘는 경우에 뉴런은 출력 스파이크를 만들수 있고 막전위 값은 0으로 초기화될 수 있다. 그리고 해당 뉴런의 출력 스파이크는 다음 층으로 전달될 수 있다. 결국, 은닉층은 스파이크의 가중치와 문턱값을 출력층으로 전달할 수 있다. 이후, 출력층은 출력 데이터인 Class1 또는 Class 2를 출력할 수 있다. 상술한 SNN의 동작 과정은 입력층부터 마지막 출력층까지 반복될 수 있다. 즉, 상술한 과정을 여러 타임 스텝동안 반복 수행함으로써 SNN의 추론 동작을 수행할 수 있다.

[0032] 여기에서, SNN 추론의 반복적인 연산 수행 중에 비활성 뉴런이 존재할 수 있다. 비활성 뉴런은 출력 스파이크를 발생시키지 않으므로 정확도에 영향이 적은 뉴런으로 가정될 수 있다.

[0033] 도 1b를 참조하면, SNN은 0과 1의 1bit 스파이크 단위로 정보를 입력받아 전달할 수 있다. 이에 따라, SNN은 하드웨어 구현에 있어 저전력으로 구동이 가능하다.

[0034] 빈도 코딩(rate coding) 기법은 입력값에 비례하여 스파이크 빈도수(spike rate)를 증가시키는 코딩 기법이다. 입력값이 클수록 빠른 시간에 잦은 스파이크를 입력하고 반면에, 입력값이 작을수록 드문 스파이크를 입력하는 코딩 기법이다. 빈도 코딩을 사용하는 경우에는 특정 값을 표현하기 위하여 시간에 대해 반복적으로 스파이크를 사용하므로, 추론 과정에 반복적인 연산으로 이한 큰 에너지 소모와 긴 시간을 야기시킬 수 있다.

[0035] 도 2a 및 도 2b는 스파이킹 뉴럴 네트워크의 복수의 뉴런에 대한 막전위값에 대한 예시이다.

[0036] 프루닝(pruning) 기법은 추론 동작 중에 피쳐 맵에서 중요하지 않은 뉴런(채널)을 선택적으로 잘라내는 기법이다. 프루닝 기법은 뉴런의 중요도에 따라 중요한 뉴런과 중요하지 않은 뉴런을 구분해서 잘라낼 수 있다. 본 발명에 따른 SNN은 프루닝이 수행될 수 있다. 예를 들어, 종래 기술에서는 프루닝 되기 전 SNN에서, 복수의 뉴런들 중에서 뉴런의 가중치가 소정의 임계값 이하인 경우, 해당 뉴런의 연산을 약화 또는 제거하는 프루닝이 수행될 수 있다. 프루닝될 수 있는 SNN의 뉴런을 결정하기 위해, SNN이 탐색될 수 있다. 이때, SNN의 정확도를 실질적으로 손상시키지 않고, SNN 파라미터들의 부분들 또는 층의 각 뉴런들을, 제거하거나 감소시킴으로써 SNN은 프루닝될 수 있다.

[0037] 하기의 식은 시냅스후 전위(PSP, Post Synaptic Potential)의 누적값으로, 막전위값을 연산하기 위해 필요한 연산이다. 종래의 막전위값(membrane potential)은 하기의 식으로 연산한다.

[0038] $$u_j^l(t)=u_j^l(t-1)+z_j^l(t)$$

[0039] 여기에서, t는 스파이크 시간이다.

[0040] 종래의 시냅스후 전위의 누적값 zjl은 하기의 식과 같이 복잡한 곱셈 연산을 포함한다.

[0041] $$z_j^l(t) = \sum_i w_{ij}^l \times \varepsilon_{IN}^l(t^{l-1} - t_{ref}^{l-1}) + b_j^l$$

[0042] 여기에서, l은 층 순서, t^{l-1}는 l-1번째 층의 스파이크 시간, t^l_{ref}는 스파이스 연산 시작 시간, w^l_{ij}는 시냅스 가중치, ε^l_{IN}는 커널의 적분, b^l_j는 수상돌기(dedrite) 함수이다.

[0043] 도 2a를 참조하면, 출력 스파이크가 출력되지 않는 비활성 뉴런들은 지속적으로 음의 막전위값을 가진다는 점을 이용하여, 추론 성능에 영향이 적은 수준에서 추론 중에 뉴런을 프루닝할 수 있다. 프루닝된 뉴런은 네트워크의 입력 데이터에 대한 추론이 끝날 때까지 추가적인 연산이 일어나지 않을 수 있다. 이에 따라, SNN 추론의 에너지 소모량과 소요 시간이 감소될 수 있다.

[0044] 뉴런은 막전위값이 소정 기준값(pruning threshold)보다 작은 경우에 프루닝될 수 있다. 도 2b를 참조하면, -2, -4, -6의 기준값을 SNN의 출력층을 제외한 모든 층에 적용한 경우에 프루닝된 뉴런의 비율에 대한 네트워크의 분류 정확도 감소량을 나타낸다. 기준값이 0에 가까울수록 높은 비율의 뉴런이 프루닝될 수 있으나 분류 정확도 또한 더 많이 감소한다. 예를 들어, 기준값이 0에 가까운 경우에 많은 뉴런이 프루닝될 수 있고, 정확도가 크게 감소될 수 있다. 반면에, 기준값이 0에 먼 경우에 적은 뉴런이 프루닝될 수 있고, 정확도가 적게 감소될 수 있다. 이에 따라, 최종 정확도에 대한 중요도를 고려하여 뉴런을 프루닝하기 위한 기준값을 선택하는 과정이 필요하다. 즉, 정확도의 감소가 적고, 프루닝 비율이 높은 기준값을 선택하는 알고리즘이 중요하다.

[0045] 본 발명의 실시예에 관한 설명에서 SNN을 실시예로 설명하였으나, 이에만 한정되는 것은 아니고, 구현예에 따라 합성곱 신경망(Convolutional Neural Network, CNN), 심층 신경망(Deep Neural Network, DNN), 재귀 신경망(Recurrent Neural Network, RNN) 등 종래 이미 공지되었거나 향후 개발될 수 있는 다양한 유형의 인공 신경망을 폭넓게 포함할 수 있다.

[0046] 도 3은 본 발명의 실시예에 따른 스파이킹 뉴럴 네트워크의 추론 가속 장치(10)를 포함하는 신경망 기반의 추론 시스템(1)의 개략적인 구성도이다.

[0047] 도 3을 참조하면, 추론 시스템(1)은 추론 가속 장치(10) 및 뉴로모픽 장치(20)를 포함할 수 있고, 추론 가속 장치(10)는 메모리(100), 제어부(200) 및 통신 모듈(300)을 포함할 수 있다.

[0048] 추론 가속 장치(10)는 메모리(100) 및 시스템 버스 또는 다른 적절한 회로를 통해 메모리(100)와 연결된 제어부(200)를 포함할 수 있다.

[0049] 추론 가속 장치(10)는 메모리(100)에 프로그램 코드를 저장할 수 있다. 제어부(200)는 시스템 버스를 통해 메모리(100)에서 불러들인 프로그램 코드를 실행함으로써 인공 신경망을 프루닝하는 동작을 처리할 수 있다.

[0050] 메모리(100)는 로컬 메모리 또는 하나 이상의 대용량 저장 장치들(bulk storage devices)과 같은 하나 이상의 물리적 메모리 장치들을 포함할 수 있다. 이 때, 로컬 메모리는 RAM(Random Access Memory) 또는 프로그램 코드를 실제로 실행하는 동안 일반적으로 사용되는 다른 휘발성 메모리 장치를 포함할 수 있다. 대용량 저장 장치는 HDD(Hard Disk Drive), SSD(Solid State Drive) 또는 다른 비휘발성 메모리 장치로 구현될 수 있다.

[0051] 또한, 추론 가속 장치(10)는 프루닝 동작을 수행하는 동안에 대용량 저장 장치들에서 프로그램 코드를 검색하는 횟수를 줄이기 위해, 적어도 일부 프로그램 코드의 임시 저장 공간을 제공하는, 하나 이상의 캐시 메모리들(미도시)을 포함할 수 있다.

[0052] 메모리(100)에 저장된 실행 가능한 프로그램 코드가 추론 가속 장치(10)에 의해 실행됨에 따라, 제어부(200)에 의해 본 개시에 기재된 다양한 동작들을 수행할 수 있다. 예를 들어, 메모리(100)는 제어부(200)가 도 4 내지 도 6에 기재된 하나 이상의 동작을 수행하도록 하기 위한 프로그램 코드를 저장할 수 있다.

[0053] 메모리(100)는 데이터를 저장하기 위한 저장 장소로서, OS(Operating System), 각종 프로그램들, 및 각종 데이터를 저장할 수 있다. 실시예에 있어서, 메모리(100)는 뉴로모픽 장치(20)의 연산 수행 과정에서 생성되는 중간 결과들 또는 연산 수행 과정에 이용되는 가중치들을 저장할 수 있다.

[0054] 메모리(100)는 DRAM일 수 있으나, 이에 한정되는 것은 아니다. 메모리(100)는 휘발성 메모리 또는 불휘발성 메

모리 중 적어도 하나를 포함할 수 있다. 불휘발성 메모리는 ROM, PROM, EPROM, EEPROM, 플래시 메모리, PRAM, MRAM, RRAM, FRAM 등을 포함한다. 휘발성 메모리는 DRAM, SRAM, SDRAM, PRAM, MRAM, RRAM, FeRAM 등을 포함한다. 실시예에 있어서, 메모리(100)는 HDD, SSD, CF, SD, Micro-SD, Mini-SD, xD 또는 Memory Stick 중 적어도 하나를 포함할 수 있다.

[0055] 구현되는 장치의 특정 유형에 따라, 추론 가속 장치(10)는 도시된 구성 요소보다 적은 구성 요소들 또는 도 3에 도시되지 않은 추가적인 구성 요소들을 포함할 수 있다. 또한, 하나 이상의 구성 요소들은 다른 구성 요소에 포함될 수 있고, 그렇지 않으면 다른 구성 요소의 일부를 형성할 수 있다.

[0056] 제어부(200)는 입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택할 수 있다. 적어도 하나 이상의 선택 알고리즘은 제1 선택 알고리즘 및 제2 선택 알고리즘을 포함할 수 있다.

[0057] 제1 선택 알고리즘은 이분법(bisection method) 기반의 선택 알고리즘을 포함할 수 있다. 제1 선택 알고리즘은 손실 함수에 기초하여 적어도 하나 이상의 층 각각에서 손실비가 가장 작은 기준값을 선택할 수 있다.

[0058] 제2 선택 알고리즘은 그리디 베스트-퍼스트 탐색(Greedy Best-First Search) 알고리즘을 포함할 수 있다. 제2 선택 알고리즘은 손실 함수에 기초하여 연산량 감소비가 가장 큰 기준값을 선택할 수 있다.

[0059] 제어부(200)는 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단할 수 있다. 여기에서, 활성 뉴런은 막전위값이 기준값 이상인 뉴런일 수 있고, 비활성 뉴런은 막전위값이 기준값 미만일 수 있다.

[0060] 제어부(200)는 활성 뉴런의 추론 연산을 수행할 수 있다. 한편, 제어부(2200)는 비활성 뉴런의 추론 연산을 중지하는 동적 프루닝(pruning)을 수행할 수 있다.

[0061] 제어부(200)의 구체적인 동작은 도 4 내지 도 6의 설명에서 구제적으로 후술하기로 한다.

[0062] 제어부(200)는 추론 시스템(1)의 전반적인 동작을 제어한다. 제어부(200)는 싱글 코어(Single Core)를 포함하거나, 멀티 코어들(Multi-Core)을 포함할 수 있다. 제어부(200)는 메모리(100)에 저장된 프로그램들 및/또는 데이터를 처리 또는 실행할 수 있다. 일부 실시예에 있어서, 제어부(200)는 메모리(100)에 저장된 프로그램들을 실행함으로써, 뉴로모픽 장치(20)의 기능을 제어할 수 있다. 또한, 제어부(200)는 뉴로모픽 장치(20)에 의해 이용되는 가중치 정보 량을 감소시키는 프루닝을 수행할 수 있다. 제어부(200)는 CPU, GPU, AP 등으로 구현될 수 있다.

[0063] 뉴로모픽 장치(20)는 수신되는 입력 데이터를 기초로 연산을 수행하고, 수행 결과를 기초로 정보 신호를 생성할 수 있다. 뉴로모픽 장치(20)는 인공 신경망 전용 하드웨어 가속기 자체 또는 이를 포함하는 장치에 해당할 수 있다. 여기에서, 정보 신호는 음성 인식 신호, 사물 인식 신호, 영상 인식 신호, 생체 정보 인식 신호 등과 같은 다양한 종류의 인식 신호 중 하나를 포함할 수 있다. 예를 들어, 뉴로모픽 장치(20)는 비디오 스트림에 포함되는 프레임 데이터를 입력 데이터로서 수신하고, 프레임 데이터로부터 프레임 데이터가 나타내는 이미지에 포함된 사물에 대한 인식 신호를 생성할 수 있다. 그러나, 이에 제한되는 것은 아니며, 추론 시스템(1)이 탑재된 전자 장치의 종류 또는 기능에 따라 뉴로모픽 장치(20)는 다양한 종류의 입력 데이터를 수신할 수 있고, 입력 데이터에 따른 인식 신호를 생성할 수 있다.통신 모듈(300)은 외부 디바이스와 통신할 수 있는 다양한 유선 또는 무선 인터페이스를 구비할 수 있다. 예를 들어, 통신 모듈(300)은 유선 근거리통신망(Local Area Network; LAN), Wi-fi(Wireless Fidelity)와 같은 무선 근거리 통신망 (Wireless Local Area Network; WLAN), 블루투스(Bluetooth)와 같은 무선 개인 통신망(Wireless Personal Area Network; WPAN), 무선 USB (Wireless Universal Serial Bus), Zigbee, NFC (Near Field Communication), RFID (Radio-frequency identification), PLC(Power Line communication), 또는 3G (3^{rd} Generation), 4G (4th Generation), LTE (Long Term Evolution) 등 이동통신망(mobile cellular network)에 접속 가능한 통신 인터페이스 등을 포함할 수 있다.

[0064] 상술한 바와 같이, 본 발명의 실시예에 따른 스파이킹 뉴럴 네트워크의 추론 가속 장치(10)는 추론 결과에 영향이 적은 뉴런을 추론 중에 판단함으로써, 판단된 뉴런의 연산을 수행하지 않으므로 추론에 소모되는 에너지 및 시간을 감소시킬 수 있다. 이에 따라 스파이킹 뉴럴 네트워크의 추론 가속 장치(10)는 저전력으로 동작하는 장치에서 구동될 수 있다.

[0066] 도 4는 본 발명의 실시예에 따른 도 3에 도시된 추론 가속 장치(10)의 동작 방법에 대한 순서도이다. 도 5a 및

도 5b는 본 발명의 실시예에 따른 제1 선택 알고리즘에 대한 예시이고, 도 6은 본 발명의 실시예에 따른 제2 선택 알고리즘에 대한 예시이다.

[0067] 도 4를 참조하면, S1100 단계에서, 입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값이 선택될 수 있다. 예를 들어, 도 3의 제어부(200)는 입력층, 은닉층 및 출력층 중 적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 뉴런의 막전위값에 대한 기준값을 선택할 수 있다.

[0068] 적어도 하나 이상의 선택 알고리즘은 제1 선택 알고리즘 및 제2 선택 알고리즘을 포함하는데, 제1 선택 알고리즘은 이분법(bisection method) 기반의 선택 알고리즘을 포함할 수 있고, 제2 선택 알고리즘은 그리디 베스트-퍼스트 탐색(Greedy Best-First Search) 알고리즘을 포함할 수 있으나, 이에 한정하지 않고 후술하는 기준값 선택에 대한 연산을 수행하는 알고리즘은 모두 가능하다.

[0069] 이분법 기반의 선택 알고리즘은 연속함수에서 해가 존재하는 폐구간에서 근을 찾는 알고리즘이다. 이러한 폐구간을 이분하는 동작을 반복함으로써, 해가 존재하는 구간 범위를 줄일 수 있다. 이는 함수의 수식이 필요 없으므로, 함수화가 어려운 네트워크에 적용 가능하다.

[0070] 도 5a를 참조하면, SNN 내 하나 층의 기준값을 찾는 과정에서 한 층의 기준값에 따른 손실 함수을 보여준다. 기준값을 선택을 진행하는 과정 중에서 초기에는 손실값이 거의 증가하지 않다가, 특정 지점(knee point)를 지난 후 급격하게 손실값이 증가할 수 있다. 제1 선택 알고리즘인 이분법 기반의 선택 알고리즘은 특정 지점 근처의 지점을 찾는 방법으로써, 이러한 방식으로 이분법 기반의 선택 알고리즘은 기준값을 빠르게 찾을 수 있다. 이를 위해, 알고리즘은 네트워크의 층별로 첫 층부터 기준값을 찾는다.

[0071] 도 5b를 참조하면, 제1 선택 알고리즘은 손실 함수에 기초하여 적어도 하나 이상의 층 각각에서 손실비가 가장 작은 기준값을 선택할 수 있다. 예를 들어, 제어부(200)는 적어도 하나 이상의 층 각각의 첫 층부터 순차적으로 제1 선택 알고리즘에 대한 적용 범위를 설정할 수 있다.

[0072] 이후, 제어부(200)는 제1 기준값을 선택할 수 있다. 여기에서, 제1 기준값은 손실값(loss)이 $(1+\beta)loss^{init}$인 막전위값일 수 있다. $loss^{init}$는 초기 손실값, β는 사용자가 설정한 임의의 출력 손실비일 수 있다. 다시 말해, 특정 층의 기준값의 손실값은 출력 손실값으로 가정될 수 있다.

[0073] 그리고 제어부(200)는 제1 기준값에서 소정의 값인 △b 만큼 감소시킨 제2 기준값을 선택할 수 있다. 여기에서, △B는 사용자가 설정한 임의의 하이퍼 파라미터일 수 있다. 제2 기준값은 손실값이 $(1+\gamma)loss^{init}$인 막전위값일 수 있고, γ는 사용자가 설정한 임의의 출력 손실비일 수 있다. 이에 따라, $0 \leq \gamma \leq \beta$이다.

[0074] 제어부(200)는 제2 기준값의 손실에 기초하여 상술한 과정을 반복 수행할 수 있다. 이 과정을 반복하면서, 소정의 값인 △B는 점차 감소시킬 수 있다. △B를 점차 감소시키면서, 최적의 기준값을 선택하기 위한 back tracking 과정을 수행할 수 있다. 제어부(200)는 상술한 과정을 SNN의 마지막 층까지 반복할 수 있다.

[0075] 결국, 제어부(200)는 손실비가 가장 작은 제2 기준값을 선택할 수 있다.

[0076] 상술한 바와 같이, 제1 선택 알고리즘은 후술하는 제2 선택 알고리즘에 비교하여 기준값을 선택하는데, 빠른 연산 처리가 가능하다.

[0078] 도 6을 참조하면, 그리디 베스트-퍼스트 탐색 알고리즘은 주어진 연산량 감소비를 최소한의 네트워크 손실 함수(loss function)값 증가로 도달' 이라는 목표로 각 층의 기준값을 찾는 알고리즘이다. 여기에서 손실 함수는 일반적으로 많이 사용하는 cross entropy loss를 사용할 수 있다.

[0079] 다시 말해, 제2 선택 알고리즘은 손실 함수에 기초하여 연산량 감소비가 가장 큰 상기 기준값을 선택할 수 있다. 여기에서, 연산량 감소비는 연산량의 감소량 대비 손실값 증가량의 비율일 수 있다. 예를 들어, 제어부(200)는 적어도 하나 이상의 층에서 제2 선택 알고리즘에 대한 제1 기준값 집합을 설정할 수 있다. 여기에서 제1 기준값 집합은 $Pth_{current}$={pth_1, pth_2, ..., pth_l}으로 나타낼 수 있다. 초기의 제1 기준값 집합은 충분히 작은 값으로 선택되어야 SNN의 성능 저하를 최소화할 수 있다.

[0080] 이후, 제어부(200)는 제1 기준값 집합 중에서 적어도 하나 이상의 기준값을 소정의 값 △G만큼 증가시킬 수 있

다. 여기에서, △G는 사용자가 설정한 임의의 하이퍼 파라미터일 수 있다.

[0081] 제어부(200)는 소정의 값 △G만큼 증가된 제2 기준값 집합에 대한 연산량 감소비를 연산할 수 있다. 여기에서, 제2 기준값 집합은 Pth_{next}일 수 있다.

[0082] 제어부(200)는 제2 기준값 집합의 각각의 기준값의 연산량에 기초하여 상술한 과정을 반복할 수 있다. 여기에서 연산량은 총 SOP(Synaptic Operation)의 수이고, SOP는 뉴런이 입력 스파이크 1개에 대해 막전위값을 갱신하는 연산이다. 즉, 제어부(200)는 제2 기준값 집합에서 연산량의 감소량과 손실값의 증가량의 비율이 가장 큰 기준값을 선택할 수 있다. 예를 들어, 현재 연산량/초기 연산량 값이 목표치(α) 미만이면 상술한 단계를 반복할 수 있다.

[0083] 제어부(200)는 연산량 감소비가 가장 큰 기준값을 선택할 수 있다. 예를 들어, 제어부(200)는 목표치(α)를 달성한 경우에는 연산량 감소비가 가장 큰 기준값을 선택할 수 있다.

[0084] 상술한 바와 같이, 제1 선택 알고리즘 및 제2 선택 알고리즘 모두 정확도 저하가 적고, 연산량 감소가 가능하다. 구체적으로, 제2 선택 알고리즘은 연산량 감소비를 설정 가능하므로, 기준값 선택 속도가 느리다. 제1 선택 알고리즘은 목표 연산량 감소비 설정이 불가능하므로, 기준값 선택 속도가 빠르다.

[0085] 제2 선택 알고리즘은 제1 선택 알고리즘에 비교하여 모든 층의 기준값을 동시에 찾기 때문에 연산 시간이 느릴 수 있다. 예를 들어, SNN의 층의 수가 커질수록 알고리즘 수행 시간이 급속도로 증가할 수 있다. 이에 따라, 초기 제1 기준값 집합의 선정이 어려운 문제가 있다. 기준값이 0에 가까운 경우 정확도가 감소하고, 기준값이 0에서 너무 먼 경우에 알고리즘 수행 시간이 증가하는 문제가 있다.

[0086] 이를 해결하기 위해, 제어부(200)는 제1 선택 알고리즘을 수행하여 제1 기준값 집합을 선택할 수 있다. 이후, 선택된 제1 기준값 집합으로 제2 선택 알고리즘을 수행함으로써, 상술한 문제점을 보완할 수 있다. 즉, 제1 선택 알고리즘 및 제2 선택 알고리즘을 순차적으로 함께 사용하는 경우에 정확도 저하가 감소될 수 있고, 연산 처리 속도가 빨라질 수 있다.

[0087] 또한 제1 선택 알고리즘 및 제2 선택 알고리즘 각각은 단독으로 사용 가능하다.

[0088] S1200 단계에서, 기준값에 기초하여 활성 뉴런 및 비활성 뉴런이 판단될 수 있다. 즉, 뉴런의 막전위값이 기준값 이상인지 판단될 수 있다. 예를 들어, 제어부(200)는 기준값에 기초하여 활성 뉴런 및 비활성 뉴런을 판단할 수 있다.

[0089] 뉴런의 막전위값이 기준값 이상이면 S1300 단계에서, 활성 뉴런으로 판단될 수 있고, 추론 연산을 수행할 수 있다.

[0090] 뉴런의 막전위값이 기준값 미만이면 S1400 단계에서, 비활성 뉴런으로 판단될 수 있고, 추론 연산을 중지하는 동적 프루닝을 수행할 수 있다.

[0091] 상술한 바와 같이, 본 발명의 실시예에 따른 추론 가속 장치(10)의 동작 방법은 추론 결과에 영향이 적은 뉴런을 추론 중에 판단함으로써, 판단된 뉴런의 연산을 수행하지 않으므로 추론에 소모되는 에너지 및 시간을 감소시킬 수 있다. 이에 따라 추론 가속 장치(10)의 동작 방법은 저전력으로 동작하는 장치에서 수행될 수 있다.

[0093] 도 7은 본 발명의 실시예에 따른 추론 가속 장치(10)의 동작 방법과 종래 기술을 비교한 그래프이다.

[0094] 도 7을 참조하면, 베이스라인은 프루닝 기법이 적용되지 않은 SNN이고, 종래 기술은 종래의 프루닝 기법이 적용된 SNN이다.

[0095] 도 7의 (a)를 참조하면, 본 발명에 따른 추론 가속 장치(10)의 면적 오버헤드는 종래 기술에 비교하여 감소하였다. 구체적으로, 종래 기술에 비교하여 추론 가속 장치(10)의 면적 오버헤드는 36%에서 7.3%로 감소한 것을 알 수 있다.

[0096] 도 7의 (b)를 참조하면, 본 발명에 따른 추론 가속 장치(10)의 에너지 감소율은 종래 기술에 비교하여 증가하였다. 구체적으로, 종래 기술에 비교하여 추론 가속 장치(10)의 에너지 감소율은 42%에서 57%로 증가한 것을 알 수 있다.

[0097] 상술한 바와 같이, 본 발명에 따른 추론 가속 장치(10)는 병렬적으로 뉴런의 연산을 처리하므로, 빠른 처리 속

도를 보여줄 수 있다. 추론 가속 장치(10)는 비활성 뉴런의 추론 연산을 생략하므로, 종래 기술에 비교하여 더 많은 처리량을 보일 수 있고, 더 낮은 에너지 소모를 보일 수 있다.

[0099] 상술한 내용은 본 발명을 실시하기 위한 구체적인 실시 예들이다. 본 발명은 상술한 실시 예들 이외에도, 단순하게 설계 변경되거나 용이하게 변경할 수 있는 실시 예들도 포함될 것이다. 또한, 본 발명은 실시 예들을 이용하여 용이하게 변형하여 실시할 수 있는 기술들도 포함될 것이다. 따라서, 본 발명의 범위는 상술한 실시 예들에 국한되어 정해져서는 안되며, 후술하는 특허청구범위뿐만 아니라 이 발명의 특허청구범위와 균등한 것들에 의해 정해져야 할 것이다.

[0100] 이상에서 설명된 시스템 또는 장치는 하드웨어 구성요소, 또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 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, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 어플리케이션을 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 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0101] 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 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록매체에 저장될 수 있다.

[0102] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수개 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 애플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0103] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 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.

부호의 설명

[0105] 1 : 추론 시스템

10 : 스파이킹 뉴럴 네트워크의 추론 가속 장치

100 : 메모리

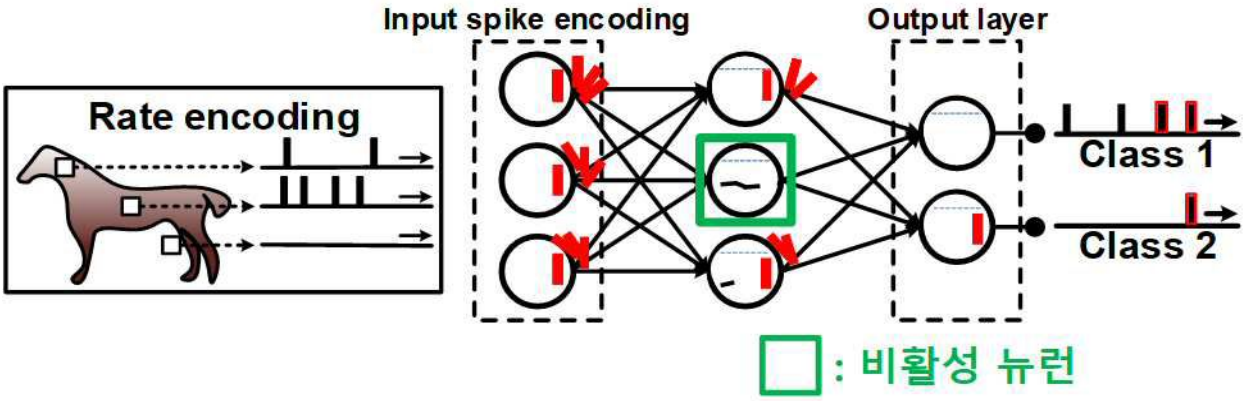
200 : 제어부

300 : 통신모듈

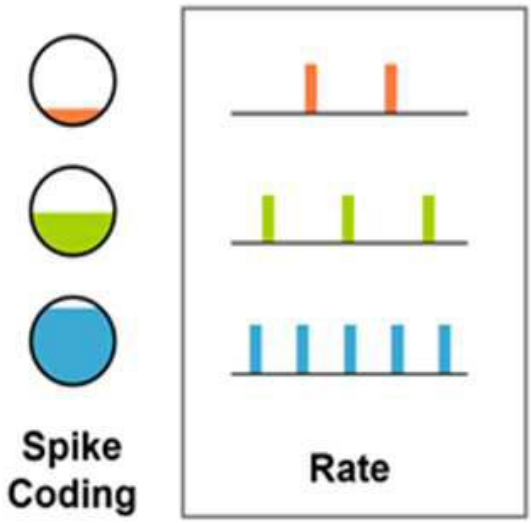
20 : 뉴로모픽 장치

도면

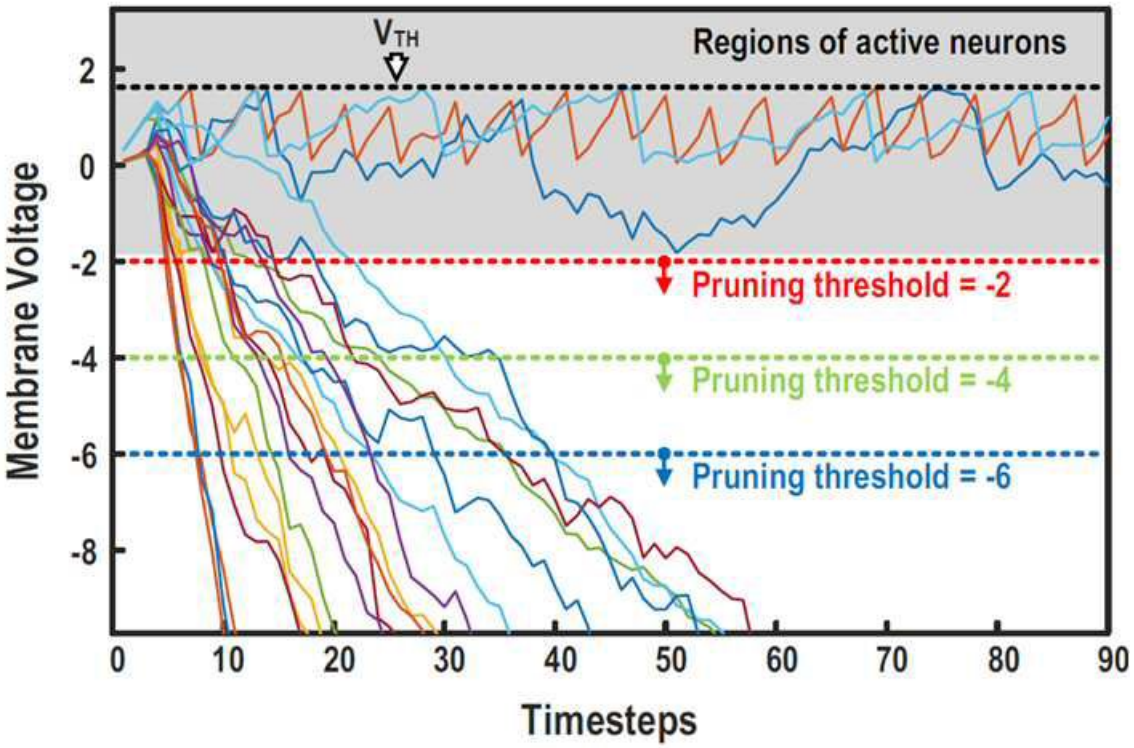
도면1a



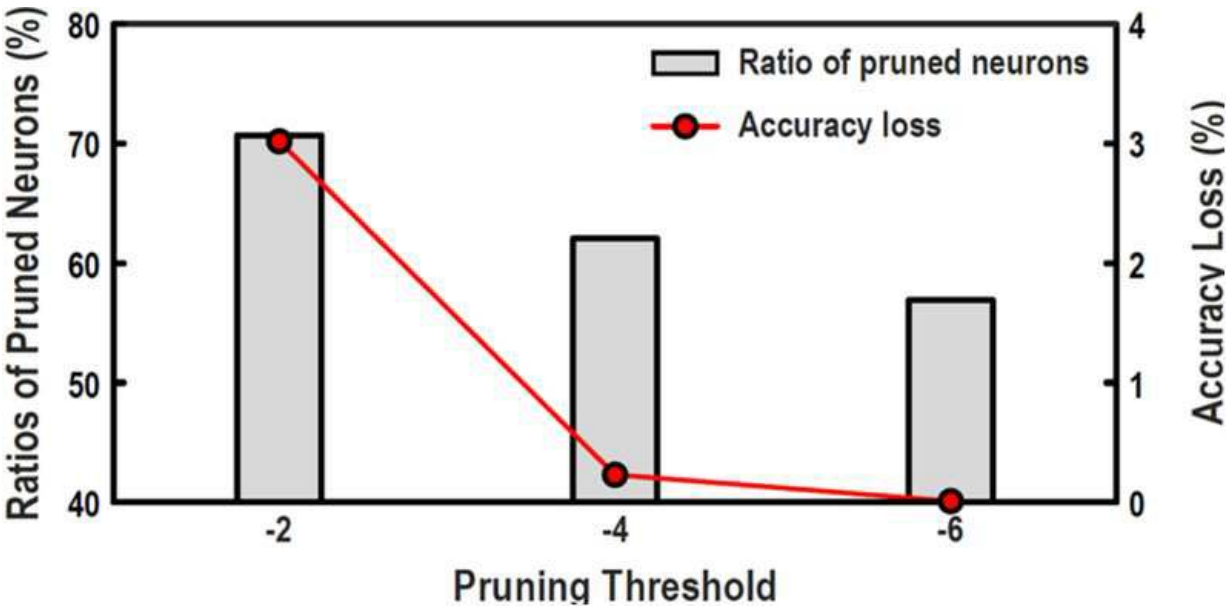
도면1b



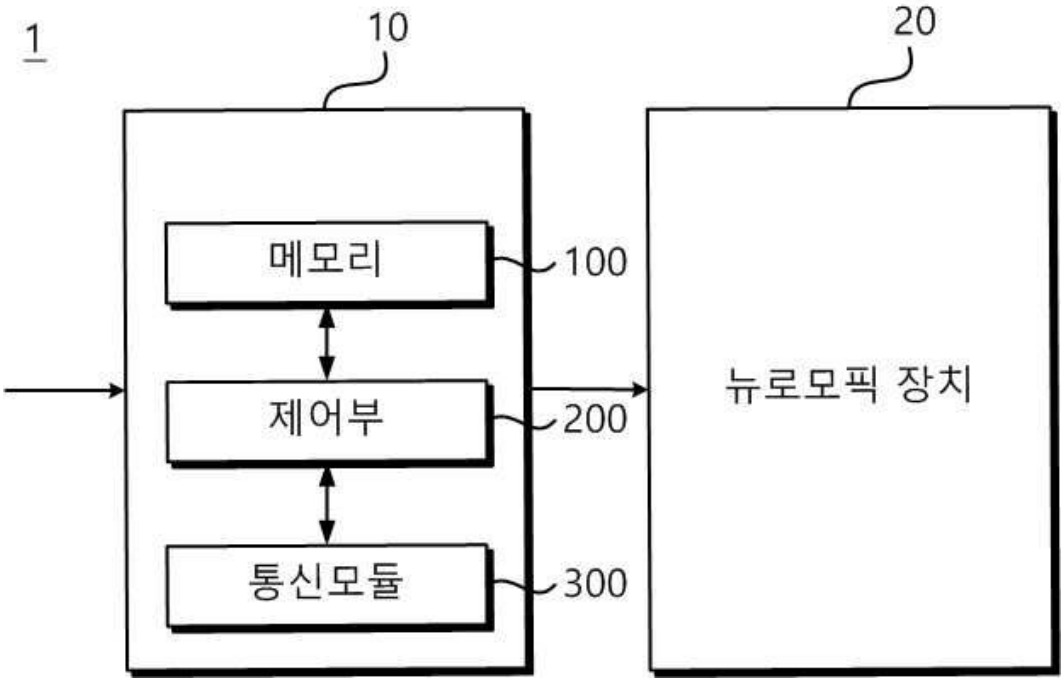
도면2a


V_{TH}
Regions of active neurons
Pruning threshold = -2
Pruning threshold = -4
Pruning threshold = -6
Membrane Voltage
Timesteps
2
0
-2
-4
-6
-8
0
10
20
30
40
50
60
70
80
90

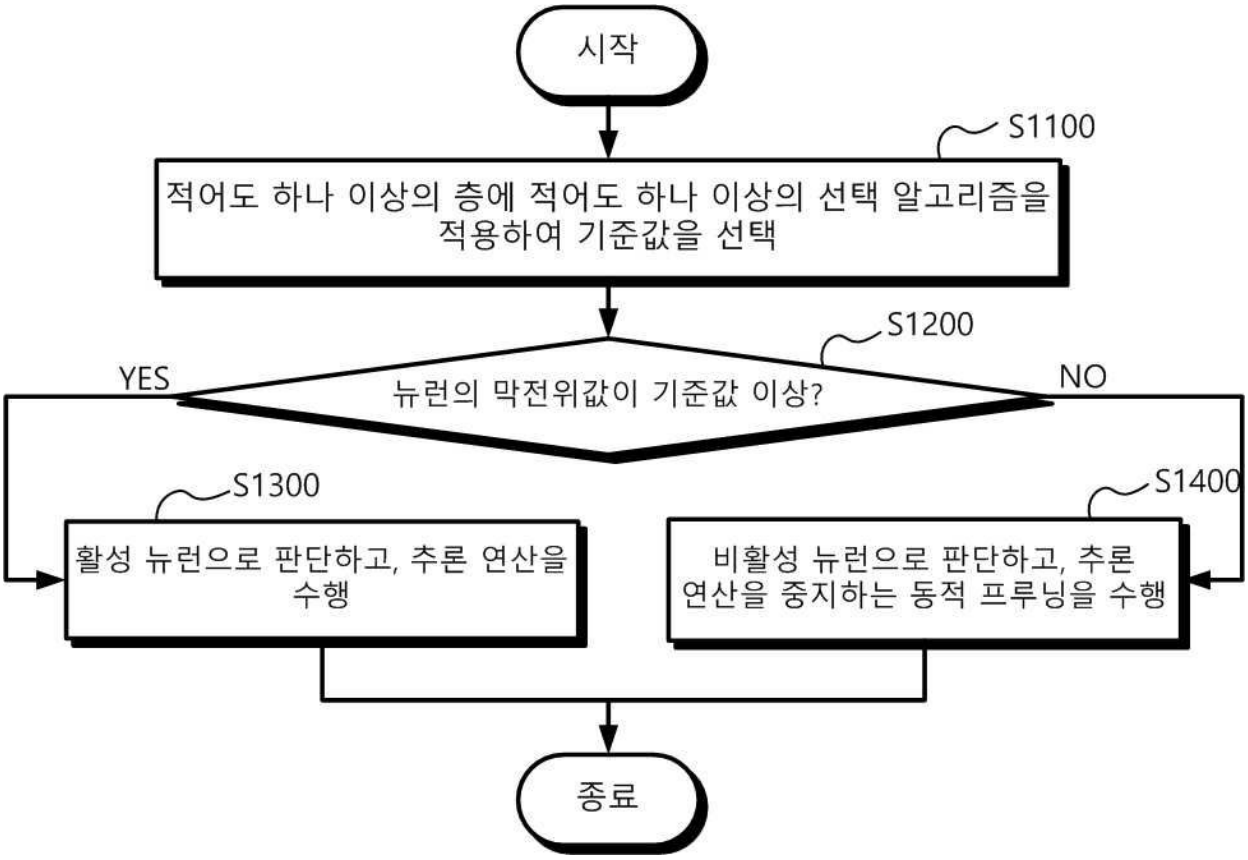
도면2b


Ratio of pruned neurons
Accuracy loss
Ratios of Pruned Neurons (%)
Accuracy Loss (%)
Pruning Threshold
80
70
60
50
40
4
3
2
1
0
-2
-4
-6

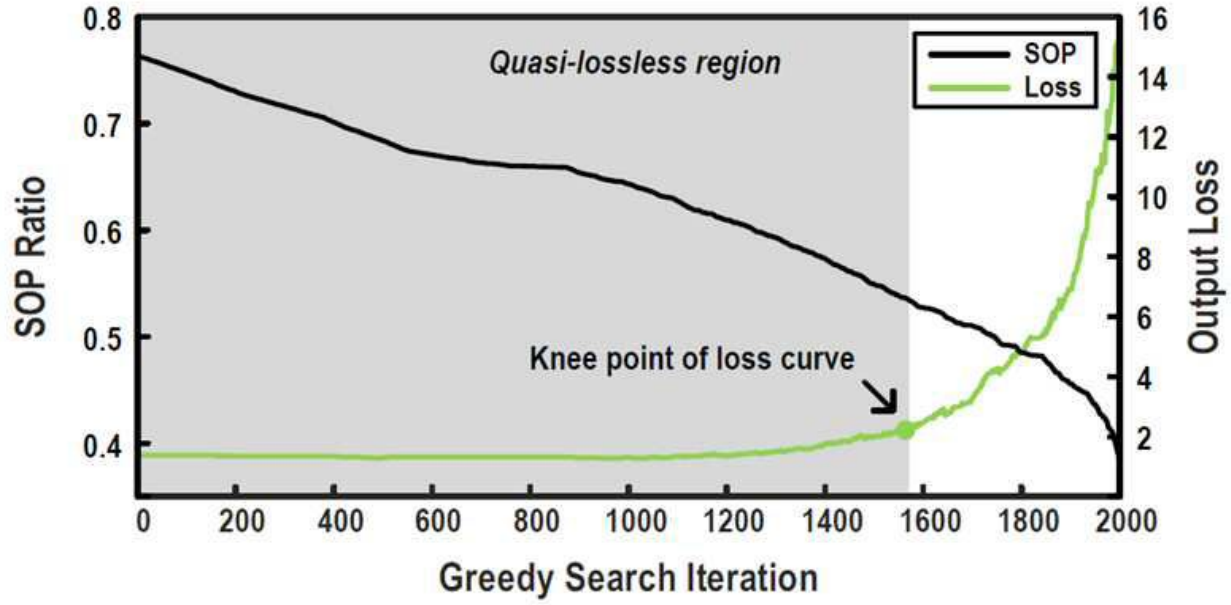
도면3


1
10
20
메모리
100
제어부
200
통신모듈
300
뉴로모픽 장치

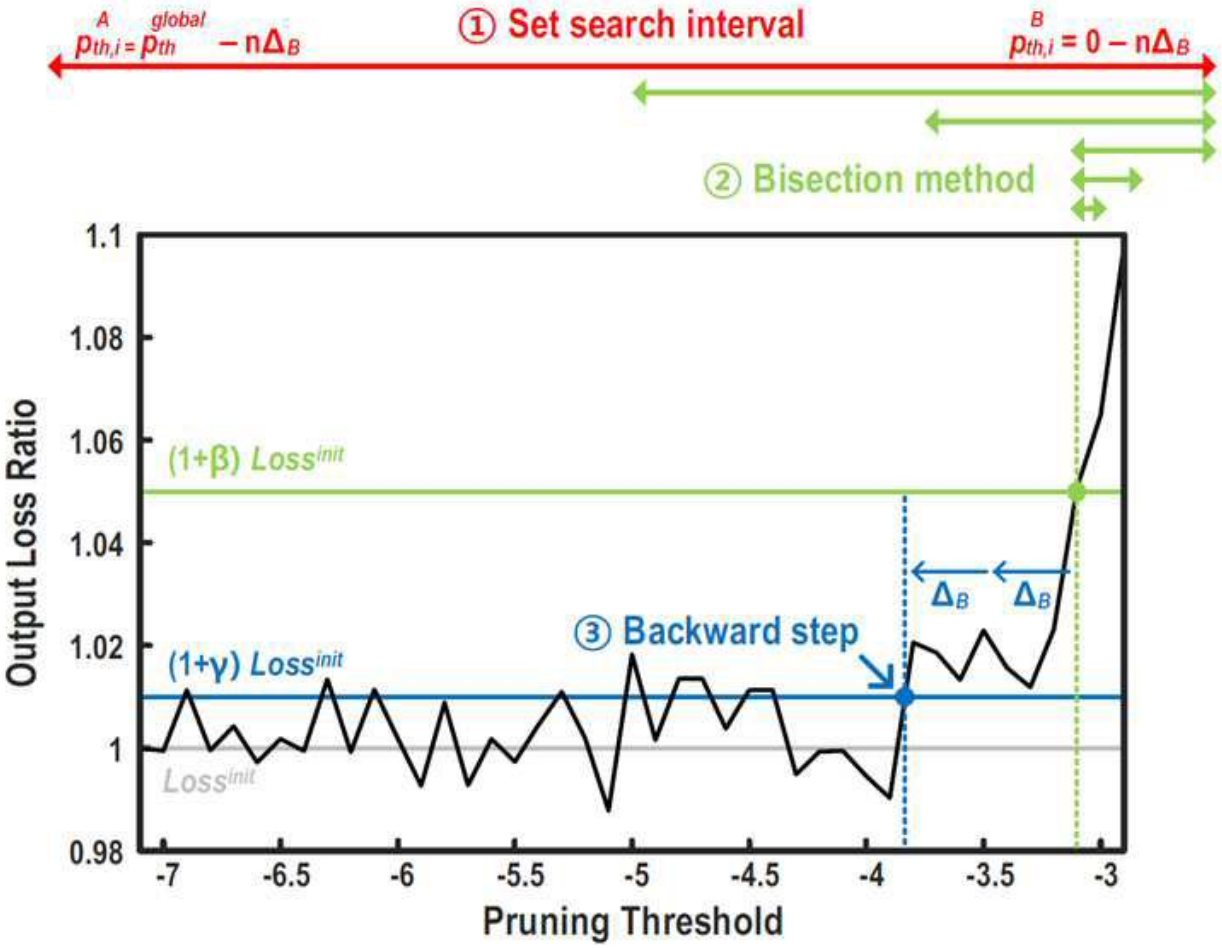
도면4


시작
S1100
적어도 하나 이상의 층에 적어도 하나 이상의 선택 알고리즘을 적용하여 기준값을 선택
S1200
YES
뉴런의 막전위값이 기준값 이상?
NO
S1300
활성 뉴런으로 판단하고, 추론 연산을 수행
S1400
비활성 뉴런으로 판단하고, 추론 연산을 중지하는 동적 프루닝을 수행
종료

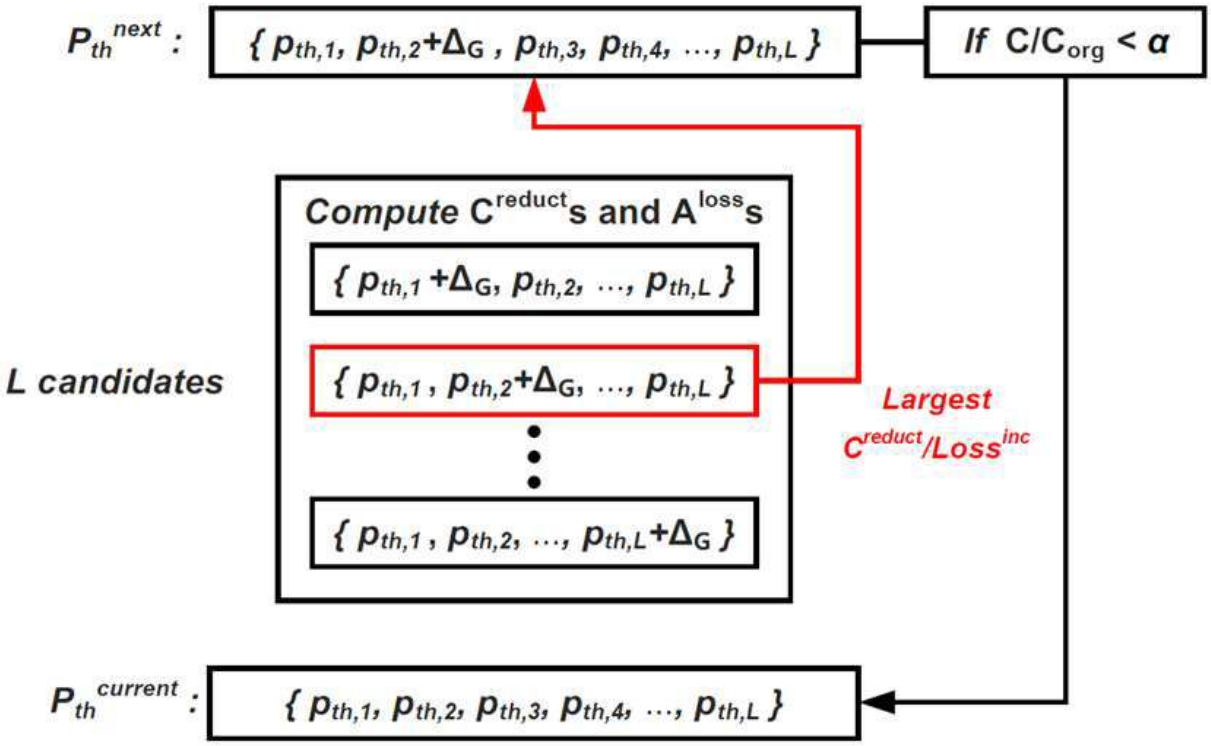
도면5a


Quasi-lossless region
SOP
Loss
Knee point of loss curve
SOP Ratio
Output Loss
Greedy Search Iteration
0.8
0.7
0.6
0.5
0.4
16
14
12
10
8
6
4
2
0
200
400
600
800
1000
1200
1400
1600
1800
2000

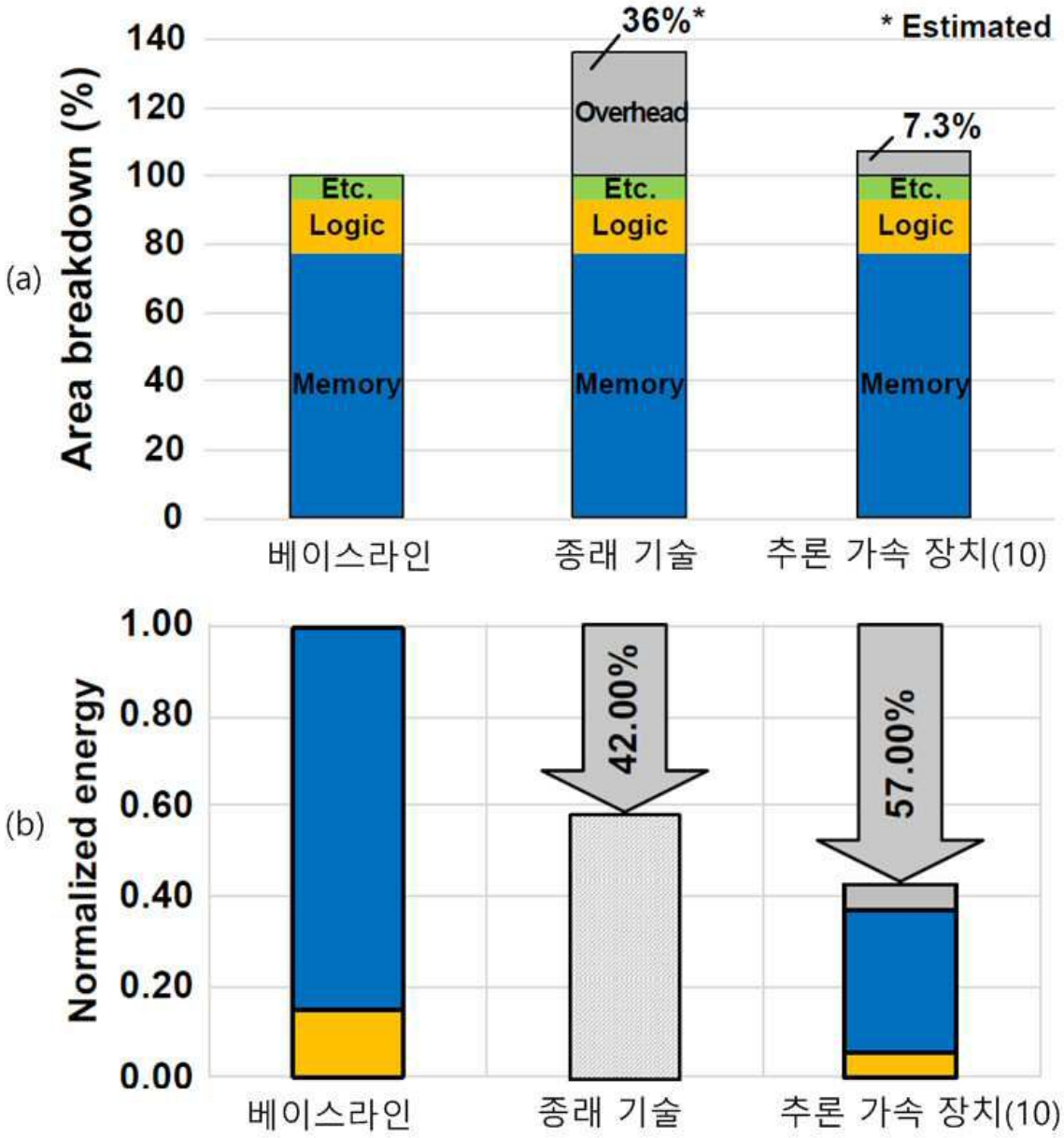
도면5b


① Set search interval
② Bisection method
③ Backward step
(1+β) Loss^init
(1+γ) Loss^init
Loss^init
Δ_B
Δ_B
Output Loss Ratio
Pruning Threshold
1.1
1.08
1.06
1.04
1.02
1
0.98
-7
-6.5
-6
-5.5
-5
-4.5
-4
-3.5
-3

도면6


P_{th}^{next} : { $p_{th,1}$, $p_{th,2}+\Delta_G$, $p_{th,3}$, $p_{th,4}$, …, $p_{th,L}$ }
If $C/C_{org} < \alpha$
Compute C^{reduct}s and A^{loss}s
{ $p_{th,1}+\Delta_G$, $p_{th,2}$, …, $p_{th,L}$ }
L candidates
{ $p_{th,1}$, $p_{th,2}+\Delta_G$, …, $p_{th,L}$ }
{ $p_{th,1}$, $p_{th,2}$, …, $p_{th,L}+\Delta_G$ }
Largest $C^{reduct}/Loss^{inc}$
$P_{th}^{current}$: { $p_{th,1}$, $p_{th,2}$, $p_{th,3}$, $p_{th,4}$, …, $p_{th,L}$ }

도면7


(a)
Area breakdown (%)
140
120
100
80
60
40
20
0
36%*
* Estimated
Overhead
7.3%
Etc.
Logic
Memory
베이스라인
종래 기술
추론 가속 장치(10)
(b)
Normalized energy
1.00
0.80
0.60
0.40
0.20
0.00
42.00%
57.00%
베이스라인
종래 기술
추론 가속 장치(10)