

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2024-0081604
(43) 공개일자 2024년06월10일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2023.01) G06F 17/13 (2006.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/049 (2023.01)
G06F 17/13 (2013.01)
(21) 출원번호 10-2022-0164253
(22) 출원일자 2022년11월30일
심사청구일자 2022년11월30일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대 학교)
(72) 발명자
서영교
경기도 부천시 길주로77번길 19-45, 920호(상동, 디아프갤러리4)
이연진
인천광역시 미추홀구 매소홀로 171, 105동 2102호(용현동, 신창미션힐)
최수름
인천광역시 연수구 선학로 100, 5동 1204호(선학동, 금호아파트)
(74) 대리인
양성보

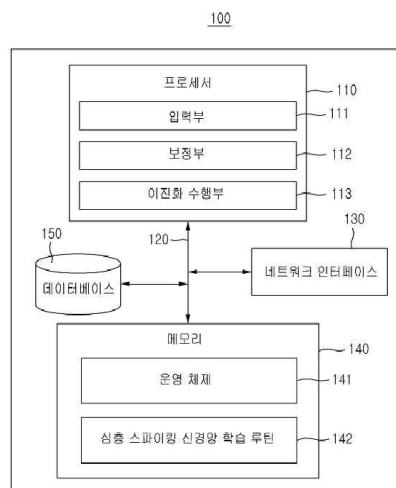
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 메모리 효율적인 심층 스파이킹 신경망

(57) 요약

메모리 효율적인 심층 스파이킹 신경망 시스템 및 그 동작 방법이 제시된다. 본 발명에서 제안하는 메모리 효율적인 심층 스파이킹 신경망 시스템은 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 입력부, 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시키는 보정부 및 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 이진화 수행부를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06N 3/084 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711173059
과제번호 2022M3I7A2085479
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 원천기술개발사업
연구과제명 [Ezbaro] (2022M3I7A2079155) 강유전체 메모리 기반 PIM 소자, 어레이, 구동회로 개

발

기 여 율 20/100
과제수행기관명 인하대학교
연구기간 2022.04.20 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1415179211
과제번호 20009972
부처명 산업통상자원부
과제관리(전문)기관명 한국산업기술평가관리원
연구사업명 차세대지능형반도체기술개발
연구과제명 고효율 초저전력 경량 엣지 디바이스용 소자회로 및 SoC 개발(3차년도)
기 여 율 20/100
과제수행기관명 한국전자기술연구원
연구기간 2022.01.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711152938
과제번호 2021-0-02052-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 대학ITRC센터육성지원사업(교육운영수익)
연구과제명 [Ezbaro][정부] 스마트 모빌리티를 위한 인공지능 시스템반도체 핵심 기술 개발 및

인력 양성(2차년도)

기 여 율 20/100
과제수행기관명 인하대학교 산학협력단
연구기간 2022.01.01 ~ 2022.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711169337
과제번호 2020R1F1A1051529
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 기본연구
연구과제명 [Ezbaro] CMOS 메모리와 차세대 메모리를 이용한 저전력 저면적 PIM (Processing In

Memory) 설계

기 여 율 20/100
과제수행기관명 인하대학교
연구기간 2022.03.01 ~ 2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711155747
과제번호 2021M3F3A2A01037531
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 신소자원천기술개발
연구과제명 [Ezbaro] Dynamic logic 기반 고성능 STMG 로직 회로 설계
기 여 율 20/100
과제수행기관명 인하대학교
연구기간 2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 입력부;

상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시키는 보정부; 및

상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 이진화 수행부

를 포함하는 심층 스파이킹 신경망 시스템.

청구항 2

제1항에 있어서,

상기 입력부는,

부동 소수점 입력값을 가지는 합성곱 신경망보다 이진화에 끼치는 영향을 감소시키도록 입력값이 이진화된 값으로 입력되는

심층 스파이킹 신경망 시스템.

청구항 3

제1항에 있어서,

상기 이진화 수행부는,

메모리 크기와 사용량을 감소시키고 연산 하드웨어의 경량화를 위해 가중치의 이진화를 적용하는

심층 스파이킹 신경망 시스템.

청구항 4

제3항에 있어서,

상기 이진화 수행부는,

딤러닝 모델을 경량화하기 위해 사인(sign) 함수를 통해 가중치를 +1 과 -1로 제한함으로써 산술 연산을 경량화된 비트 간의 연산으로 대체하는

심층 스파이킹 신경망 시스템.

청구항 5

입력부가 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 단계;

보정부가 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시키는 단계; 및

이진화 수행부가 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 단계

를 포함하는 심층 스파이킹 신경망의 동작 방법.

청구항 6

제5항에 있어서,

상기 입력부가 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 단계는,

부동 소수점 입력값을 가지는 합성곱 신경망보다 이진화에 끼치는 영향을 감소시키도록 입력값이 이진화된 값으로 입력되는

심층 스파이킹 신경망의 동작 방법.

청구항 7

제5항에 있어서,

상기 이진화 수행부가 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 단계는,

메모리 크기와 사용량을 감소시키고 연산 하드웨어의 경량화를 위해 가중치의 이진화를 적용하는

심층 스파이킹 신경망의 동작 방법.

청구항 8

제7항에 있어서,

상기 이진화 수행부가 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 단계는,

딥러닝 모델을 경량화하기 위해 사인(sign) 함수를 통해 가중치를 +1 과 -1로 제한함으로써 산술 연산을 경량화된 비트 간의 연산으로 대체하는

심층 스파이킹 신경망의 동작 방법.

발명의 설명

기술 분야

[0001] 본 발명은 메모리 효율적인 심층 스파이킹 신경망 시스템 및 그 동작 방법에 관한 것이다.

배경 기술

[0002] 기존의 딥러닝 모델은 목표 정확도에 도달하기까지 수많은 튜닝을 거쳐 학습 과정을 지속적으로 반복해야 하는 특징이 있다. 이로 인해 학습과 추론 과정에서 고비용의 전산 자원 및 많은 연산량을 필요로 하게 되며, 이에 따른 전력 소모량의 증가 및 학습 시간의 증가로 인해 개발 지연 등의 문제를 가지게 된다[1].

[0003] 더불어 최근의 임베디드 환경으로 인해 제한된 연산 능력, 메모리 전력 하에서 딥러닝 알고리즘 경량화의 필요성 또한 증가하고 있다[2].

[0004] 이에 대한 방안으로 제시된 XNOR-Net은 가중치를 이진화함으로써 저장 공간에 있어 적은 메모리를 차지하면서도 고속으로 추론이 가능하도록 한 효율적인 네트워크이다[3].

[0005] 하지만 XNOR-Net의 첫 번째 레이어의 가중치의 경우 정확도 하락의 문제로 이진화하지 못하여 극대화된 메모리 절약(Memory Saving)을 이루지 못하고 있다.

선행기술문헌

비특허문헌

[0006] (비특허문헌 0001) [1] E. H. Kim, K. H. Lee, and W. K. Seong, "딥러닝 모델의 경량화 기술 동향", Communications of KIISE, pp. 18-29. 2020.

(비특허문헌 0002) [2] Y. J. Lee, Y. H. Moon, J. Y. Park and O. G. Min, "Recent R&D Trends for Lightweight Deep Learning", Electronics and Telecommunications Trends, vol. 34, no. 2, pp. 40-50. 2019.

(비특허문헌 0003) [3] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnornet: Imagenet classification using binary convolutional neural networks", European Conference on Computer Vision,

pp. 525-542. 2016.

(비특허문헌 0004) [4] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks". International journal of neural systems, pp. 295-308. 2009.

(비특허문헌 0005) [5] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier, "STDP-based spiking deep convolutional neural networks for object recognition". Neural Netw., vol. 99, pp. 56-67. 2017.

(비특허문헌 0006) [6] C. K. Lee, SS. Sarwar, P. Panda and G. Srinivasan. "Enabling Spike-Based Backpropagation for Training Deep Neural Network Architectures", Front. Neurosci., February 2020. Transaction on Image Processing, vol. 10, no. 11, pp. 1670-1675. 2001.

발명의 내용

해결하려는 과제

[0007] 본 발명이 이루고자 하는 기술적 과제는 첫 번째 레이어의 가중치를 이진화하여 메모리 세이빙(Memory Saving)을 더욱 극대화 시키고, 정확도 하락을 감소시키기 위한 네트워크인 첫 번째 레이어 이진화 BSNN(1st layer binarized Binary Spiking Neural Network) 시스템 및 그 동작 방법을 제공하는데 있다.

과제의 해결 수단

[0008] 일 측면에 있어서, 본 발명에서 제안하는 메모리 효율적인 심층 스파이킹 신경망 시스템은 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 입력부, 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시키는 보정부 및 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 이진화 수행부를 포함한다.

[0009] 상기 입력부는 부동 소수점 입력값을 가지는 합성곱 신경망보다 이진화에 끼치는 영향을 감소시키도록 입력값이 이진화된 값으로 입력된다.

[0010] 상기 이진화 수행부는 메모리 크기와 사용량을 감소시키고 연산 하드웨어의 경량화를 위해 가중치의 이진화를 적용한다.

[0011] 상기 이진화 수행부는 딥러닝 모델을 경량화하기 위해 사인(sign) 함수를 통해 가중치를 +1 과 -1로 제한함으로써 산술 연산을 경량화된 비트 간의 연산으로 대체한다.

[0012] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 메모리 효율적인 심층 스파이킹 신경망의 동작 방법은 입력부가 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 단계, 보정부가 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시키는 단계 및 이진화 수행부가 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 단계를 포함한다.

발명의 효과

[0013] 본 발명의 실시예들에 따르면 첫 번째 레이어 이진화 BSNN(1st layer binarized Binary Spiking Neural Network) 시스템 및 그 동작 방법을 통해 첫 번째 레이어의 가중치를 이진화하여 메모리 세이빙(Memory Saving)을 더욱 극대화 시키고, 정확도 하락을 감소시킬 수 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시예에 따른 메모리 효율적인 심층 스파이킹 신경망 시스템의 구성을 나타내는 도면이다.

도 2는 본 발명의 일 실시예에 따른 메모리 효율적인 심층 스파이킹 신경망의 동작 방법을 설명하기 위한 흐름도이다.

도 3은 본 발명의 일 실시예에 따른 첫 번째 레이어 이진화 BSNN 및 종래기술의 정확도 및 메모리 세이빙을 비교하는 표이다.

발명을 실시하기 위한 구체적인 내용

- [0015] 최근의 임베디드 환경으로 인해 제한된 연산 능력, 메모리 전력 하에서 딥러닝 모델 추론이 요구되고 있으며, 경량화에 집중한 여러 뉴럴 네트워크들이 제안되고 있다. 경량화 네트워크인 XNOR-Net은 가중치 이진화를 통해 네트워크 경량화를 이루어내었지만 정확도 하락 등의 문제로 첫 번째 레이어의 가중치를 이진화하지 못함으로써 추가적인 메모리 세이빙(Memory Saving)을 이루어내지 못하고 있다. 따라서 본 발명에서는 첫 번째 레이어의 가중치를 이진화함으로써 추가적인 메모리 세이빙을 이루어 낼 뿐 아니라 정확도도 XNOR-Net보다 더 높은 첫 번째 레이어 이진화 BSNN(1st layer binarized Binary Spiking Neural Network)을 제안하고자 한다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.
- [0017] 도 1은 본 발명의 일 실시예에 따른 메모리 효율적인 심층 스파이킹 신경망 시스템의 구성을 나타내는 도면이다.
- [0018] 본 실시예에 따른 심층 스파이킹 신경망 시스템(100)은 프로세서(110), 버스(120), 네트워크 인터페이스(130), 메모리(140) 및 데이터베이스(150)를 포함할 수 있다. 메모리(140)는 운영체제(141) 및 메모리 효율적인 심층 스파이킹 신경망 학습 루틴(142)을 포함할 수 있다. 프로세서(110)는 입력부(111), 보정부(112) 및 이진화 수행부(113)를 포함할 수 있다. 다른 실시예들에서 심층 스파이킹 신경망 시스템(100)은 도 1의 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, 심층 스파이킹 신경망 시스템(100)은 디스플레이나 트랜시버(transceiver)와 같은 다른 구성요소들을 포함할 수도 있다.
- [0019] 메모리(140)는 컴퓨터에서 판독 가능한 기록 매체로서, RAM(random access memory), ROM(read only memory) 및 디스크 드라이브와 같은 비소멸성 대용량 기록장치(permanent mass storage device)를 포함할 수 있다. 또한, 메모리(140)에는 운영체제(141)와 메모리 효율적인 심층 스파이킹 신경망 학습 루틴(142)을 위한 프로그램 코드가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 드라이브 메커니즘(drive mechanism, 미도시)을 이용하여 메모리(140)와는 별도의 컴퓨터에서 판독 가능한 기록 매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록 매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록 매체(미도시)를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록 매체가 아닌 네트워크 인터페이스(130)를 통해 메모리(140)에 로딩될 수도 있다.
- [0020] 버스(120)는 심층 스파이킹 신경망 시스템(100)의 구성요소들간의 통신 및 데이터 전송을 가능하게 할 수 있다. 버스(120)는 고속 시리얼 버스(high-speed serial bus), 병렬 버스(parallel bus), SAN(Storage Area Network) 및/또는 다른 적절한 통신 기술을 이용하여 구성될 수 있다.
- [0021] 네트워크 인터페이스(130)는 심층 스파이킹 신경망 시스템(100)을 컴퓨터 네트워크에 연결하기 위한 컴퓨터 하드웨어 구성요소일 수 있다. 네트워크 인터페이스(130)는 심층 스파이킹 신경망 시스템(100)을 무선 또는 유선 커넥션을 통해 컴퓨터 네트워크에 연결시킬 수 있다.
- [0022] 데이터베이스(150)는 메모리 효율적인 심층 스파이킹 신경망 학습을 위해 필요한 모든 정보를 저장 및 유지하는 역할을 할 수 있다. 도 1에서는 심층 스파이킹 신경망 시스템(100)의 내부에 데이터베이스(150)를 구축하여 포함하는 것으로 도시하고 있으나, 이에 한정되는 것은 아니며 시스템 구현 방식이나 환경 등에 따라 생략될 수 있고 혹은 전체 또는 일부의 데이터베이스가 별개의 다른 시스템 상에 구축된 외부 데이터베이스로서 존재하는 것 또한 가능하다.
- [0023] 프로세서(110)는 기본적인 산술, 로직 및 심층 스파이킹 신경망 시스템(100)의 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(140) 또는 네트워크 인터페이스(130)에 의해, 그리고 버스(120)를 통해 프로세서(110)로 제공될 수 있다. 프로세서(110)는 입력부(111), 보정부(112) 및 이진화 수행부(113)를 위한 프로그램 코드를 실행하도록 구성될 수 있다. 이러한 프로그램 코드는 메모리(140)와 같은 기록 장치에 저장될 수 있다.
- [0024] 입력부(111), 보정부(112) 및 이진화 수행부(113)는 도 2의 단계들(210~230)을 수행하기 위해 구성될 수 있다.

- [0025] 심층 스파이킹 신경망 시스템(100)은 입력부(111), 보정부(112) 및 이진화 수행부(113)를 포함할 수 있다.
- [0026] 입력부(111)는 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는다.
- [0027] 본 발명의 실시예에 따르면, 합성곱 신경망(Convolution Neural Network; CNN)이 아닌 스파이킹 기반 신경망(Spiking Neural Network; SNN)을 사용하였다. 스파이킹 기반 신경망은 인간의 뇌의 동작 방식을 모방한 네트워크로 네트워크의 입력은 일련의 푸아송 분포를 따르는 스파이크 (0,1) 형태로 전달 되게 된다[4]. 즉, 입력값이 합성곱 신경망과 다르게 이진화된 값으로 들어오므로 부동 소수점 입력값을 가지는 합성곱 신경망보다 이진화에 영향을 덜 끼치게 된다.
- [0028] 보정부(112)는 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시킨다.
- [0029] 일반적으로 스파이킹 신경망은 활성화 함수의 불연속성으로 인해 역전파를 통한 신경망 학습이 불가능하여 STDP 학습방식을 사용한다[5]. 하지만 본 발명에서 제안한 네트워크는 식(1)에서 표현한 LIF 근사 미분 방정식을 통해 역전파 기법 기반의 신경망 학습을 사용하여 높은 정확성을 가질 수 있도록 하였다[6].
- $$\frac{\partial a_{IF}}{\partial net} \approx \frac{1}{V_{th}} \cdot 1 = \frac{1}{V_{th}} \quad (1)$$
- [0030]
- [0031] 이진화 수행부(113)는 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화한다.
- [0032] 본 발명의 실시예에 따르면, 메모리 크기와 사용량을 줄이고 연산 하드웨어의 경량화를 위해서 가중치의 이진화를 적용했다.
- $$W_F * s \approx (\alpha * W_B) * s, W_B = \text{sign}(W) \quad (2)$$
- [0033]
- [0034] 식(2)에서 표현한 사인(sign) 함수를 통해 가중치를 +1 과 -1로 제한하여 부동 소수점 기반의 기존 신경망과 달리 산술 연산을 경량화된 비트 간의 연산으로 대체하여 딥러닝 모델을 경량화하였다. 따라서 이진화된 역전파 기반의 스파이킹 신경망을 통해서 적은 메모리를 사용하면서 딥러닝 모델의 정확성 또한 XNOR-Net보다 크며, 부동 소수점 기반의 기존 합성곱 신경망에 근접한 매우 높은 정확성을 가질 수 있다.
- [0036] 도 2는 본 발명의 일 실시예에 따른 메모리 효율적인 심층 스파이킹 신경망의 동작 방법을 설명하기 위한 흐름도이다.
- [0037] 제안하는 메모리 효율적인 심층 스파이킹 신경망의 동작 방법은 입력부가 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는 단계(210), 보정부가 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시키는 단계(220) 및 이진화 수행부가 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화하는 단계(230)를 포함한다.
- [0038] 단계(210)에서, 입력부가 인간의 뇌의 동작 방식을 모방한 네트워크인 스파이킹 신경망을 통해 일련의 푸아송(Poisson) 분포를 따르는 스파이크 형태의 입력을 받는다.
- [0039] 본 발명의 실시예에 따르면, 합성곱 신경망(Convolution Neural Network; CNN)이 아닌 스파이킹 기반 신경망(Spiking Neural Network; SNN)을 사용하였다. 스파이킹 기반 신경망은 인간의 뇌의 동작 방식을 모방한 네트워크로 네트워크의 입력은 일련의 푸아송 분포를 따르는 스파이크 (0,1) 형태로 전달 되게 된다.
- [0040] 단계(220)에서, 보정부가 상기 스파이킹 신경망의 불연속성으로 인해 불가능한 역전파의 학습 방식을 LIF 근사 미분 방정식을 통해 가능하게 하여 정확도를 향상 시킨다.
- [0041] 본 발명의 실시예에 따르면, 네트워크는 식(1)에서 표현한 LIF 근사 미분 방정식을 통해 역전파 기법 기반의 신경망 학습을 사용하여 높은 정확성을 가질 수 있도록 하였다.
- [0042] 단계(230)에서, 이진화 수행부가 상기 스파이킹 신경망의 은닉층과 첫 번째 레이어의 가중치들을 이진화한다.
- [0043] 본 발명의 실시예에 따르면, 메모리 크기와 사용량을 줄이고 연산 하드웨어의 경량화를 위해서 가중치의 이진화

를 적용했다.

[0044] 사인(sign) 함수를 통해 가중치를 +1 과 -1로 제한하여 부동 소수점 기반의 기존 신경망과 달리 산술 연산을 경량화된 비트 간의 연산으로 대체하여 딥러닝 모델을 경량화하였다.

[0046] 도 3은 본 발명의 일 실시예에 따른 첫 번째 레이어 이진화 BSNN 및 종래기술의 정확도 및 메모리 세이빙을 비교하는 표이다.

[0047] 제안한 첫 번째 레이어의 가중치가 이진화된 심층 스파이킹 신경망은 VGG8 네트워크에서 훈련하였고 표준 이미지 데이터 셋인 CIFAR-10 데이터셋을 이 용하여 진행하였다. 비교군인 XNOR-Net은 추가적인 하드웨어와 메모리를 필요로 하는 배치 정규화 계층을 없앤 네트워크로 실험을 진행하였다. 실험 결과 표준 합성곱 신경망(Standard CNN)에서는 91.8%의 정확도를 보였다. 이에 반해 1st layer binarized XNOR-Net은 첫 번째 레이어의 가중치의 이진화로 인해 정확도가 크게 하락한 70.1%의 분류 정확도를 보였다. 이는 표준 합성곱 신경망보다 크게 정확성이 감소한 값이다. 하지만 가중치의 이진화로 인해 표준 합성곱 신경망보다 30.9배 높은 메모리 세이빙을 가진다. 이와 비교해 본 발명에서 제안하는 첫 번째 레이어 이진화 BSNN은 89.1%라는 높은 정확성과 30.9x라는 높은 메모리 세이빙을 동시에 가진다. 이는 첫 번째 레이어의 가중치의 이진화로인한 정확도 하락이 일어나지 않아 부동 소수점을 사용하는 표준 합성곱 신경망과 비교해서도 매우 근접한 높은 정확성을 가지며, 경량화된 네트워크인 1st layer binarized XNOR-NET과도 동일한 메모리 세이빙을 가지고 있다.

[0048] 이와 같이, 본 발명에서는 첫 번째 레이어의 가중치가 이진화된 스파이킹 신경망을 제안하였다. 합성곱 신경망이 아닌 스파이크 기반 신경망을 사용하며, 가중치의 이진화, LIF 근사 방정식을 통한 스파이킹 신경망의 역전과 학습이 가능하도록 하여 경량화와 동시에 높은 정확도를 유지할 수 있도록 하였다. 실험 결과 1st layer binarized XNOR-Net과 달리 정확도 하락이 크지 않았고 표준 합성곱 신경망에 비해서 30.9배 높은 메모리 세이빙을 달성하였다.

[0050] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/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 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 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0051] 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, 가상장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0052] 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. 상기 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

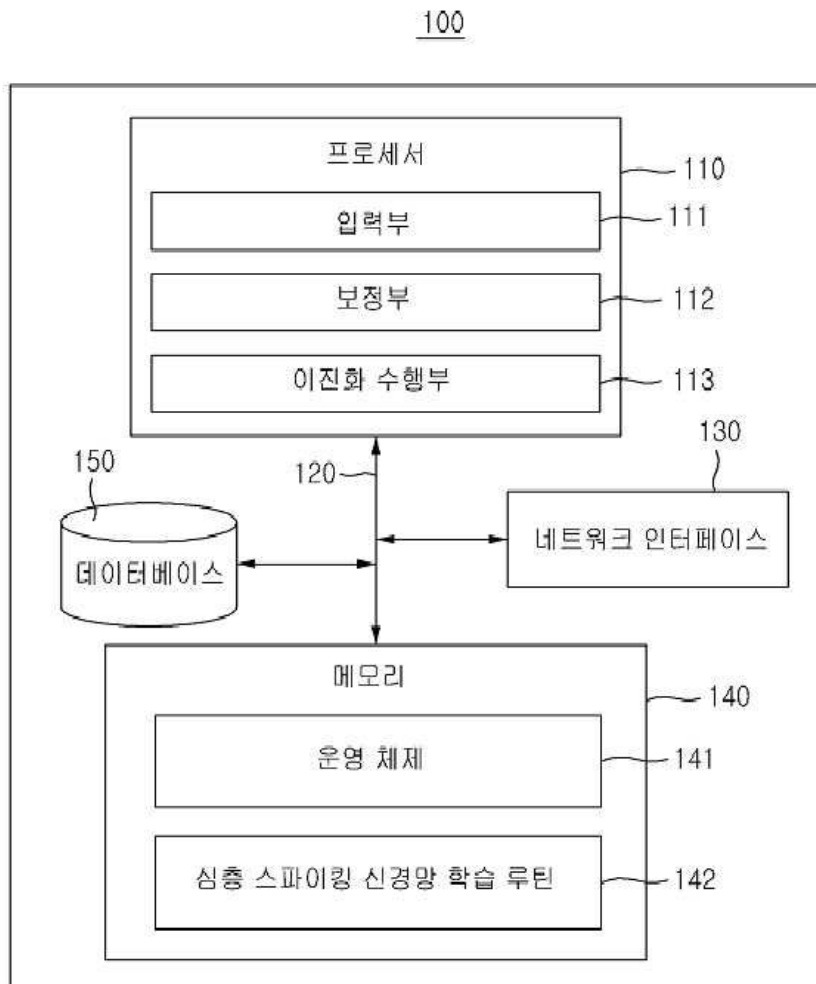
(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0053] 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.

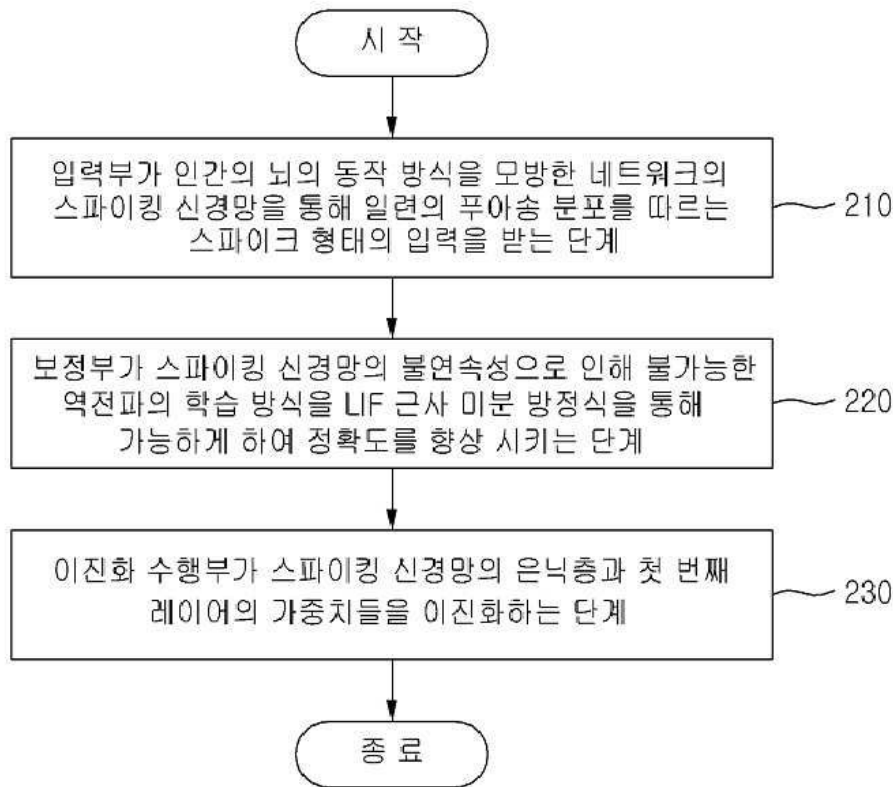
[0054] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

도면1



도면2



도면3

Network	Accuracy (CIFAR-10)	Memory Saving
Standard CNN (without Batchnorm)	91.8%	1x
1st layer binarized XNOR-Net (without Batchnorm)	70.2%	30.9x
This work (1st layer binarized BSNN)	89.1%	30.9x