



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0010383
(43) 공개일자 2022년01월25일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) G06N 3/04 (2006.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2020-0089210
(22) 출원일자 2020년07월17일
심사청구일자 2020년07월17일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암
동5가)
(72) 발명자
박중선
서울특별시 서초구 신반포로 9 반포아파트
신동엽
서울특별시 강서구 초록마을로26길 49-11 두산빌
라
(뒷면에 계속)
(74) 대리인
이기성

전체 청구항 수 : 총 20 항

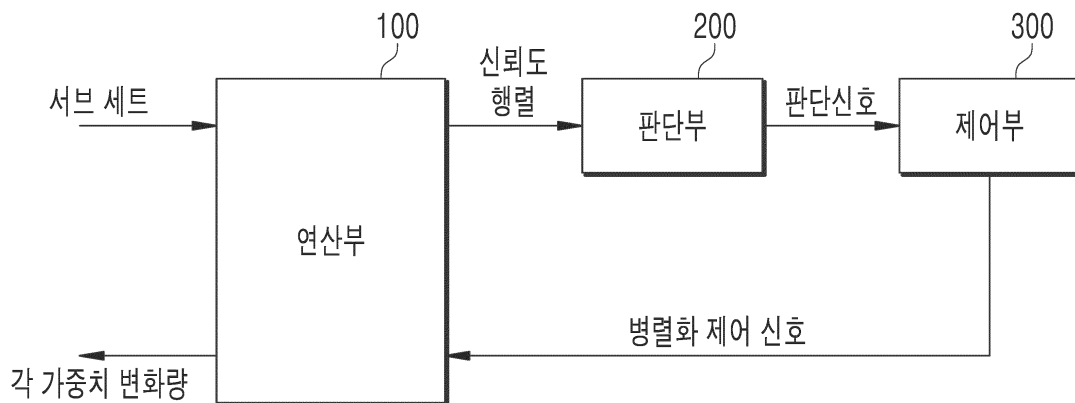
(54) 발명의 명칭 심층 신경망 학습 가속 장치 및 그 동작 방법

(57) 요약

본 출원의 실시예에 따른 심층 신경망 학습 가속 장치는 미니-배치 경사 하강법에 따라, 서브 세트의 복수의 입력 데이터들에 대한 제1 및 제2 연산을 순차적으로 수행하는 연산부, 상기 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 상기 복수의 입력 데이터들 각각을 스킵 데이터와 학습 데이터 중 어느 하나로 판단하는 판단부 및 상기 스킵 데이터에 대한 상기 제2 연산을 스킵시키도록 상기 연산부를 제어하는 제어부를 포함한다.

대표도 - 도1

10



(52) CPC특허분류

G06N 3/084 (2013.01)

(72) 발명자

김건호

경기도 군포시 고산로643번길 10 신안모란아파트

조중호

서울특별시 동대문구 무학로 193

이 발명을 지원한 국가연구개발사업

과제고유번호	1711115231
과제번호	2020R1A2C3014820
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	온-디바이스 개인화를 위한 에너지 효율적인 딥러닝 가속기 설계
기 여 율	60/100
과제수행기관명	고려대학교
연구기간	2020.03.01 ~ 2021.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116203
과제번호	10077445
부처명	과학기술정보통신부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	전자정보디바이스산업원천기술개발(R&D)
연구과제명	스마트 모바일 및 IoT 디바이스를 위한 뉴럴셀(Spiking Neural Cell)기반 SNP SoC
원천기술 개발	
기 여 율	25/100
과제수행기관명	한국전자기술연구원
연구기간	2020.01.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116159
과제번호	2018-0-01433-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	스마트 농축산 IoA(Internet of Agriculture) 연구센터
기 여 율	15/100
과제수행기관명	국립강원대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

명세서

청구범위

청구항 1

미니-배치 경사 하강법에 따라, 서버 세트의 복수의 입력 데이터들에 대한 제1 및 제2 연산을 순차적으로 수행하는 연산부;

상기 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 상기 복수의 입력 데이터들 각각을 스킵 데이터와 학습 데이터 중 어느 하나로 판단하는 판단부; 및

상기 스킵 데이터에 대한 상기 제2 연산을 스킵시키도록 상기 연산부를 제어하는 제어부를 포함하는, 심층 신경망 학습 가속 장치.

청구항 2

제1항에 있어서,

상기 연산부는 상기 제1 연산으로부터 일정 시간 이후에, 상기 학습 데이터에 대한 상기 제2 연산을 수행하는, 심층 신경망 학습 가속 장치.

청구항 3

제1항에 있어서,

상기 제1 연산은 순전파 알고리즘을 이용하는 상기 미니-배치 경사 하강법의 제1 학습 단계인, 심층 신경망 학습 가속 장치.

청구항 4

제1항에 있어서,

상기 제2 연산은 역전파 알고리즘 및 가중치 업데이트 알고리즘을 순차적으로 이용하는 상기 미니-배치 경사 하강법의 제2 학습 단계인, 심층 신경망 학습 가속 장치.

청구항 5

제1항에 있어서,

상기 판단부는 상기 신뢰도 행렬의 최대 크기 요소와 기설정된 임계값 사이의 대소 여부를 비교하는 비교기로 구현되는, 심층 신경망 학습 가속 장치.

청구항 6

제5항에 있어서,

상기 비교기는 상기 최대 크기 요소의 크기가 기설정된 임계값 이상인 경우, 상기 스킵 데이터에 대응되는 로우 신호를 상기 제어부로 출력하는, 심층 신경망 학습 가속 장치.

청구항 7

제5항에 있어서,

상기 비교기는 상기 신뢰도 행렬의 최대 크기 요소가 기설정된 임계값 미만인 경우, 상기 학습 데이터에 대응되는 하이 신호를 상기 제어부로 출력하는, 심층 신경망 학습 가속 장치.

청구항 8

제6항에 있어서,

상기 제어부는 상기 로우 신호에 기초하여, 상기 학습 데이터에 대한 상기 제2 연산을 병렬화시키는, 심층 신경망 학습 가속 장치.

청구항 9

제6항에 있어서,

상기 로우 신호의 개수는 상기 제2 연산의 동작 시간에 반비례하는 관계인, 심층 신경망 학습 가속 장치.

청구항 10

제1항에 있어서,

전체 입력 데이터들로부터 임의로 선택된 상기 복수의 입력 데이터들 각각을 상기 연산부에 할당시키는 입력부; 및

상기 연산부를 통해 출력받는 각 가중치 변화량을 합하여, 상기 서브 세트에 대한 기울기에 대응되는 출력 가중치 변화량을 출력하는 출력부를 더 포함하는, 심층 신경망 학습 가속 장치.

청구항 11

제1항에 있어서,

상기 연산부는 시스틀릭 어레이 구조로 형성되고, 상기 제1 및 제2 연산을 순차적으로 수행하는 복수의 연산장치들을 포함하는, 심층 신경망 학습 가속 장치.

청구항 12

제11항에 있어서,

상기 연산부는 상기 제어부로부터 출력받는 병렬화 제어 신호에 응답하여, 상기 복수의 연산장치들 중 상기 스킵 데이터에 대응되는 어느 하나의 연산장치를 초기화하는, 심층 신경망 학습 가속 장치.

청구항 13

제12항에 있어서,

상기 연산부는 상기 복수의 연산장치들 중 나머지 연산장치에 할당된 학습 데이터의 일부를 상기 어느 하나의 연산장치에 재할당하는, 심층 신경망 학습 가속 장치.

청구항 14

제11항에 있어서,

상기 제어부는 각 학습 데이터로부터 분할되는 복수의 서브 데이터들을 데이터 플로우에 따라 복수의 연산장치들에 재할당시키는, 심층 신경망 학습 가속 장치.

청구항 15

제14항에 있어서,

상기 데이터 플로우는 데이터를 리드하고 저장하기 위한 데이터의 이동 경로를 나타내는, 심층 신경망 학습 가속 장치.

청구항 16

심층 신경망 학습 가속 장치의 동작 방법으로서,

연산부가 미니-배치 경사 하강법에 따라, 서브 세트의 복수의 입력 데이터들에 대한 제1 연산을 수행하는 단계;

판단부가 상기 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 상기 복수의 입력 데이터들을 스킵 데이터와 학습 데이터 중 하나로 각각 판단하는 단계;

제어부가 상기 스킵 데이터에 응답하여, 상기 스킵 데이터에 대한 제2 연산을 스킵시키기 위한 병렬화 제어 신

호를 출력하는 단계; 및

상기 연산부가 상기 병렬화 제어 신호에 기초하여, 상기 스킵 데이터에 대한 상기 제2 연산을 스킵하고, 상기 학습 데이터에 대한 상기 제2 연산을 수행하는 단계를 포함하는, 심층 신경망 학습 가속 장치의 동작 방법.

청구항 17

제16항에 있어서,

상기 제1 연산은 순전파 알고리즘을 이용하는 상기 미니-배치 경사 하강법의 제1 학습 단계인, 심층 신경망 학습 가속 장치의 동작 방법.

청구항 18

제16항에 있어서,

상기 제2 연산은 역전파 알고리즘 및 가중치 업데이트 알고리즘을 순차적으로 이용하는 상기 미니-배치 경사 하강법의 제2 학습 단계인, 심층 신경망 학습 가속 장치의 동작 방법.

청구항 19

제16항에 있어서,

상기 판단부는 상기 신뢰도 행렬의 최대 크기 요소와 기설정된 임계값 사이의 대소 여부를 비교하는 비교기로 구현되는, 심층 신경망 학습 가속 장치의 동작 방법.

청구항 20

제16항에 있어서,

상기 제2 연산을 수행하는 단계는 상기 연산부가 상기 병렬화 제어 신호에 기초하여, 복수의 연산장치들 중 상기 스킵 데이터에 대응되는 어느 하나의 연산장치를 초기화하는 단계;

상기 연산부가 나머지 연산장치에 할당된 상기 학습 데이터의 각 일부를 상기 어느 하나의 연산장치에 재할당하는 단계; 및

상기 연산부가 상기 제1 연산으로부터 일정 시간 이후에 상기 복수의 연산장치들을 통해 상기 학습 데이터에 대한 상기 제2 연산을 병렬로 처리하는 단계를 포함하는, 심층 신경망 학습 가속 장치의 동작 방법.

발명의 설명

기술 분야

[0001] 본 출원은 심층 신경망 학습 가속 장치 및 그 동작 방법에 관한 것으로서, 특히, 미니 배치 경사 하강법에 따른 예측 신뢰도에 기초하는 심층 신경망 학습 가속 장치 및 그 동작 방법에 관한 것이다.

배경 기술

[0002] 심층 신경망(Deep Neural Network, DNN)은 많은 파라미터와 연산량을 기반으로 이미지 인식/분류, 객체 탐지 등 여러 응용 분야에서 최고의 성능을 보이고 있다.

[0003] 그러나, 심층 신경망(DNN)의 수많은 파라미터를 학습시키기 위해서는 많은 연산량이 필요하기 때문에, 심층 신경망(DNN)의 학습에는 일/주 단위의 긴 시간을 필요로 하기도 한다. 이러한 심층 신경망(DNN)의 학습 시간 및 학습 에너지 소모를 줄이기 위해서는, 학습에 필요한 연산량 자체를 줄이는 것이 효과적이다.

[0004] 심층 신경망(DNN)의 학습에는 일반적으로 미니 배치 경사 하강법(Mini-batch Gradient Descent)에 기반한 심층 신경망(DNN)의 학습 방법이 사용된다. 여기서, 미니 배치 경사 하강법은 본질적으로 노이즈(Noise)를 포함하기 때문에, 학습에 필요한 연산을 반드시 정확하게 계산할 필요가 없이 근사하는 것이 가능하다.

[0005] 그러나, 미니 배치 경사 하강법에 기반한 심층 신경망(DNN)의 학습 방법 또한, 긴 학습 시간과 많은 학습 에너지

지를 필요로 하기 때문에, 학습에 중요한 연산과 그렇지 않은 연산을 잘 구분하여, 상대적으로 중요하지 않은 연산에 대해 효과적인 근사를 적용할 수 있는 방안이 필요하다.

발명의 내용

해결하려는 과제

[0006] 본 출원의 목적은 학습 속도를 증가시키고, 학습 에너지를 감소시킬 수 있는 심층 신경망 학습 가속 장치 및 그 동작 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0007] 본 출원의 실시예에 따른 심층 신경망 학습 가속 장치는 미니-배치 경사 하강법에 따라, 서버 세트의 복수의 입력 데이터들에 대한 제1 및 제2 연산을 순차적으로 수행하는 연산부, 상기 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 상기 복수의 입력 데이터들 각각을 스킵 데이터와 학습 데이터 중 어느 하나로 판단하는 판단부 및 상기 스킵 데이터에 대한 상기 제2 연산을 스킵시키도록 상기 연산부를 제어하는 제어부를 포함한다.

[0008] 실시예에 따라, 상기 연산부는 상기 제1 연산으로부터 일정 시간 이후에, 상기 학습 데이터에 대한 상기 제2 연산을 수행한다.

[0009] 실시예에 따라, 상기 제1 연산은 순전파 알고리즘(FP)을 이용하는 상기 미니-배치 경사 하강법의 제1 학습 단계이다.

[0010] 실시예에 따라, 상기 제2 연산은 역전파 알고리즘 및 가중치 업데이트 알고리즘을 순차적으로 이용하는 상기 미니-배치 경사 하강법의 제2 학습 단계이다.

[0011] 실시예에 따라, 상기 판단부는 상기 신뢰도 행렬의 최대 크기 요소와 기설정된 임계값 사이의 대소 여부를 비교하는 비교기로 구현된다.

[0012] 실시예에 따라, 상기 비교기는 상기 최대 크기 요소의 크기가 기설정된 임계값 이상인 경우, 상기 스킵 데이터에 대응되는 로우 신호를 상기 제어부로 출력한다.

[0013] 실시예에 따라, 상기 비교기는 상기 신뢰도 행렬의 최대 크기 요소가 기설정된 임계값 미만인 경우, 상기 학습 데이터에 대응되는 하이 신호를 상기 제어부로 출력한다.

[0014] 실시예에 따라, 상기 제어부는 상기 로우 신호에 기초하여, 상기 학습 데이터에 대한 상기 제2 연산을 병렬화시킨다.

[0015] 실시예에 따라, 상기 로우 신호의 개수는 상기 제2 연산의 동작 시간에 반비례하는 관계이다.

[0016] 실시예에 따라, 전체 입력 데이터들로부터 임의로 선택된 상기 복수의 입력 데이터들 각각을 상기 연산부에 할당시키는 입력부 및 상기 연산부를 통해 출력받는 각 가중치 변화량을 합하여, 상기 서버 세트에 대한 기울기에 대응되는 출력 가중치 변화량을 출력하는 출력부를 더 포함한다.

[0017] 실시예에 따라, 상기 연산부는 시스틀릭 어레이 구조로 형성되고, 상기 제1 및 제2 연산을 순차적으로 수행하는 복수의 연산장치들을 포함한다.

[0018] 실시예에 따라, 상기 연산부는 상기 제어부로부터 출력받는 병렬화 제어 신호에 응답하여, 상기 복수의 연산장치들 중 상기 스킵 데이터에 대응되는 어느 하나의 연산장치를 초기화한다.

[0019] 실시예에 따라, 상기 연산부는 상기 복수의 연산장치들 중 나머지 연산장치에 할당된 학습 데이터의 일부를 상기 어느 하나의 연산장치에 재할당한다.

[0020] 실시예에 따라, 상기 제어부는 각 학습 데이터로부터 분할되는 복수의 서버 데이터들을 데이터 플로우에 따라 복수의 연산장치들에 재할당시킨다.

[0021] 실시예에 따라, 상기 데이터 플로우는 데이터를 리드하고 저장하기 위한 데이터의 이동 경로를 나타낸다.

[0022] 본 출원의 실시예에 따른 심층 신경망 학습 가속 장치의 동작 방법으로서, 연산부가 미니-배치 경사 하강법에 따라, 서버 세트의 복수의 입력 데이터들에 대한 제1 연산을 수행하는 단계, 판단부가 상기 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 상기 복수의 입력 데이터들을 스킵 데이터와 학습 데이터 중 하나로 각각 판단

하는 단계, 제어부가 상기 스킵 데이터에 응답하여, 상기 스킵 데이터에 대한 상기 제2 연산을 스킵시키기 위한 병렬화 제어 신호를 출력하는 단계 및 상기 연산부가 상기 병렬화 제어 신호에 기초하여, 상기 스킵 데이터에 대한 상기 제2 연산을 스킵하고, 상기 학습 데이터에 대한 상기 제2 연산을 수행하는 단계를 포함한다.

[0023] 실시예에 따라, 상기 제1 연산은 순전파 알고리즘(FP)을 이용하는 상기 미니-배치 경사 하강법의 제1 학습 단계이다.

[0024] 실시예에 따라, 상기 제2 연산은 역전파 알고리즘 및 가중치 업데이트 알고리즘을 순차적으로 이용하는 상기 미니-배치 경사 하강법의 제2 학습 단계이다.

[0025] 실시예에 따라, 상기 판단부는 상기 신뢰도 행렬의 최대 크기 요소와 기설정된 임계값 사이의 대소 여부를 비교하는 비교기로 구현된다.

[0026] 실시예에 따라, 상기 제2 연산을 수행하는 단계는 상기 연산부가 상기 병렬화 제어 신호에 기초하여, 복수의 연산장치들 중 상기 스킵 데이터에 대응되는 어느 하나의 연산장치를 초기화하는 단계, 상기 연산부가 나머지 연산장치에 할당된 상기 학습 데이터의 각 일부를 상기 어느 하나의 연산장치에 재할당하는 단계 및 상기 연산부가 상기 제1 연산으로부터 일정 시간 이후에 상기 복수의 연산장치들을 통해 상기 학습 데이터에 대한 상기 제2 연산을 병렬로 처리하는 단계를 포함한다.

발명의 효과

[0027] 본 출원의 실시예에 따른 심층 신경망 학습 가속 장치 및 그 동작 방법은 가중치 변화량을 출력하기 위한 전체 학습 연산량을 획기적으로 감소시켜 에너지 소모를 감소시킬 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 출원의 일 실시예에 따른 심층 신경망 학습 가속 장치에 대한 블록도이다.

도 2는 미니-배치 경사 하강법을 설명하기 위한 도이다.

도 3은 도 1의 심층 신경망 학습 가속 장치의 실시 예를 나타내는 도이다.

도 4는 도 3의 연산부가 제1 연산을 수행하는 동작 예이다.

도 5는 도 3의 연산부가 제2 연산을 수행하는 동작 예이다.

도 6은 도 1의 심층 신경망 학습 가속 장치의 동작 프로세스이다.

도 7은 도 5의 연산부가 제2 연산을 수행하는 동작 프로세스이다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하, 구체적인 실시형태 및 첨부된 도면을 참조하여 본 출원의 실시형태를 설명한다. 그러나, 본 출원의 실시형태는 여러 가지 다른 형태로 변형될 수 있으며, 본 출원의 범위가 이하 설명하는 실시형태로 한정되는 것은 아니다. 또한, 본 출원의 실시형태는 통상의 기술자에게 본 발명을 더욱 완전하게 설명하기 위해서 제공되는 것이다. 따라서, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있으며, 도면상의 동일한 부호로 표시되는 요소는 동일한 요소이다.

[0030] 그리고 도면에서 본 출원을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하고, 여러 층 및 영역을 명확하게 표현하기 위하여 두께를 확대하여 나타내었으며, 동일한 사상의 범위 내의 기능이 동일한 구성요소는 동일한 참조부호를 사용하여 설명한다. 나아가, 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.

[0031] 도 1은 본 출원의 일 실시예에 따른 심층 신경망 학습 가속 장치(10)에 대한 블록도이고, 도 2는 미니-배치 경사 하강법을 설명하기 위한 도이다.

[0032] 도 1과 도 2를 참조하면, 심층 신경망 학습 가속 장치(10)는 연산부(100), 판단부(200) 및 제어부(300)를 포함할 수 있다.

[0033] 먼저, 연산부(100)는 미니-배치 경사 하강법(Mini-batch Gradient Decent, MGD)에 따라, 서브 세트의 복수의

입력 데이터들에 대한 제1 및 제2 연산을 순차적으로 수행할 수 있다.

- [0034] 여기서, 서버 세트는 전체 입력 데이터들로부터 임의로 선택된 복수의 입력 데이터들을 포함할 수 있다. 예를 들면, 복수의 입력 데이터들은 이미지를 분류 및 검출하기 위한 데이터, 의료적 정보를 판단하기 위한 데이터, 자율 주행을 위한 데이터 및 보안 및 시스템 관리를 위한 데이터 중 어느 하나를 포함할 수 있다.
- [0035] 도 2에 도시된 바와 같이, 미니-배치 경사 하강법(MGD)은 서버 세트의 각 입력 데이터에 대한 가중치 변화량을 학습하도록 제1 및 제2 연산을 순차적으로 수행하기 위한 일련의 신경망 학습 스케줄일 수 있다. 여기서, 각 입력 데이터의 가중치 변화량의 합은 서버 세트의 출력 가중치 변화량일 수 있다.
- [0036] 이러한 미니-배치 경사 하강법(MGD)은 순전파 알고리즘(Foward Propagation algorithm, FP), 역전파 알고리즘(Backward Propagation algorithm, BP) 및 가중치 업데이트 알고리즘(Weight Update algorithm, WU)을 순차적으로 이용하는 학습 방법을 포함할 수 있다.
- [0037] 이때, 제1 연산은 순전파 알고리즘(FP)을 이용하는 미니-배치 경사 하강법(MGD)의 제1 학습 단계이고, 제2 연산은 역전파 알고리즘(BP) 및 가중치 업데이트 알고리즘(WU)을 이용하는 미니-배치 경사 하강법(MGD)의 제2 학습 단계일 수 있다. 즉, 제2 연산은 제1 연산으로부터 일정 시간 이후에 수행될 수 있다.
- [0038] 다음으로, 판단부(200)는 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 복수의 입력 데이터들 각각을 스kip 데이터와 학습 데이터 중 어느 하나로 판단할 수 있다.
- [0039] 도 2에 도시된 바와 같이, 신뢰도 행렬의 최대 크기 요소는 학습 기여도에 대응되는 손실 상태를 결정할 수 있다.
- [0040] 예를 들면, 도 2에 도시된 바와 같이, 신뢰도 행렬의 최대 크기 요소가 기설정된 임계치 이상인 경우, 해당 신뢰도 행렬은 낮은 학습 기여도에 대응되는 낮은 손실 상태를 결정할 수 있다. 또한, 신뢰도 행렬의 최대 크기 요소가 기설정된 임계치 미만인 경우, 해당 신뢰도 행렬은 높은 학습 기여도에 대응되는 높은 손실 상태를 결정할 수 있다.
- [0041] 실시예에 따른 판단부(200)는 신뢰도 행렬의 요소들 중 최대 크기 요소와 기설정된 임계값 사이의 대소 여부를 비교하는 비교기로 구현될 수 있다. 예를 들면, 도 3에 도시된 바와 같이, 비교기의 임계값은 0.9의 크기로 기설정될 수 있다.
- [0042] 이러한 비교기는 상기 비교 결과에 기초하여, 입력 데이터를 스kip 데이터와 학습 데이터 중 어느 하나로 판단하는 판단 신호를 제어부(300)에 출력할 수 있다. 여기서, 판단신호는 어느 하나의 로우 신호와 어느 하나의 하이 신호를 포함할 수 있다.
- [0043] 구체적으로, 신뢰도 행렬의 요소들 중 최대 크기 요소가 기설정된 임계값 미만인 경우, 판단부(200)는 신뢰도 행렬에 대응되는 입력 데이터를 학습 데이터로 판단하는 하이 신호를 출력할 수 있다. 또한, 신뢰도 행렬의 요소들 중 최대 크기 요소가 기설정된 임계값 이상인 경우, 판단부(200)는 신뢰도 행렬에 대응되는 입력 데이터를 스kip 데이터로 판단하는 로우 신호를 출력할 수 있다.
- [0044] 다음으로, 제어부(300)는 판단부(200)를 통해 판단된 판단 신호에 기초하여, 스kip 데이터에 대한 제2 연산을 스kip시킬 수 있도록 연산부(100)를 제어할 수 있다.
- [0045] 구체적으로, 제어부(300)는 판단부(200)를 통해 판단된 스kip 데이터에 대응되는 로우 신호를 출력받을 수 있다. 이때, 제어부(300)는 로우 신호에 응답하여, 해당 스kip 데이터에 대한 제2 연산을 스kip시키기 위한 병렬화 제어 신호를 연산부(100)에 출력함으로써, 연산부(100)의 제2 연산에 대해 제어할 수 있다. 이에, 연산부(100)는 병렬화 제어 신호에 기초하여, 스kip 데이터에 대한 제2 연산을 스kip할 수 있다.
- [0046] 또한, 제어부(300)는 로우 신호에 기초하여, 적어도 하나의 학습 데이터에 대한 제2 연산을 병렬화시킬 수 있다.
- [0047] 구체적으로, 제어부(300)는 적어도 하나의 학습 데이터의 각 일부를 스kip 데이터가 할당된 연산장치에 재할당시킴으로써, 적어도 하나의 학습 데이터에 대한 제2 연산을 병렬 처리시킬 수 있다.
- [0048] 예를 들면, 제어부(300)는 적어도 하나의 학습 데이터를 제1 및 제2 복수의 서버 데이터들로 분할하고, 제1 및 제2 복수의 서버 데이터들 중 하나의 복수의 서버 데이터들을 스kip 데이터가 할당된 연산장치에 재할당시킬 수 있다.

- [0049] 이때, 로우 신호의 개수는 제2 연산의 동작 시간에 반비례하는 관계일 수 있다. 예를 들면, 로우 신호의 개수가 증가함에 따라 연산부(100)가 제2 연산을 수행하는 동작 시간은 감소될 수 있다. 또한, 로우 신호의 개수가 감소됨에 따라 연산부(100)가 제2 연산을 수행하는 동작 시간은 증가될 수 있다.
- [0050] 실시예에 따라, 로우 신호의 개수가 하이 신호 개수보다 적은 경우, 제어부(300)는 데이터 크기가 가장 큰 순서대로 학습 데이터의 일부를 스킵 데이터가 할당된 연산장치에 재할당시킬 수 있다.
- [0051] 본 출원의 실시예에 따른 심층 신경망 학습 가속 장치(10)는 신뢰도 행렬에 기초하여 복수의 입력 데이터들을 판단부(200)를 통해 스킵 데이터와 학습 데이터로 각각 식별함으로써, 각 입력 데이터에 대한 학습 기여도를 판단할 수 있다. 이때, 심층 신경망 학습 가속 장치(10)는 제어부(100)를 통해 스킵 데이터에 대한 연산부(100)의 제2 연산을 스킵시킴으로써, 출력 가중치 변화량을 출력하기 위한 학습 연산량을 획기적으로 감소시킬 수 있다.
- [0052] 도 3은 도 1의 심층 신경망 학습 가속 장치(11)의 실시 예를 나타내고, 도 4는 도 3의 연산부(100)가 제1 연산을 수행하는 동작 예이며, 도 5는 도 3의 연산부(100)가 제2 연산을 수행하는 동작 예이다.
- [0053] 도 1 내지 도 5를 참조하면, 심층 신경망 학습 가속 장치(11)는 연산부(100), 판단부(200), 제어부(300), 입력부(400) 및 출력부(500)를 포함할 수 있다. 이하, 도 1 내지 도 2에서 설명된 동일한 부재번호의 연산부(100), 판단부(200) 및 제어부(300)에 대한 중복된 설명은 생략될 것이다.
- [0054] 먼저, 입력부(400)는 전체 입력 데이터들로부터 임의로 선택된 서브 세트의 복수의 입력 데이터들 각각을 연산장치에 개별적으로 할당시킬 수 있다. 여기서, 복수의 입력 데이터들의 개수는 연산장치의 개수에 일대일 대응될 수 있다.
- [0055] 구체적으로, 입력부(400)는 전체 입력 데이터들을 입력받고, 미니-배치 경사 하강법(MGD)에 따라 전체 입력 데이터들로부터 임의로 선택된 서브 세트의 복수의 입력 데이터들을 연산부(100)의 연산장치에 개별적으로 출력할 수 있다. 즉, 각 연산장치에는 서로 다른 입력 데이터들이 할당될 수 있다.
- [0056] 다음으로, 연산부(100)는 복수의 연산장치들(예컨대, 110_1~110_5)을 포함할 수 있다.
- [0057] 실시예에 따른 복수의 연산장치들(예컨대, 110_1~110_5)은 시스톨릭 어레이 구조로 형성되고, 제1 및 제2 연산을 순차적으로 각각 수행할 수 있다. 이때, 복수의 연산장치들(예컨대, 110_1~110_5) 각각은 메모리 인 컴퓨팅으로 구현될 수 있다.
- [0058] 구체적으로, 복수의 연산장치들(예컨대, 110_1~110_5)은 입력부(400)를 통해 할당된 복수의 입력 데이터들에 대하여, 제1 연산을 각각 수행할 수 있다. 즉, 복수의 연산장치들(예컨대, 110_1~110_5)은 순전파 알고리즘(FP)을 이용하여, 복수의 입력 데이터들에 대한 제1 연산을 동시에 분산 처리할 수 있다.
- [0059] 예를 들면, 순전파 알고리즘(FP)은 다음의 계산식(1)에 대응되고, 여기서, 계산식(1)은 $A_{in} \otimes W = A_{out}$ (1)이고, 이때, A_{in} 은 순전파 알고리즘의 활성화 입력이고, W 는 가중치이며, A_{out} 는 활성화 출력일 수 있다.
- [0060] 도 4에 도시된 바와 같이, 복수의 연산장치들(예컨대, 110_1~110_5)은 제1 연산을 수행함에 따라 획득되는 복수의 신뢰도 행렬들에 기초하여, 손실 상태를 출력할 수 있다.
- [0061] 또한, 도 5에 도시된 바와 같이, 복수의 연산장치들(예컨대, 110_1~110_5)은 제1 연산 시점으로부터 일정 시간 이후에 복수의 입력 데이터들에 대하여, 제2 연산을 수행할 수 있다. 즉, 복수의 연산장치들(예컨대, 110_1~110_5)은 역전파 알고리즘(BP) 및 가중치 업데이트 알고리즘(WU)을 이용하여, 제2 연산을 동시에 분산 처리할 수 있다.
- [0062] 예를 들면, 역전파 알고리즘(BP)은 다음의 계산식(2)에 대응되고, 여기서, 계산식(2)는 $L_{in} \otimes W^T = L_{out}$ (2)이고, 이때, L_{in} 은 손실 입력이고, W^T 는 전치(Transpose)된 가중치이며, L_{out} 는 손실 출력일 수 있다.
- [0063] 또한, 가중치 업데이트 알고리즘(WU)은 다음의 계산식(3)에 대응되고, 여기서, 계산식(3)은 $A_{in} \otimes L_{in} = W_G$ (3)이고, 이때, A_{in} 은 순전파 시의 활성화 입력이고, L_{in} 은 손실 입력이며, W_G 는 가중치 변화량 일 수 있다.

- [0064] 실시예에 따라, 연산부(100)는 병렬화 제어 신호에 응답하여, 복수의 연산장치들(예컨대, 110_1~110_5) 중 스kip 데이터에 대응되는 어느 하나의 연산장치(예컨대, 110_2, 110_5)을 초기화할 수 있다.
- [0065] 이때, 연산부(100)는 나머지 연산장치(예컨대, 110_1, 110_3, 및 110_5)에 할당된 학습 데이터의 일부를 어느 하나의 연산장치(예컨대, 110_2, 110_5)에 재할당할 수 있다.
- [0066] 이후, 연산부(100)는 제1 연산으로부터 일정 시간 이후에, 복수의 연산장치들(예컨대, 110_1~110_5)을 통해 학습 데이터에 대한 제2 연산을 병렬로 처리함으로써, 연산 속도를 증가시킬 수 있다. 여기서, 어느 하나의 연산장치(예컨대, 110_2, 110_5)의 개수는 학습 데이터의 일부인 병렬화 데이터의 개수에 일대일 대응될 수 있다.
- [0067] 실시예에 따라, 제어부(300)는 각 학습 데이터로부터 분할되는 복수의 서브 데이터들을 데이터 플로우에 따라 복수의 연산장치들(예컨대, 110_1~110_5)에 재할당시킬 수 있다. 여기서, 데이터 플로우는 데이터를 리드하고 저장하기 위한 데이터의 이동 경로를 의미할 수 있다.
- [0068] 즉, 제어부(300)는 각 학습 데이터를 복수의 연산장치들(예컨대, 110_1~110_5)에 재할당시킴으로써, 학습 데이터에 대한 제2 연산을 분산 처리시킬 수 있다.
- [0069] 다음으로, 출력부(500)는 연산부(100)를 통해 출력받는 각 가중치 변화량에 기초하여, 서브 세트에 대한 기울기 (gradient)를 의미하는 출력 가중치 변화량을 출력할 수 있다. 여기서, 출력 가중치 변화량은 각 가중치 변화량의 합일 수 있다.
- [0070] 도 6은 도 1의 심층 신경망 학습 가속 장치(10)의 동작 프로세스이다.
- [0071] 도 1과 도 6을 참조하면, S110 단계에서, 연산부(100)는 미니-배치 경사 하강법에 따라 서브 세트의 복수의 입력 데이터들에 대한 제1 연산을 수행할 수 있다.
- [0072] 그런 다음, S120 단계에서, 판단부(200)는 제1 연산에 따라 획득되는 신뢰도 행렬에 기초하여, 복수의 입력 데이터들 각각을 스kip 데이터와 학습 데이터 중 어느 하나로 판단할 수 있다.
- [0073] 그런 다음, S130 단계에서, 제어부(300)는 판단부(200)를 통해 판단된 스kip 데이터에 응답하여, 병렬화 제어 신호를 연산부(100)로 출력할 수 있다.
- [0074] 이때, S140 단계에서, 연산부(100)는 병렬화 제어 신호에 기초하여, 스kip 데이터에 대한 제2 연산을 스kip하고, 학습 데이터에 대한 제2 연산을 수행할 수 있다.
- [0075] 도 7은 도 5의 연산부(100)가 제2 연산을 수행하는 동작 프로세스이다.
- [0076] 도 5와 도 7을 참조하면, 먼저 S210 단계에서, 연산부(100)는 병렬화 제어 신호에 기초하여, 복수의 연산장치들 (예컨대, 110_1~110_5) 중 스kip 데이터에 대응되는 어느 하나의 연산장치(예컨대, 110_2, 110_5)을 초기화할 수 있다.
- [0077] 이때, S220 단계에서, 연산부(100)는 나머지 연산장치(예컨대, 110_1, 110_3, 및 110_5)에 할당된 학습 데이터의 일부를 어느 하나의 연산장치(예컨대, 110_2, 110_5)에 분배할 수 있다.
- [0078] 이후, S230 단계에서, 연산부(100)는 제1 연산으로부터 일정 시간 이후에, 복수의 연산장치들(예컨대, 110_1~110_5)을 통해 학습 데이터에 대한 제2 연산을 병렬로 처리할 수 있다.
- [0079] 본 출원은 도면에 도시된 일 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 출원의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

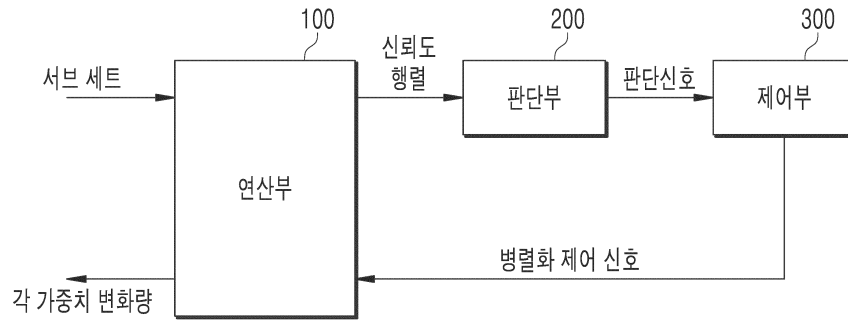
- [0080] 10, 11: 심층 신경망 학습 가속 장치
- 100: 연산부
- 200: 판단부
- 300: 제어부
- 400: 입력부

500: 출력부

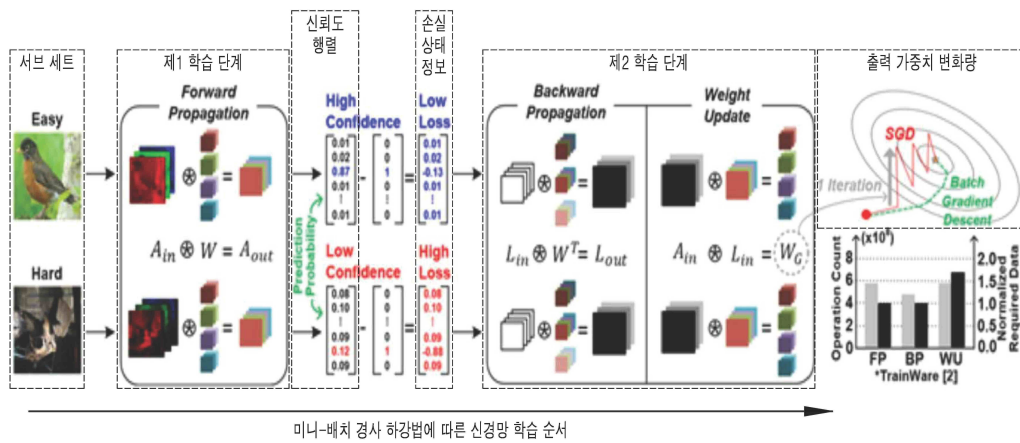
도면

도면1

10

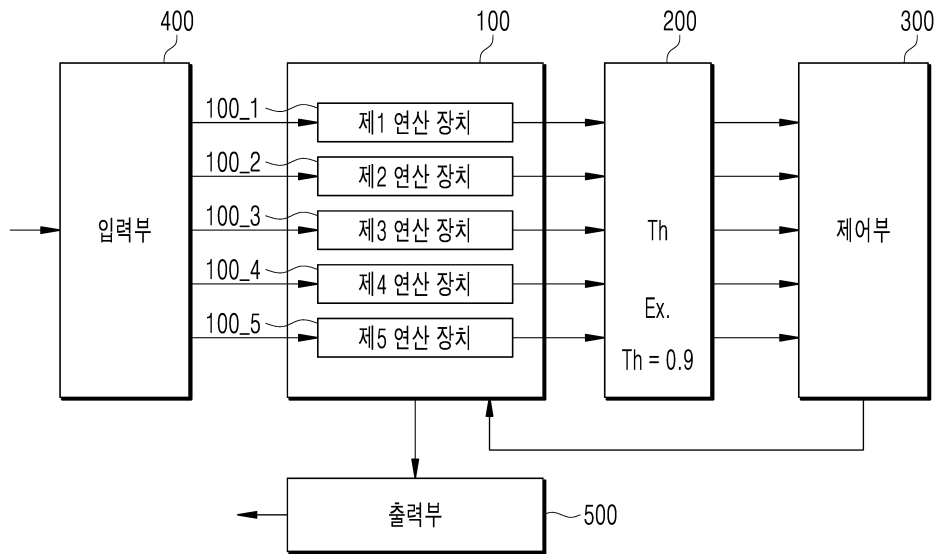


도면2

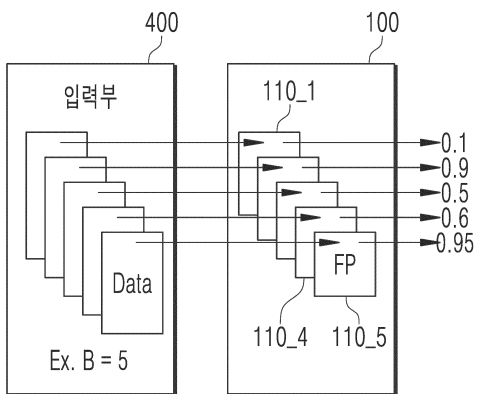


도면3

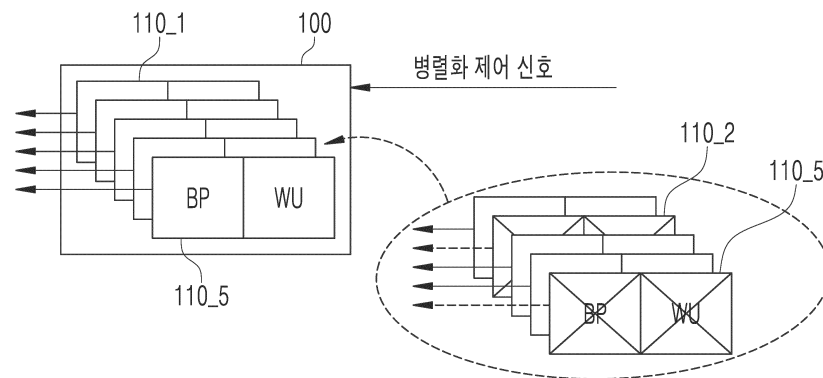
11



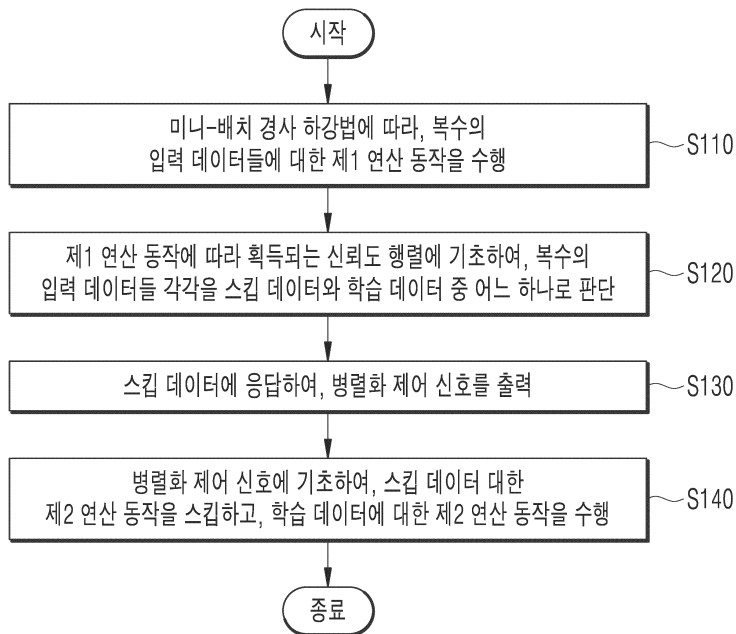
도면4



도면5



도면6



도면7

