(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년12월16일
(11) 등록번호 10-2340091
(24) 등록일자 2021년12월13일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2021-0039857
(22) 출원일자 2021년03월26일
심사청구일자 2021년03월26일
(56) 선행기술조사문헌
KR1020180043154 A*
KR101987475 B1
KR1020180013674 A
KR1020210023912 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
이성주
서울특별시 광진구 뚝섬로35길 32, 308-1110
이준표
서울특별시 광진구 능동로28길 23 202호
(74) 대리인
이강민, 안준형, 남승희

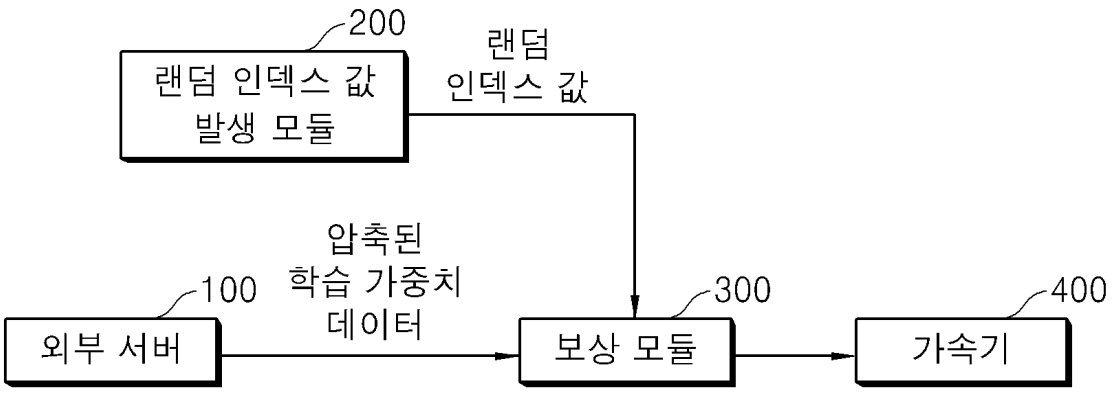
전체 청구항 수 : 총 12 항 심사관 : 송근배

(54) 발명의 명칭 **인공신경망의 양자화 오차 보상 시스템 및 그 방법**

(57) 요 약

본 발명은 인공신경망의 양자화 오차 보상 시스템 및 그 방법에 관한 것으로서, 외부 서버(PC)와 가속기(Accelerator) 사이에서 학습된 가중치의 압축 시 발생한 오차를 보상하는 하드웨어 모듈 기반의 보상 모듈을 구성하여, 학습된 가중치의 압축 시 발생한 양자화 오차를 보정하는 인공신경망의 양자화에서 발생하는 오차 보상 시스템 및 그 방법에 관한 것이다.

대 표 도 - 도2



(52) CPC특허분류
G06N 3/063 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711116145
과제번호 2018-0-01423-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 대학ICT연구센터지원사업
연구과제명 지능형 비행로봇 융합기술 연구
기 여 율 5/10
과제수행기관명 세종대학교 산학협력단
연구기간 2021.01.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345321135
과제번호 2020R1A6A1A0303854011
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 대학중점연구소지원사업
연구과제명 자율지능무인비행체연구소
기 여 율 1/10
과제수행기관명 세종대학교 산학협력단
연구기간 2021.03.01 ~ 2022.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711108024
과제번호 2020R1A2C1007546
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 실내보안용 초고해상도 지능형 레이다센서 신호처리 연구
기 여 율 4/10
과제수행기관명 세종대학교 산합협력단
연구기간 2021.03.01 ~ 2022.02.28

명 세 서

청구범위

청구항 1

인공신경망(artificial neural network)의 양자화(quantization)에서 발생하는 오차를 보상하는 시스템에 있어서,

학습된 가중치 값들을 양자화하여 압축시킨 압축된 학습 가중치 데이터를 보상 모듈로 전달하는 외부 서버;

소정의 주기 간격으로 랜덤 인덱스 값을 발생시켜 보상 모듈로 입력하는 랜덤 인덱스 값 발생 모듈;

미리 저장된 보정데이터를 기반으로 상기 전달받은 학습 가중치 데이터를 압축시 발생한 양자화 오차를 보상하여 보정하고, 보정된 학습 가중치 데이터를 가속기로 전달하는 보상모듈;을 포함하여 구성되며,

상기 보상모듈은,

서로 다른 값을 가지는 적어도 둘 이상의 보정 가중치 값이 배열되고 있고, 각 배열마다 인덱스 값이 할당된 룩업 테이블로 구성되는 보정 데이터를 저장하는 제1 메모리 모듈;

외부 서버로부터 전달되는 압축된 학습 가중치 데이터를 임시 저장하는 제2 메모리 모듈;

상기 랜덤 인덱스 값 발생 모듈에서 발생시킨 인덱스 값을 이용하여 상기 제1 메모리 모듈의 보정 데이터 및 제2 메모리 모듈의 압축된 학습 가중치로부터 보정 가중치 값 및 학습 가중치 값을 각각 추출하고, 상기 추출된 보정 가중치 값을 상기 추출된 학습 가중치 값에 더하거나 차감하여 상기 추출된 학습 가중치 값을 보정하는 에러 보정 모듈;

을 포함하여 구성되며,

상기 보정 데이터는,

상기 학습 가중치 값들의 양자화 압축 시 손실된 값인 보정 가중치 값들이 각각의 손실 빈도수에 대응하는 비율로 구성되는 것;

을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 시스템.

청구항 2

제1항에 있어서,

상기 보상 모듈로부터 전달되는 보정된 학습 가중치 데이터를 이용하여 추론 결과를 출력하는 가속기;

를 더 포함하여 구성되는 인공신경망의 가중치 양자화 오차 보상 시스템.

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 제2 메모리 모듈에 저장되는 압축된 학습 가중치 데이터는,

각 클러스터(cluster) 별로 학습 가중치 값이 배열되어 있고, 각 배열마다 인덱스 값이 할당된 형태로 구성되는 것을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 시스템.

청구항 5

제1항 또는 제4항에 있어서,

상기 보상 모듈은,

상기 제1 메모리 모듈에 저장된 보정 데이터 및 제2 메모리 모듈에 저장된 압축된 학습 가중치 데이터로부터, 상기 랜덤 인덱스 값 발생 모듈에 의해 발생된 인덱스 값에 해당하는 보정 가중치 값 및 학습 가중치 값을 각각 추출하는 가중치 값 추출 모듈;

상기 가중치 값 추출 모듈로부터 임의의 인덱스 값에 대응하는 보정 가중치 값 및 학습 가중치 값이 추출되면, 상기 학습 가중치 값에 상기 보정 가중치 값을 더하거나 또는 차감하여 상기 학습 가중치 값의 양자화 오차를 보정하는 연산 모듈;

을 포함하여 구성되는 것을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 시스템.

청구항 6

제5항에 있어서,

상기 보정 데이터를 구성하는 각각의 보정 가중치 값은,

상기 외부 서버에서 압축된 학습 가중치 데이터 획득하는 과정에서 발생된 양자화 오차 값이며,

상기 룩업 테이블에 상기 발생된 빈도 수에 대응하는 비율에 따라 배열되는 것; 을 특징으로 인공신경망의 가중치 양자화 오차 보상 시스템.

청구항 7

제5항에 있어서,

상기 연산 모듈은,

상기 추출된 학습 가중치 값이 양수인 경우, 상기 학습 가중치 값에서 상기 추출된 보정 가중치 값을 차감하고,

상기 추출된 학습 가중치 값이 음수인 경우, 상기 학습 가중치 값에 상기 추출된 보정 가중치 값을 더하는 것;

을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 시스템.

청구항 8

인공신경망(artificial neural network)의 양자화(quantization)에서 발생하는 오차를 보상하는 방법에 있어서,

외부 서버와 가속기 사이에 구성되는 보상 모듈에, 소정의 보정 데이터를 미리 획득하여 룩업 테이블로 저장하는 보정 데이터 저장 단계;

보상 모듈에서, 외부 서버로부터 압축된 학습 가중치 데이터를 수신하는 압축된 학습 가중치 데이터 수신 단계;

보상 모듈에서, 상기 압축된 학습 가중치 데이터 수신 단계에서 수신한 외부 서버로부터의 압축된 학습 가중치 데이터를 임시 저장하는 압축된 학습 가중치 데이터 저장 단계;

랜덤 인덱스 값 발생 모듈에서, 소정의 주기 간격으로 랜덤 인덱스 값을 발생시키는 랜덤 인덱스 값 발생 단계;

보상 모듈은, 상기 랜덤 인덱스 값 발생 단계에 의해 인덱스 값이 발생되면, 상기 발생된 인덱스 값을 이용하여 상기 보정 데이터 저장 단계에서 저장된 보정 데이터 및 상기 압축된 학습 가중치 데이터 저장 단계에서 저장된 압축된 학습 가중치 데이터로부터 보정 가중치 값 및 학습 가중치 값을 각각 추출하여, 상기 추출한 보정 가중치 값을 상기 추출한 학습 가중치 값에 더하거나 차감하는 방식을 통해 상기 압축된 학습 가중치 데이터의 압축 시 발생한 양자화 오차를 보정하는 압축된 학습 가중치 데이터 보정 단계;

를 포함하여 구성되며,

상기 보정데이터는,

상기 학습 가중치 값들의 양자화 압축시 손실된 값인 보정 가중치 값들이 각각의 손실 빈도수에 대응하는 비율로 구성되는 것;

을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 방법.

청구항 9

제8항에 있어서,

상기 보정 데이터는,

서로 다른 값을 가지는 적어도 둘 이상의 보정 가중치 값이 배열되고 있고, 각 배열마다 인덱스 값이 할당된 룩업 테이블로 구성되는 것; 을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 방법.

청구항 10

제8항에 있어서,

상기 압축된 학습 가중치 데이터는,

각 클러스터(cluster) 별로 학습 가중치 값이 배열되고 있고, 각 배열마다 인덱스 값이 할당된 형태로 구성되는 것; 을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 방법.

청구항 11

제8항 내지 제10항 중 어느 한 항에 있어서,

상기 압축된 학습 가중치 데이터 보정 단계는,

상기 보정 데이터 저장 단계에서 저장된 보정 데이터 및 상기 압축된 학습 가중치 데이터 저장 단계에서 저장된 압축된 학습 가중치 데이터로부터, 상기 랜덤 인덱스 값 발생 단계에서 발생된 인덱스 값에 해당하는 보정 가중치 값 및 학습 가중치 값을 각각 추출하는 가중치 값 추출 단계;

상기 가중치 값 추출 단계에서 추출된 상기 추출된 학습 가중치 값에 상기 추출된 보정 가중치 값을 더하거나 또는 차감하여 상기 학습 가중치 값을 보정하는 연산 단계;

를 포함하여 구성되는 것을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 방법.

청구항 12

제11항에 있어서,

상기 연산 단계에서,

상기 추출된 학습 가중치 값이 양수인 경우, 상기 학습 가중치 값에서 상기 추출된 보정 가중치 값을 차감하고,

상기 추출된 학습 가중치 값이 음수인 경우, 상기 학습 가중치 값에 상기 보정 가중치 값을 더하는 것;

을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 방법.

청구항 13

제11항에 있어서,

상기 보정 데이터를 구성하는 각각의 보정 가중치 값은,

외부 서버에서 압축된 학습 가중치 데이터를 획득하는 과정에서 발생된 양자화 오차 값이며,

상기 룩업 테이블에 상기 발생된 빈도 수에 대응하는 비율에 따라 배열되는 것; 을 특징으로 하는 인공신경망의 가중치 양자화 오차 보상 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공신경망의 양자화 오차 보상 시스템 및 그 방법에 관한 것으로서, 외부 서버(PC)와 가속기(Accelerator) 사이에서 학습된 가중치의 압축 시 발생한 오차를 보상하도록 구성된 하드웨어 모듈을 포함하는 인공신경망의 양자화에서 발생하는 오차 보상 시스템 및 그 방법에 관한 것이다.

배 경 기 술

[0002] 인공지능(Artificial Intelligence)의 한 분야인 딥러닝(Deep Learning)은 복잡한 데이터의 패턴을 인식하고 정교한 예측을 가능하게 한다는 점에서 4차 산업혁명 시대의 핵심 기술로서 다양한 분야에서 활용되고 있다. 딥러닝은 인간의 생물학적 신경 세포의 특성을 수학적 표현에 의해 모델링 한 인공신경망(artificial neural network)을 깊게 구성하여 학습하는 방법을 말한다.

[0003] 일반적으로 딥러닝은 학습용 데이터를 활용하여 인공신경망을 학습시키는 학습 단계(training)와, 학습이 완료된 인공신경망 모델(trained model)에 새로운 데이터를 입력하여 출력을 얻는 추론 단계(inference)로 구성된다. 이러한 딥러닝은 인공신경망을 깊게 구성할수록 더 정교한 예측이 가능하여 성능을 끌어올릴 수 있는 반면, 이는 막대한 연산량으로 인해 더 많은 전력을 필요로 하고, 속도가 저하되는 문제로 이어지게 된다. 이러한 문제를 해결하기 위해 비슷한 성능을 유지한 채 더 적은 파라미터 수와 연산량을 가지는 모델을 만드는 인공신경망 모델 경량화 기술이 사용되고 있다.

[0004] 이러한 인공신경망 모델 경량화 기술은 크게 알고리즘 자체를 적은 연산과 효율적인 구조로 설계하는 경량 알고리즘 연구 방식과 만들어진 모델의 파라미터들을 줄이는 모델 압축과 같은 기법을 적용하는 알고리즘 경량화 방식으로 나눠진다.

[0005] 알고리즘 경량화 방식은 인공신경망을 압축하는 여러 기술이 적용될 수 있는데, 주로 양자화(quantization) 하여 가중치(weight)로 저장하는 bit를 최소화 하는 방법을 사용하고 있다. 보다 구체적으로 도 1의 예시를 참조하여 양자화의 과정을 설명하면, 도 1의 (a)와 같이 32 bit로 표현되는 가중치(weights)를 클러스터링(clustering)하여 (b)에 보이는 것처럼 그 클러스터(cluster)의 개수로 가중치를 표현한다. 그리고, (c)와 같이 클러스터별로의 중심(centroid)을 클러스터를 대표하는 값으로 활용하고, 이 값을 재학습을 통해 (d)와 같은 보정된 값을 획득하는 것으로 구성된다. 이 때, 도 1의 (b)에서 나타내는 가중치 값은 신경망이 가지고 있는 모든 가중치들을 클러스터링과 양자화를 통해 압축시킨 것이므로 정보의 손실이 존재한다.

[0006] 이와 같은 정보의 손실을 최소화하기 위해, 종래에는 (d)의 보정된 가중치 값을 획득하고자 할 때, 도 1의 (e)와 같은 기울기(gradient) 값을 구하여 이들 값들을 (f)처럼 그룹화한 후 그룹별로의 중심(centroid)을 (g)와 같이 그룹의 기울기를 대표하는 값으로 하고, 이를 이용하여 (c)의 클러스터별 가중치 값을 보정하여 (d)와 같은 보정된 가중치 값을 획득하는 소프트웨어 방식을 사용함으로써 압축 시 발생한 정보의 손실을 최소화 시켰다.

[0007] 그런데, 이러한 소프트웨어 방식의 경우, (e)의 기울기 값들을 구하기 위해 재훈련을 시키는 과정이 필요로 하기 때문에 이로 인해 많은 시간이 소모되는 문제점을 가지고 있다.

[0008] (특허문헌 1) KR10-2019-0130455 A

발명의 내용

해결하려는 과제

[0009] 본 발명은 상술한 문제점을 해결하고자 하는 것으로서, 외부 서버(PC)에서 가속기(Accelerator)로 압축한 학습된 가중치가 전달될 때, 상기 압축한 학습된 가중치를 확률적 접근 방식으로 보정하여 가속기에 입력하는 하드웨어 모듈을 구성하여, 압축에 의한 정보 손실을 향상된 속도로 최소화 할 수 있는 인공신경망의 양자화에서 발생하는 오차 보상 시스템 및 그 방법을 제공하고자 한다.

과제의 해결 수단

[0010] 본 발명에 따른 인공신경망(artificial neural network)의 양자화(quantization)에서 발생하는 오차를 보상하는 시스템은, 학습된 가중치 값들을 양자화하여 압축시킨 압축된 학습 가중치 데이터를 보상 모듈로 전달하는 외부 서버; 소정의 주기 간격으로 임의의 인덱스 값을 발생시켜 보상 모듈로 입력하는 랜덤 인덱스 값 발생 모듈; 소정의 보정 데이터를 미리 구비하며, 상기 랜덤 인덱스 값 발생 모듈에 의해 발생된 인덱스 값을 이용하여 상기 외부 서버로부터 전달 받은 압축된 학습 가중치 데이터의 압축 시 발생한 양자화 오차를 상기 보정 데이터로 보상하여 가속기로 전달하는 보상 모듈; 를 포함하여 구성된다.

[0011] 또한, 상기 보상 모듈로부터 전달되는 보정된 학습 가중치 데이터를 이용하여 추론 결과를 출력하는 가속기; 를

더 포함하여 구성된다.

[0012] 한편, 상기 보상 모듈은, 서로 다른 값을 가지는 적어도 둘 이상의 보정 가중치 값이 배열되어 있고, 각 배열마다 인덱스 값이 할당된 룩업 테이블로 구성되는 소정의 보정 데이터를 저장하는 제1 메모리 모듈; 외부 서버로부터 전달되는 압축된 학습 가중치 데이터를 임시 저장하는 제2 메모리 모듈; 상기 랜덤 인덱스 값 발생 모듈에서 발생시킨 인덱스 값을 이용하여 상기 제1 메모리 모듈의 보정 데이터 및 제2 메모리 모듈의 압축된 학습 가중치로부터 보정 가중치 값 및 학습 가중치 값을 각각 추출하고, 상기 추출된 보정 가중치 값을 이용하여 상기 추출된 학습 가중치 값을 보정하는 에러 보정 모듈; 을 포함하여 구성되는 것을 특징으로 한다.

[0013] 여기서, 상기 제2 메모리 모듈에 저장되는 압축된 학습 가중치 데이터는, 각 클러스터(cluster) 별로 학습 가중치 값이 배열되어 있고, 각 배열마다 인덱스 값이 할당된 형태로 구성되는 것을 특징으로 한다.

[0014] 한편, 상기 보상 모듈은, 상기 제1 메모리 모듈에 저장된 보정 데이터 및 제2 메모리 모듈에 저장된 압축된 학습 가중치 데이터로부터, 상기 랜덤 인덱스 값 발생 모듈에 의해 발생된 인덱스 값에 해당하는 보정 가중치 값 및 학습 가중치 값을 각각 추출하는 가중치 값 추출 모듈; 상기 가중치 값 추출 모듈로부터 임의의 인덱스 값에 대응하는 보정 가중치 값 및 학습 가중치 값이 추출되면, 상기 학습 가중치 값에 상기 보정 가중치 값을 더하거나 또는 차감하여 상기 학습 가중치 값의 양자화 오차를 보정하는 연산 모듈; 을 포함하여 구성되는 것을 특징으로 한다.

[0015] 여기서, 상기 보정 데이터를 구성하는 각각의 보정 가중치 값은, 상기 외부 서버에서 압축된 학습 가중치 데이터 획득하는 과정에서 발생된 양자화 오차 값이며, 상기 룩업 테이블에 상기 발생된 빈도 수에 대응하는 비율에 따라 배열되는 것; 을 특징으로 한다.

[0016] 한편, 상기 연산 모듈은, 상기 추출된 학습 가중치 값이 양수인 경우, 상기 학습 가중치 값에서 상기 추출된 보정 가중치 값을 차감하고, 상기 추출된 학습 가중치 값이 음수인 경우, 상기 학습 가중치 값에 상기 추출된 보정 가중치 값을 더하는 것; 을 특징으로 한다.

[0017] 본 발명에 따른 인공신경망(artificial neural network)의 양자화(quantization)에서 발생하는 오차를 보상하는 방법은, 외부 서버와 가속기 사이에 구성되는 보상 모듈에, 룩업 테이블로 구성된 소정의 보정 데이터를 미리 획득하여 저장하는 보정 데이터 저장 단계; 보상 모듈에서, 외부 서버로부터 압축된 학습 가중치 데이터를 수신하는 압축된 학습 가중치 데이터 수신 단계; 보상 모듈에서, 상기 압축된 학습 가중치 데이터 수신 단계에서 수신한 외부 서버로부터의 압축된 학습 가중치 데이터를 임시 저장하는 압축된 학습 가중치 데이터 저장 단계; 랜덤 인덱스 값 발생 모듈에서, 소정의 주기 간격으로 임의의 인덱스 값을 발생시키는 랜덤 인덱스 값 발생 단계; 보상 모듈은, 상기 랜덤 인덱스 값 발생 단계에 의해 인덱스 값이 발생되면, 상기 발생된 인덱스 값을 이용하여 상기 보정 데이터 저장 단계에서 저장된 보정 데이터 및 상기 압축된 학습 가중치 데이터 저장 단계에서 저장된 압축된 학습 가중치 데이터로부터 보정 가중치 값 및 학습 가중치 값을 각각 추출하여, 상기 추출한 보정 가중치 값으로 상기 추출한 학습 가중치 값을 보정하는 압축된 학습 가중치 데이터 보정 단계; 를 포함하여 구성된다.

[0018] 여기서, 상기 보정 데이터는, 서로 다른 값을 가지는 적어도 둘 이상의 보정 가중치 값이 배열되고 있고, 각 배열마다 인덱스 값이 할당된 룩업 테이블로 구성되는 것; 을 특징으로 한다.

[0019] 또한, 상기 압축된 학습 가중치 데이터는, 각 클러스터(cluster) 별로 학습 가중치 값이 배열되고 있고, 각 배열마다 인덱스 값이 할당된 형태로 구성되는 것; 을 특징으로 한다.

[0020] 한편, 상기 압축된 학습 가중치 데이터 보정 단계는, 상기 보정 데이터 저장 단계에서 저장된 보정 데이터 및 상기 압축된 학습 가중치 데이터 저장 단계에서 저장된 압축된 학습 가중치 데이터로부터, 상기 랜덤 인덱스 값 발생 단계에서 발생된 인덱스 값에 해당하는 보정 가중치 값 및 학습 가중치 값을 각각 추출하는 가중치 값 추출 단계; 상기 가중치 값 추출 단계에서 추출된 상기 추출된 학습 가중치 값에 상기 추출된 보정 가중치 값을 더하거나 또는 차감하여 상기 학습 가중치 값을 보정하는 연산 단계; 를 포함하여 구성되는 것을 특징으로 한다.

[0021] 여기서, 상기 연산 단계에서, 상기 추출된 학습 가중치 값이 양수인 경우, 상기 학습 가중치 값에서 상기 추출된 보정 가중치 값을 차감하고, 상기 추출된 학습 가중치 값이 음수인 경우, 상기 학습 가중치 값에 상기 보정 가중치 값을 더하는 것; 을 특징으로 한다.

[0022] 한편, 상기 보정 데이터를 구성하는 각각의 보정 가중치 값은, 외부 서버에서 압축된 학습 가중치 데이터를 획

득하는 과정에서 발생된 양자화 오차 값이며, 상기 룩업 테이블에 상기 발생된 빈도 수에 대응하는 비율에 따라 배열되는 것; 을 특징으로 한다.

발명의 효과

[0023] 본 발명은 외부 서버(PC)와 가속기(Accelerator) 사이에 외부 서버로부터 압축하여 전달되는 학습된 가중치를 압축 시 발생한 오차를 보정하여 가속기로 입력하는 하드웨어 보상 모듈을 구성하여, 종래의 외부 서버(PC)에서 소프트웨어 방식으로 보정하였던 것에 비해 향상된 보정 속도로 정보의 손실을 최소화할 수 있다.

도면의 간단한 설명

[0024] 도 1은 종래의 양자화 기술을 사용한 인공신경망 모델의 압축 과정의 예시를 보여주는 도면이다.

도 2는 본 발명에 따른 인공신경망의 양자화 오차 보상 시스템의 전체적인 구성을 도시한 도면이다.

도 3은 도 2의 보상 모듈의 세부 구성을 도시한 도면이다.

도 4는 본 발명의 양자화 오차 보정 원리를 보여주는 도면이다.

도 5는 본 발명에 따른 인공신경망의 양자화 오차 보상 방법의 흐름도를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0025] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시 예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시 예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면부호를 붙였다.

[0026] 이하, 도면을 참조하여 본 발명에 대하여 상세하게 설명한다.

[0027] 1. 본 발명에 따른 양자화 오차 시스템

[0028] 도 2는 본 발명의 양자화 오차 시스템의 전체적인 구성을 도시한 도면이고, 도 3은 보상 모듈의 세부 구성을 도시한 도면이다.

[0029] 도 2 및 3을 참조하면, 본 발명의 크게 외부 서버(100), 랜덤 인덱스 값 발생 모듈(200), 보상 모듈(300), 가속기(400)를 포함하여 구성된다.

[0030] 1.1. 외부 서버(100)

[0031] 외부 서버(100)는, 학습된 가중치 데이터를 다양한 공지의 인공신경망의 압축 기법 중 하나로서 양자화(quantization) 기법을 사용하여 압축시켜 후술하는 보상 모듈(300)로 전달한다.

[0032] 인공신경망(artificial neural network)은 기본적으로 노드와 엣지로 구성되어 있으며, 엣지에 가중치(weight)가 부여되어 있는 형태를 가지는 하나의 수학적인 함수이다. 인공신경망을 학습한다는 것은, 그 함수의 최적의 계수인 가중치를 찾아가는 것으로 이해할 수 있다. 이에 인공신경망이 깊게 구성될수록 가중치의 수는 많아질 것이고, 이는 곧 연산량 증가로 이어져 전력 소모가 높아지게 된다. 이에, 일반적으로 외부 서버는 인공신경망을 학습하여 학습된 가중치(weight)들을 획득한 후, 이들을 다양한 압축 기술을 사용하여 학습된 가중치들을 압축시켜 가속기(accelerator)로 전달하도록 구성된다.

[0033] 본 발명의 외부 서버(100)는, 학습된 가중치 값들을 다양한 공지의 압축 기술 중 양자화(quantization) 기술을 사용하여 압축시키도록 구성되며, 본 명세서에서는 양자화하여 압축시킨 학습된 가중치(weight) 값들을 압축된 학습 가중치 데이터로 지칭하여 설명한다.

[0034] 이러한 압축된 학습 가중치 데이터는, 예를 들어 도 1의 (c)와 같은 형태로 구성될 수 있다. 즉, 클러스터(cluster) 별로 중심(centroid) 값이 해당 클러스터의 가중치 값으로 설정되어 있는 형태인 것이다. 이 때, 각 클러스터 별로 해당 학습 가중치 값을 가리키는 인덱스(index) 값을 할당되어, 후술하는 보상 모듈(300)에서 인덱스 값을 이용하여 그에 해당하는 가중치 값을 찾아 읽어 들일 수 있다.

[0035] 외부 서버(100)는, 이와 같이 각 클러스터의 학습 가중치 값 별로 인덱스가 할당되어 있는 형태로 구성된 압축된 학습 가중치 데이터를 후술하는 보상 모듈(300)로 전달한다.

[0036] 1.2. 랜덤 인덱스 값 발생 모듈(200)

[0037] 랜덤 인덱스 값 발생 모듈(200)은, 임의의 인덱스 값을 랜덤으로 발생시켜 후술하는 보상 모듈(300)로 입력할 수 있다. 이 때, 소정의 주기 간격으로 임의의 인덱스 값을 발생시킬 수 있다.

[0038] 랜덤 인덱스 값 발생 모듈(200)는, 예를 들어 난수 발생기와 같은 형태로 구성될 수 있다.

[0039] 여기서, 도면에는 랜덤 인덱스 값 발생 모듈(200)을 별도의 하드웨어 모듈 형태로 도시하였지만, 이에 한정하는 것은 아니며 랜덤 인덱스 값 발생 모듈(200)은 외부 서버(100) 자체에 구현될 수도 있고, 보상 모듈(300) 내부에 구현될 수도 있다. 후술하는 보상 모듈(300)의 에러 보정 모듈(330)에서 보정 가중치 값 및 학습 가중치 값을 추출할 수 있도록, 랜덤으로 인덱스 값을 발생시켜 입력해줄 수 있는 형태라면 구현 가능하다.

[0040] 1.3. 보상 모듈(300)

[0041] 보상 모듈은, 미리 저장된 소정의 보정 데이터와 상기 외부 서버(100)로부터 전달 받은 압축된 학습 가중치 데이터를 기반으로, 상기 랜덤 인덱스 값 발생 모듈(200)에 의해 발생된 인덱스 값에 해당하는 보정 데이터를 이용하여 압축된 학습 가중치 데이터를 보정하도록 구성된다. 이러한 보상 모듈(300)은, 아래와 같은 세부 구성을 포함하여 구성될 수 있다.

[0042] 가. 제1 메모리 모듈(310)

[0043] 제1 메모리 모듈(310)은, 미리 획득한 소정의 보정 데이터를 룩업 테이블 형태로 저장하는 구성이다.

[0044] 소정의 보정 데이터는, 예를 들어 도 4의 (e)와 같은 룩업 테이블 형태로 구성될 수 있으며, 여기서 '0',' a', 'b', 'c' 각각을 보정 가중치 값으로 지칭한다. 이 때, 각각의 보정 가중치 값은 앞서 설명한 외부 서버(100)에서 학습된 가중치 값들을 양자화하여 압축할 시 발생하였던 양자화 오차의 발생 빈도 수에 대응하는 비율에 따라 룩업 테이블에 구성된다.

[0045] 도 4를 참조하여 이러한 보정 데이터를 획득하는 과정 및 오차 보상 원리에 대하여 설명하도록 한다.

[0046] 먼저, 학습된 가중치 값들을 함수 그래프로 나타내보면, 도 4의 (a)와 같은 가우시안(정규) 분포를 이루고 있고, 가중치가 음수에서는 0에 가까워질수록 선형(linear)으로 상승하고, 양수에서는 선형으로 감소하는 형태를 보인다.

[0047] 이와 같이 정규분포를 이루는 학습된 가중치 값들을 양자화 기술을 사용하여 압축하는 과정으로서 일정 구간 별로 그 사이에 있는 가중치 값들을 특정 가중치 값으로 압축하는 클러스터링(clustering)을 하면, 도 4의 (b)와 같은 형태의 가우시안(정규) 분포로 표현된다. 이는 외부 서버(100)에서 보상 모듈(300)로 전달되는 압축된 학습 가중치 데이터를 함수 그래프로 표현하였을 때의 상태이다.

[0048] 도 4의 (a)에서 클러스터링(clustering)을 통해 (b)가 되는 과정에서, 도 4의 (c)에 보이는 것처럼 특정 구간 S1과 S2 사이에 속하는 가중치 값들을 S1에 해당하는 값으로 클러스터링, 쉽게 말해 압축하는 것이므로 손실된 값이 발생하게 되고, 이를 본 명세서에서는 양자화 오차로 지칭한다.

[0049] 이는 도 4의 (d)처럼 특정 구간 S1와 S2 사이의 값들을 'a', 'b', 'c'로 지칭하였을 때 이들 'a', 'b', 'c' 값들을 양자화 오차로 이해할 수 있다. 여기서, 도 4의 (b)와 같은 압축된 학습 가중치 값들에 대한 가우시안(정규) 분포에 기반하면 'a', 'b', 'c' 값들 각각의 발생 빈도 수, 즉 손실된 횟수 정도를 획득하는 것이 가능하다.

[0050] 이에 따라, 룩업 테이블에 양자화 오차 'a', 'b', 'c' 값들 각각의 손실 빈도 수에 대응하는 비율로 'a', 'b', 'c'를 보정 가중치 값으로서 구성한다. 이 때, 각 배열마다 해당 보정 가중치 값을 가리키는 인덱스 값이 할당된다.

[0051] 예를 들어, 압축된 학습 가중치 데이터에 대한 전체 가우시안(정규) 분포로부터 구간 사이의 양자화 오차 'a', 'b', 'c' 값들 각각의 손실 빈도 수가 'a < b < c' 인 것을 획득하였다면, 그 빈도 수에 따라 룩업 테이블에 보정 가중치 값으로서 'c' > 'b' > 'a' 순서의 비율로 넣어 배열하고, 각 배열에 인덱스 값 할당한 도 4의 (e)와 같은 룩업 테이블 형태의 보정 데이터를 마련하는 것이다.

[0052] 이러한 보정 데이터를 이용하여, 랜덤 인덱스 값 발생 모듈(200)에 의해 발생되는 임의의 인덱스 값에 따라 후술하는 제2 메모리 모듈(320)에 저장된 압축된 학습 가중치 데이터에서 해당 인덱스 값이 가리키는 학습 가중치 값을 보정 데이터에서 해당 인덱스 값이 가리키는 보정 가중치 값으로 보정하도록 구성되는데, 이 때 임의의 인

덱스 값이 랜덤하게 발생하더라도, 룩업 테이블에 보정 가중치 값 'a', 'b', 'c' 값을 각각의 손실된 횟수에 대응하는 비율로 구성하였기 때문에 가장 많은 비율로 구성되어 있는 'c' 값이 가장 높은 확률로 보정 가중치 값으로서 기능할 것이고, 이에 따라 전체적으로 보면 학습된 가중치 데이터의 압축 시 발생한 양자화 오차를 효과적으로 보상할 수 있다.

[0053] 이와 같은 원리로 획득되는 보정 데이터를 이용하여 압축된 학습 가중치 데이터를 보정하는 방식에 대한 설명은, 후술하는 에러 보정 모듈(330)에서 구체적으로 설명하도록 한다.

[0054] 나. 제2 메모리 모듈(320)

[0055] 제2 메모리 모듈(320)은, 상기 외부 서버(100)로부터 전달되는 압축된 학습 가중치 데이터를 임시 저장하는 임시 저장소로서, 예를 들어 BRAM으로 구현되거나 혹은 버퍼 형태로도 구현될 수 있다. 즉, 외부 서버(100)로부터의 압축된 학습 가중치 데이터를 임시 저장할 수 있는 형태라면 구현 가능하다.

[0056] 이와 같은 제2 메모리 모듈(320)에 저장되는 압축된 학습 가중치 데이터는 앞서 설명한 것과 같이 도 1의 (c)와 같이 각 클러스터 별 학습 가중치 값을 포함하고, 클러스터 별 학습 가중치 값을 가리키는 인덱스 값이 할당되어 있는 형태로 구성되어 있다.

[0057] 예를 들어, 도 1의 (c)의 블록 하단부터 '-1.00', '0.00', '1.50', '2.00'은 각 클러스터 별 학습 가중치 값이고, '0', '1', '2', '3'은 각 클러스터 별 학습 가중치 값을 가리키는 인덱스 값이다.

[0058] 다. 에러 보정 모듈(330)

[0059] 에러 보정 모듈(330)은, 상기 랜덤 인덱스 값 발생 모듈(200)에서 발생시킨 인덱스 값으로 상기 제1 메모리 모듈(310)에 저장된 보정 데이터 및 제2 메모리 모듈(320)에 저장된 압축된 학습 가중치로부터 보정 가중치 값 및 학습 가중치 값을 각각 추출하여, 상기 추출된 보정 가중치 값을 이용하여 상기 추출된 학습 가중치 값을 보정하도록 구성된다.

[0060] (1) 가중치 값 추출 모듈(332)

[0061] 가중치 값 추출 모듈(332)는, 상기 제1 메모리 모듈(310)에 저장된 보정 데이터 및 제2 메모리 모듈(320)에 저장된 압축된 학습 가중치 데이터로부터 상기 랜덤 인덱스 값 발생 모듈(200)에 의해 발생된 인덱스 값에 대응하는 보정 가중치 값 및 학습 가중치 값을 각각 추출할 수 있다.

[0062] 예를 들어, 랜덤 인덱스 값 발생 모듈(200)에서 발생시킨 인덱스 값이 '1' 이라고 한다면, 제2 메모리 모듈(320)에 저장된 도 1의 (c)와 같은 형태의 압축된 학습 가중치 데이터에서 인덱스 값 '1' 이 가리키는 '0.00' 값을 학습 가중치 값으로 추출하고, 제1 메모리 모듈(310)에 저장된 도 4의 (e)와 같은 형태의 보정 데이터에서 인덱스 값 '1' 이 가리키는 'c' 값을 보정 가중치 값으로 추출하는 것이다.

[0063] (2) 연산 모듈(334)

[0064] 연산 모듈(334)은, 상기 가중치 값 추출 모듈(332)에서 임의의 인덱스 값에 대응하는 보정 가중치 값 및 학습 가중치 값을 추출하면, 상기 학습 가중치 값에 상기 보정 가중치 값을 더하거나 혹은 차감하는 방식으로 상기 학습 가중치 값의 양자화 오차를 보정할 수 있다.

[0065] 보다 구체적으로, 추출된 학습 가중치 값이 양수인 경우이면, 상기 학습 가중치 값에서 추출된 보정 가중치 값을 차감하는 방식으로 보정한다.

[0066] 한편, 추출된 학습 가중치 값이 음수인 경우이면, 상기 학습 가중치 값에서 추출된 보정 가중치 값을 더하여 주는 방식으로 보정한다.

[0067] 예를 들어, 가중치 값 추출부(332)에서 학습 가중치 값으로 '0.00' 이 추출되고, 보정 가중치 값으로 'c' 가 추출된 경우, '0.00 - c' 를 하여 해당 학습 가중치 값의 양자화 오차를 보정하는 것이다.

[0068] 다른 예로, 가중치 값 추출부(332)에서 학습 가중치 값으로 '-1.00' 이 추출되고, 보정 가중치 값으로 'c' 가 추출된 경우, '-1.00 + c' 를 하여 해당 학습 가중치 값의 양자화 오차를 보정하는 것이다.

[0069] 한편, 보상 모듈(300)은 상기와 같이 압축된 학습 가중치 데이터를 보정한 후, 가속기(400)로 입력한다.

[0070] 1.4. 가속기(400)

[0071] 가속기(400)는, 학습된 데이터를 가지고 추론을 하도록 인공신경망(artificial neural network) 연산에 최적화

되도록 설계된 공지의 연산 장치로서, 상술한 보상 모듈(300)로부터 전달되는 보정된 학습 가중치 데이터를 이용하여 추론하고자 하는 새로운 입력 데이터에 대한 추론 결과를 출력한다.

[0072] 한편, 이와 같은 보상 모듈(300)과 가속기(400)는 하나의 칩(chip)에 구현될 수 있다.

[0073] 2. 본 발명에 따른 양자화 오차 방법

[0074] 도 5는 본 발명에 따른 양자화 오차 방법의 흐름도를 도시한 도면이다. 도 5를 참조하면, 본 발명의 양자화 오차 방법은 하기의 단계를 포함하여 구성될 수 있다.

[0075] 2.1. 보정 데이터 저장 단계(S100)

[0076] 보정 데이터 저장 단계는, 외부 서버(100)와 가속기(400) 에 구성되는 하드웨어 기반의 보상 모듈(300)에 룩업 테이블 형태로 구성되는 소정의 보정 데이터를 미리 획득하여 저장하는 단계이다. 보다 구체적으로는, 보상 모듈(300)의 제1 메모리 모듈(310)에 저장한다.

[0077] 소정의 보정 데이터는, 예를 들어 도 4의 (e)와 같은 룩업 테이블 형태로서, 외부 서버(100)에서 학습된 가중치 값들을 양자화 하여 압축시켜 압축된 학습 가중치 데이터를 획득하는 과정에서 발생하였던 각 양자화 오차의 발생 빈도 수에 대응하는 비율에 따라 구성된 각각의 보정 가중치 값을 포함하여 구성된다.

[0078] 이와 같은 보정 데이터를 획득하는 원리는, 위 시스템에서 보상 모듈(300)의 제1 메모리 모듈(310) 부분에서 상세하게 설명하였으므로 구체적인 설명은 생략한다.

[0079] 2.2. 압축된 학습 가중치 데이터 수신 단계(S200)

[0080] 압축된 학습 가중치 데이터 수신 단계는, 보상 모듈(300)에서 외부 서버(100)로부터 압축된 학습 가중치 데이터를 수신하는 단계이다.

[0081] 여기서, 압축된 학습 가중치 데이터는, 클러스터(cluster) 별로 중심(centroid) 값이 학습 가중치 값으로 설정되어 있고, 각 학습 가중치 값을 가리키는 인덱스 값이 할당되어 있는 형태로서, 예를 들어 도 1의 (c)와 같은 형태일 수 있다.

[0082] 2.3. 압축된 학습 가중치 데이터 저장 단계(S300)

[0083] 보상 모듈(300)은, 상기 압축된 학습 가중치 데이터 수신 단계(S200)에서 수신한 외부 서버(100)로부터의 압축된 학습 가중치 데이터를 임시 저장하는 압축된 학습 가중치 데이터 저장 단계를 수행할 수 있다.

[0084] 보다 구체적으로는, 보상 모듈(300)의 제2 메모리 모듈(320)에 압축된 학습 가중치 데이터를 임시로 저장할 수 있다.

[0085] 2.4. 랜덤 인덱스 값 발생 단계(S400)

[0086] 랜덤 인덱스 값 발생 단계는, 임의의 인덱스 값을 랜덤으로 발생시키는 단계로서, 랜덤 인덱스 값 발생 모듈(200)에 의해 수행될 수 있다. 이 단계에서 발생된 임의의 인덱스 값은 보상 모듈(300)로 입력되어 후술하는 압축된 학습 가중치 데이터 보정 단계(S500)가 수행될 수 있다.

[0087] 2.5. 압축된 학습 가중치 데이터 보정 단계(S500)

[0088] 보상 모듈(300)은, 상기 랜덤 인덱스 값 발생 단계(S400)를 통해 임의의 인덱스 값이 발생되면, 상기 발생된 인덱스 값으로 상기 보정 데이터 저장 단계(S100)에서 저장된 보정 데이터 및 상기 압축된 학습 가중치 데이터 저장 단계(S300)에서 저장된 압축된 학습 가중치 데이터로부터 보정 가중치 값 및 학습 가중치 값을 각각 추출하여, 상기 추출한 보정 가중치 값을 이용하여 상기 추출한 학습 가중치 값을 보정하도록 구성된다.

[0089] 가. 가중치 값 추출 단계(S510)

[0090] 먼저, 보상 모듈(300)의 가중치 값 추출 모듈(332)에 의해 임의의 인덱스 값을 이용하여 보정 가중치 값 및 학습 가중치 값을 각각 추출하는 가중치 값 추출 단계가 수행될 수 있다.

[0091] 보다 구체적으로, 상기 보정 데이터 저장 단계(S100)에서 저장된 보정 데이터 및 상기 압축된 학습 가중치 데이터 저장 단계(S300)에서 저장된 압축된 학습 가중치 데이터로부터 상기 랜덤 인덱스 값 발생 단계(S400)에서 발생된 인덱스 값이 가리키는 보정 가중치 값 및 학습 가중치 값을 각각 추출할 수 있다.

[0092] 예를 들자면, 상기 랜덤 인덱스 값 발생 단계(S400)에서 발생한 인덱스 값이 ‘1’ 이라고 하면, 상기 보정 데

이터 저장 단계(S100)에서 저장된 도 4의 (e)와 같은 형태의 보정 데이터에서 인덱스 값 '1' 이 가리키는 'c' 값을 보정 가중치 값으로 추출하고, 상기 압축된 학습 가중치 데이터 저장 단계(S300)에서 도 1의 (a)와 같은 형태의 압축된 학습 가중치 데이터에서 인덱스 값 '1' 이 가리키는 '0.00' 값을 학습 가중치 값으로 추출하는 것이다.

[0093] 나. 연산 단계(S520)

[0094] 상기 가중치 값 추출 단계(S510)를 통해 보정 가중치 값 및 학습 가중치 값이 각각 추출되면, 보상 모듈(300)의 연산 모듈(334)은 상기 추출된 학습 가중치 값에 상기 추출된 보정 가중치 값을 더하거나 혹은 차감하는 방식으로 상기 학습 가중치 값의 양자화 오차를 보정하는 연산 단계(S520)를 수행할 수 있다.

[0095] 보다 구체적으로, 상기 가중치 값 추출 단계(S510)에서 추출된 학습 가중치 값이 양수인 경우이면, 상기 학습 가중치 값에서 추출된 보정 가중치 값을 차감하는 방식으로 보정한다.

[0096] 한편, 상기 가중치 값 추출 단계(S510)에서 추출된 학습 가중치 값이 음수인 경우이면, 상기 학습 가중치 값에서 추출된 보정 가중치 값을 더하여 주는 방식으로 보정한다.

[0097] 예를 들어, 가중치 값 추출 단계(S510)에서 학습 가중치 값으로 '0.00' 이 추출되고, 보정 가중치 값으로 'c' 가 추출된 경우, '0.00 - c' 를 하여 해당 학습 가중치 값의 양자화 오차를 보정하는 것이다.

[0098] 다른 예로, 가중치 값 추출 단계(S510)에서 학습 가중치 값으로 '-1.00' 이 추출되고, 보정 가중치 값으로 'c' 가 추출된 경우, '-1.00 + c' 를 하여 해당 학습 가중치 값의 양자화 오차를 보정하는 것이다.

[0099] 이와 같은 방식으로 압축된 학습 가중치 데이터의 압축 시 발생한 양자화 오차를 보정한 후, 보상 모듈(300)은 보정된 학습 가중치 데이터를 가속기(400)로 전달해주는 보정된 학습 가중치 데이터 전달 단계(미도시)를 수행할 수 있다.

[0100] 한편, 본 발명의 기술적 사상은 상기 실시 예에 따라 구체적으로 기술되었으나, 상기 실시 예는 그 설명을 위한 것이며, 그 제한을 위한 것이 아님을 주지해야 한다. 또한, 본 발명의 기술분야에서 당업자는 본 발명의 기술사상의 범위 내에서 다양한 실시 예가 가능함을 이해할 수 있을 것이다.

부호의 설명

[0101] 100: 외부 서버

200: 랜덤 인덱스 값 발생 모듈

300: 보상 모듈

310: 제1 메모리 모듈

320: 제2 메모리 모듈

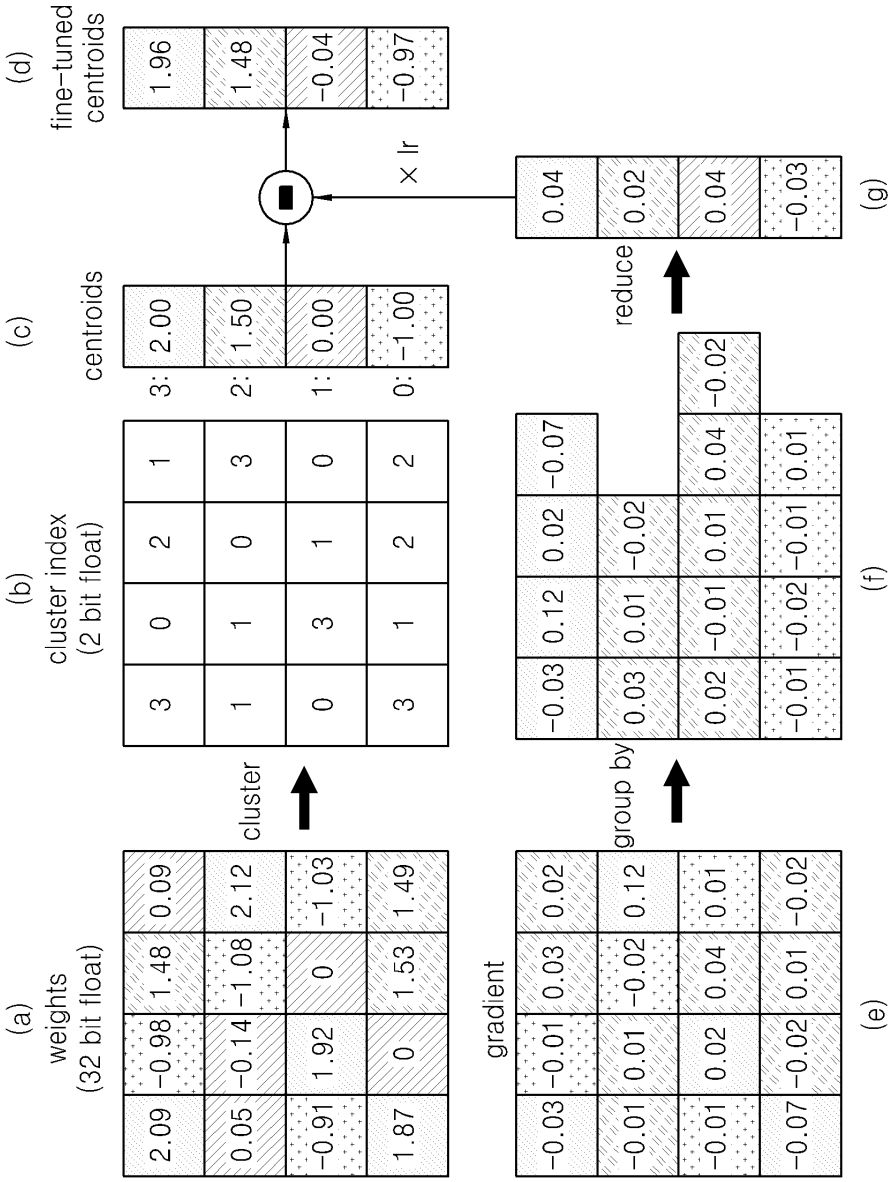
330: 에러 보정 모듈

332: 가중치 값 추출 모듈

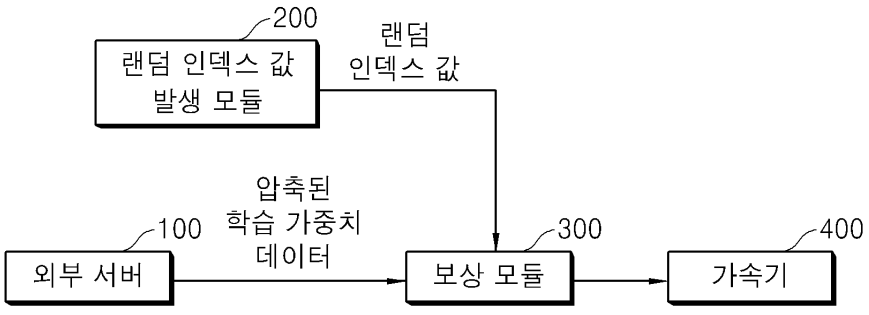
334: 연산 모듈

도면

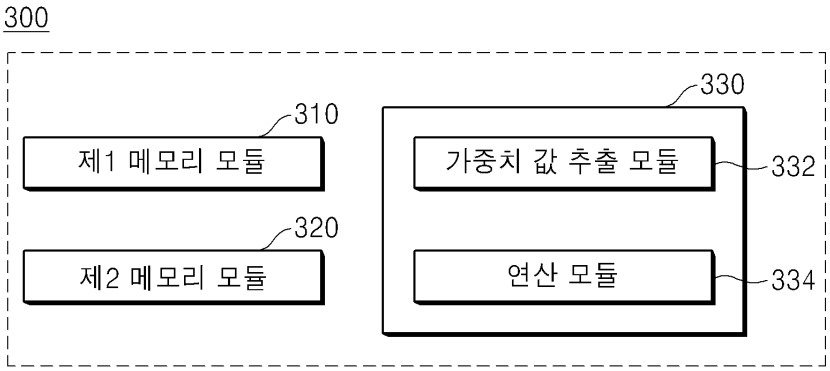
도면1


(a)
weights
(32 bit float)
2.09 -0.98 1.48 0.09
0.05 -0.14 -1.08 2.12
-0.91 1.92 0 -1.03
1.87 0 1.53 1.49
cluster
(b)
cluster index
(2 bit float)
3 0 2 1
1 1 0 3
0 3 1 0
3 1 2 2
(c)
centroids
3: 2.00
2: 1.50
1: 0.00
0: -1.00
(d)
fine-tuned
centroids
1.96 1.48 -0.04 -0.97
× lr
0.04 0.02 0.04 -0.03
(g)
reduce
(e)
gradient
-0.03 -0.01 0.03 0.02
-0.01 0.01 -0.02 0.12
-0.01 0.02 0.04 0.01
-0.07 -0.02 0.01 -0.02
group by
(f)
-0.03 0.12 0.02 -0.07
0.03 0.01 -0.02
0.02 -0.01 0.01 0.04 -0.02
-0.01 -0.02 -0.01 0.01

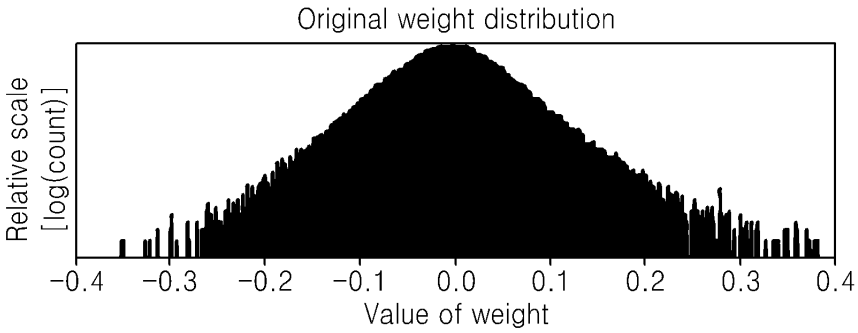
도면2


200
랜덤 인덱스 값
발생 모듈
랜덤
인덱스 값
압축된
학습 가중치
데이터
100
외부 서버
300
보상 모듈
400
가속기

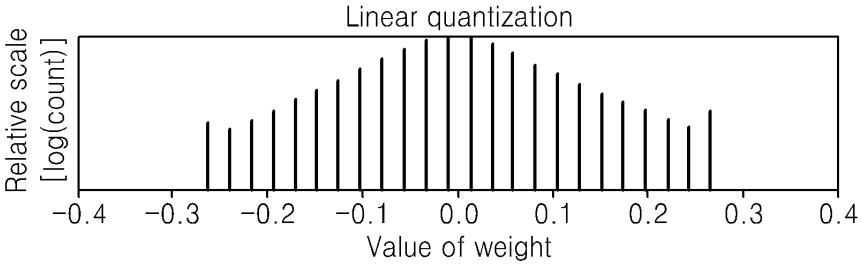
도면3



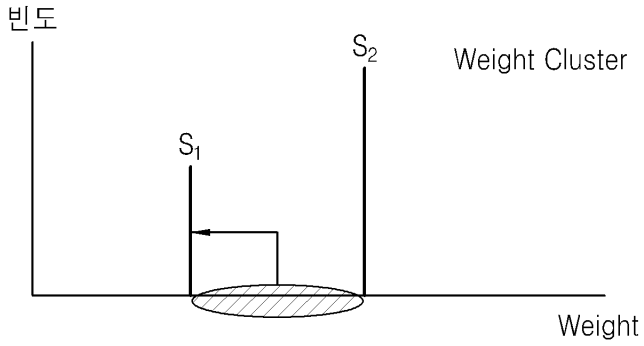
도면4a



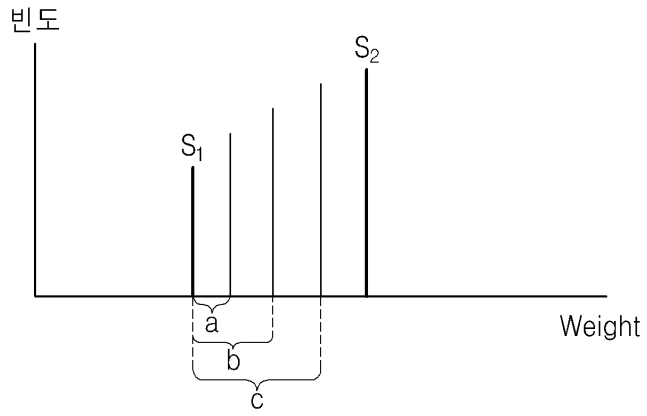
도면4b



도면4c



도면4d



도면4e

9:	0
8:	a
7:	a
6:	b
5:	b
4:	b
3:	c
2:	c
1:	c
0:	c

도면5