(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0089555
(43) 공개일자 2023년06월20일

(51) 국제특허분류(Int. Cl.)
G06N 3/0495 (2023.01) *G06N 3/063* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/0495 (2023.01)
G06N 3/063 (2013.01)
(21) 출원번호 10-2022-0172924
(22) 출원일자 2022년12월12일
심사청구일자 2022년12월12일
(30) 우선권주장
1020210177672 2021년12월13일 대한민국(KR)

(71) 출원인
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
박은혁
경상북도 포항시 남구 청암로 77
(74) 대리인
특허법인이상

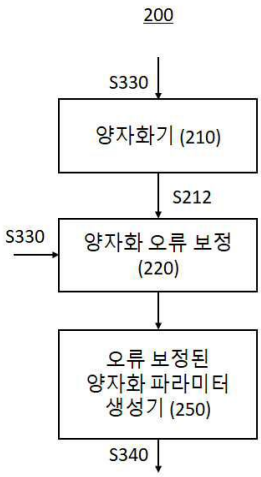
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 **뉴럴 네트워크 양자화 오류 보정 장치 및 방법**

(57) 요 약

본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치는, 뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하고, 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하고, 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정하고, 보정된 제2 파라미터를 제3 파라미터로서 출력한다.

대 표 도 - 도4



(52) CPC특허분류
G06N 3/082 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711134476
과제번호	2021-0-00310-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	인공지능 학습/추론 효율성 향상을 위한 서버용 SW 프레임워크 개발
기 여 율	1/1
과제수행기관명	에스케이텔레콤(주)
연구기간	2021.04.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

프로세서(processor); 및

프로세서를 통해 실행되는 적어도 하나의 명령이 저장된 메모리(memory); 를 포함하고,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하고,

상기 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하고,

상기 제1 파라미터의 통계적 정보와 상기 제2 파라미터의 통계적 정보에 기반하여 상기 제2 파라미터를 보정하고,

상기 보정된 제2 파라미터를 제3 파라미터로서 출력하는,

뉴럴 네트워크 양자화 오류 보정 장치.

청구항 2

제1항에 있어서,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이를 최소화하도록 상기 제2 파라미터 각각을 보정하는,

뉴럴 네트워크 양자화 오류 보정 장치.

청구항 3

제2항에 있어서,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이에 기반하여 상기 제2 파라미터 각각을 보정하는,

뉴럴 네트워크 양자화 오류 보정 장치.

청구항 4

제1항에 있어서,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이를 최소화하고, 상기 제1 파라미터의 표준편차와 상기 제2 파라미터의 표준편차 간의 차이를 최소화하도록 상기 제2 파라미터 각각을 보정하는,

뉴럴 네트워크 양자화 오류 보정 장치.

청구항 5

제4항에 있어서,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 제1 차이; 및

상기 제1 파라미터의 표준편차와 상기 제2 파라미터의 표준편차 간의 제2 차이;

에 기반하여 상기 제2 파라미터 각각을 보정하는,

뉴럴 네트워크 양자화 오류 보정 장치.

청구항 6

메모리(memory)에 저장되는 적어도 하나의 명령을 실행하는 프로세서(processor)에 의하여 수행되는 방법으로서,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하는 단계;

상기 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하는 단계;

상기 제1 파라미터의 통계적 정보와 상기 제2 파라미터의 통계적 정보에 기반하여 상기 제2 파라미터를 보정하는 단계; 및

상기 보정된 제2 파라미터를 제3 파라미터로서 출력하는 단계;

를 포함하는,

뉴럴 네트워크 양자화 오류 보정 방법.

청구항 7

제6항에 있어서,

상기 제2 파라미터를 보정하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이를 최소화하도록 상기 제2 파라미터 각각을 보정하는 단계;

를 포함하는,

뉴럴 네트워크 양자화 오류 보정 방법.

청구항 8

제7항에 있어서,

상기 제2 파라미터 각각을 보정하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이에 기반하여 상기 제2 파라미터 각각을 보정하는,

뉴럴 네트워크 양자화 오류 보정 방법.

청구항 9

제6항에 있어서,

상기 제2 파라미터를 보정하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이를 최소화하고, 상기 제1 파라미터의 표준편차와 상기 제2 파라미터의 표준편차 간의 차이를 최소화하도록 상기 제2 파라미터 각각을 보정하는 단계;

를 포함하는,

뉴럴 네트워크 양자화 오류 보정 방법.

청구항 10

제9항에 있어서,

상기 제2 파라미터 각각을 보정하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 제1 차이; 및

상기 제1 파라미터의 표준편차와 상기 제2 파라미터의 표준편차 간의 제2 차이;

에 기반하여 상기 제2 파라미터 각각을 보정하는,

뉴럴 네트워크 양자화 오류 보정 방법.

청구항 11

메모리(memory)에 저장되는 적어도 하나의 명령을 실행하는 프로세서(processor)에 의하여 수행되는 방법으로서,

상기 프로세서가 상기 적어도 하나의 명령을 실행함으로써,

뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하는 단계;

상기 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하는 단계;

상기 제1 파라미터의 통계적 정보와 상기 제2 파라미터의 통계적 정보에 기반하여 상기 제2 파라미터를 보정함으로써 제3 파라미터를 생성하는 단계; 및

상기 제3 파라미터에 기반하여 입력 데이터에 대한 추론 결과를 생성하는 단계;

를 포함하는,

양자화 오류가 보정된 뉴럴 네트워크의 동작 방법.

청구항 12

제11항에 있어서,

상기 제3 파라미터에 기반하여 새로운 뉴럴 네트워크를 생성하는 단계;

를 더 포함하고,

상기 추론 결과를 생성하는 단계는

상기 새로운 뉴럴 네트워크에 상기 입력 데이터를 입력하는 단계; 및

상기 새로운 뉴럴 네트워크의 출력을 상기 추론 결과로서 생성하는 단계;

를 포함하는,

양자화 오류가 보정된 뉴럴 네트워크의 동작 방법.

청구항 13

제11항에 있어서,

상기 제3 파라미터를 상기 뉴럴 네트워크에 전달하는 단계; 및

상기 제3 파라미터에 기반하여 상기 뉴럴 네트워크의 모든 파라미터를 갱신하는 단계;

를 더 포함하고,

상기 추론 결과를 생성하는 단계는

상기 모든 파라미터가 갱신된 뉴럴 네트워크에 상기 입력 데이터를 입력하는 단계; 및

상기 뉴럴 네트워크의 출력을 상기 추론 결과로서 생성하는 단계;

를 포함하는,

양자화 오류가 보정된 뉴럴 네트워크의 동작 방법.

청구항 14

제11항에 있어서,

상기 제3 파라미터를 생성하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이를 최소화하도록 상기 제2 파라미터 각각을 보정하는 단계;

를 포함하는,

양자화 오류가 보정된 뉴럴 네트워크의 보정 방법.

청구항 15

제14항에 있어서,

상기 제2 파라미터 각각을 보정하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이에 기반하여 상기 제2 파라미터 각각을 보정하는,

양자화 오류가 보정된 뉴럴 네트워크의 보정 방법.

청구항 16

제11항에 있어서,

상기 제3 파라미터를 생성하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 차이를 최소화하고, 상기 제1 파라미터의 표준편차와 상기 제2 파라미터의 표준편차 간의 차이를 최소화하도록 상기 제2 파라미터 각각을 보정하는 단계;

를 포함하는,

양자화 오류가 보정된 뉴럴 네트워크의 보정 방법.

청구항 17

제16항에 있어서,

상기 제2 파라미터 각각을 보정하는 단계는,

상기 제1 파라미터의 평균과 상기 제2 파라미터의 평균 간의 제1 차이; 및

상기 제1 파라미터의 표준편차와 상기 제2 파라미터의 표준편차 간의 제2 차이;

에 기반하여 상기 제2 파라미터 각각을 보정하는,

양자화 오류가 보정된 뉴럴 네트워크의 보정 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 양자화를 통해 뉴럴 네트워크를 최적화/경량화하는 시스템에 관한 것으로, 뉴럴 네트워크의 양자화 오류 보정 기능을 포함하는 뉴럴 네트워크, 뉴럴 네트워크의 양자화, 및 양자화된 뉴럴 네트워크를 이용한 추론 프로세스에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 실시예에 대한 배경 정보를 제공할 뿐 종래 기술을 구성하는 것은 아니다.

[0003] 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. 딥

러닝은 인간의 생물학적 신경 세포의 특성을 수학적 표현에 의해 모델링 한 인공신경망(artificial neural network)을 깊게 구성하여 학습하는 방법을 말한다.

[0004] 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.

[0005] 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 구분된다.

[0006] 알고리즘 경량화 방식은 인공신경망을 압축하는 여러 기술이 적용될 수 있는데, 주로 양자화(quantization) 하여 가중치(weight)로 저장하는 bit를 최소화하는 방법을 사용하고 있다.

[0007] 양자화 과정에서는 floating point 표현형으로 표현된 가중치를 보다 작은 bit의 표현형으로 양자화하는 과정이 대표적으로 이용된다. 이때 양자화된 이후의 가중치는 원본 가중치의 정보를 일부 손실하므로, 양자화 과정은 일종의 손실 압축에 해당한다고 볼 수 있다.

[0008] 이처럼 양자화 과정에서 가중치 정보의 손실이 발생하므로, 종래 기술들은 양자화 이후의 가중치 파라미터를 통계적인 보조 정보로 기술하고자 노력하고 있다. 예를 들어 평균값, 중간값, 분산/표준편차 등이 양자화된 이후의 가중치 집합을 기술하기 위해 부가적으로 포함되고 있다.

[0009] 일반적으로 뉴럴 네트워크 학습 후 양자화가 이루어지는 경우 정확도 손실을 만회하기 위한 방법은 양자화 후 학습(quantization-aware training) 또는 양자화 후 별도의 학습 없이 에러를 보정하여 활용하는 학습 후 양자화(post-training quantization)으로 나눌 수 있다.

[0010] 이 중 양자화 후 학습 방법의 경우 학습을 통해 양자화된 네트워크의 오류를 보정하고 normalization layer의 running mean/variance를 보정하는 단계가 포함되어 있어 양자화 오류를 보정하기 위한 과정이 복잡하며 이로 인하여 경량화라는 목적에 부합하지 못하는 문제가 있다.

[0011] 양자화 오류를 보정하는 또 다른 시도로서 [1] 한국공개특허 KR 10-2021-0035702호 "인공 신경망의 양자화 방법 및 인공 신경망을 이용한 연산 방법" 등이 제안되기도 하였다.

[0012] 상기 선행문헌 [1]에서는 인공 신경망을 구동하는 신경망 시스템 및 인공 신경망을 양자화하는 양자화 시스템을 포함하고, 양자화 시스템은, 인공 신경망의 파라미터들을 양자화함으로써, 인공 신경망의 양자화된 파라미터들을 생성하고, 인공 신경망의 파라미터들 및 양자화된 파라미터들을 기초로 인공 신경망의 파라미터들의 양자화 오차를 생성하고, 양자화된 파라미터들 및 인공 신경망의 파라미터들의 양자화 오차를 기초로 보정 바이어스를 생성하고, 생성한 양자화된 파라미터들 및 보정 바이어스를 신경망 시스템에 전송할 수 있다.

[0013] 선행문헌 [1]에서는 양자화된 입력 샘플 및 양자화된 파라미터들을 기초로 제1 MAC(multiply-accumulate) 연산을 수행하고, 제1 MAC 연산의 결과에 보정 바이어스를 반영함으로써 최종 연산 결과를 생성하는 구성을 제안한다.

[0014] 이 방식은 양자화 파라미터를 직접적으로 보정하는 것이 아니고 양자화 오차에 기반하여 추론 결과가 어떻게 영향을 받을 지를 별도로 학습하거나 분석하는 복잡한 과정을 필요로 한다. 따라서 대부분의 종래 기술들처럼 양자화 오류를 보정하기 위한 과정이 오히려 신경망을 더욱 복잡화하는 문제점이 반복되고 있다.

선행기술문헌

특허문헌

[0015] (특허문헌 0001) 한국공개특허 KR 10-2021-0035702호 "인공 신경망의 양자화 방법 및 인공 신경망을 이용한 연산 방법" (공개일 2021년 4월 1일)

발명의 내용

해결하려는 과제

[0016] 상기와 같은 문제점을 해결하기 위한 본 발명의 목적은, 학습 후 양자화 (post-training quantization) 기법으로 양자화된 이후의 파라미터를 이용한 별도의 학습이 필요하지 않으면서도 성능을 향상한 뉴럴 네트워크 양자화 오류 보정 장치 및 방법을 제공하는 것이다.

[0017] 본 발명의 목적은 양자화 이후 가중치의 양자화에서 발생하는 통계적 분포의 오류를 보정하는 알고리즘을 통해 정확도를 보전하면서 성능을 향상한 뉴럴 네트워크 양자화 오류 보정 장치 및 방법을 제공하는 것이다.

[0018] 본 발명의 목적은 제안된 양자화 오류 기법을 적용한 뉴럴 네트워크 및 그 운용 방법을 제공하는 것이다.

과제의 해결 수단

[0019] 본 발명의 목적을 달성하기 위한 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치는, 프로세서(processor); 및 프로세서를 통해 실행되는 적어도 하나의 명령이 저장된 메모리(memory)를 포함하고, 프로세서가 적어도 하나의 명령을 실행함으로써, 뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하고, 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하고, 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정하고, 보정된 제2 파라미터를 제3 파라미터로서 출력한다.

[0020] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하도록 제2 파라미터 각각을 보정할 수 있다.

[0021] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다.

[0022] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하고, 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 차이를 최소화하도록 제2 파라미터 각각을 보정할 수 있다.

[0023] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 제1 차이; 및 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 제2 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다.

[0024] 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 방법은 메모리(memory)에 저장되는 적어도 하나의 명령을 실행하는 프로세서(processor)에 의하여 수행되는 방법으로서, 프로세서가 적어도 하나의 명령을 실행함으로써, 뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하는 단계; 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하는 단계; 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정하는 단계; 및 보정된 제2 파라미터를 제3 파라미터로서 출력하는 단계를 포함한다.

[0025] 제2 파라미터를 보정하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하도록 제2 파라미터 각각을 보정하는 단계를 포함할 수 있다.

[0026] 제2 파라미터 각각을 보정하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다.

[0027] 제2 파라미터를 보정하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하고, 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 차이를 최소화하도록 제2 파라미터 각각을 보정하는 단계를 포함할 수 있다.

[0028] 제2 파라미터 각각을 보정하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 제1 차이; 및 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 제2 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다.

[0029] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 메모리(memory)에 저장되는 적어도 하나의 명령을 실행하는 프로세서(processor)에 의하여 수행되는 방법으로서, 프로세서가 적어도 하나의 명령을 실행함으로써, 뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하는 단계; 뉴럴 네트워크의 양자화된 이후의 제2 파라미터를 수신하는 단계; 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정함으로써 제3 파라미터를 생성하는 단계; 및 제3 파라미터에 기반하여 입력 데이터에 대

한 추론 결과를 생성하는 단계를 포함한다.

[0030] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 제3 파라미터에 기반하여 새로운 뉴럴 네트워크를 생성하는 단계를 더 포함할 수 있다.

[0031] 추론 결과를 생성하는 단계는 새로운 뉴럴 네트워크에 입력 데이터를 입력하는 단계; 및 새로운 뉴럴 네트워크의 출력을 추론 결과로서 생성하는 단계를 포함할 수 있다.

[0032] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 제3 파라미터를 뉴럴 네트워크에 전달하는 단계; 및 제3 파라미터에 기반하여 뉴럴 네트워크의 모든 파라미터를 갱신하는 단계를 더 포함할 수 있다.

[0033] 추론 결과를 생성하는 단계는 모든 파라미터가 갱신된 뉴럴 네트워크에 입력 데이터를 입력하는 단계; 및 뉴럴 네트워크의 출력을 추론 결과로서 생성하는 단계를 포함할 수 있다.

[0034] 제3 파라미터를 생성하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하도록 제2 파라미터 각각을 보정하는 단계를 포함할 수 있다.

[0035] 제2 파라미터 각각을 보정하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다.

[0036] 제3 파라미터를 생성하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하고, 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 차이를 최소화하도록 제2 파라미터 각각을 보정하는 단계를 포함할 수 있다.

[0037] 제2 파라미터 각각을 보정하는 단계는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 제1 차이; 및 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 제2 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다.

발명의 효과

[0038] 본 발명의 실시예에 따르면, 양자화 전 후 가중치 양자화로 인해 발생하는 평균 및 분산의 오차를 보정함으로써 양자화 이후의 정확도 손실을 최소화할 수 있다.

[0039] 본 발명의 실시예에 따르면, 학습 후 양자화 (post-training quantization) 기법으로 양자화된 이후의 파라미터를 이용한 별도의 학습이 필요하지 않으면서도 뉴럴 네트워크의 성능을 향상할 수 있다.

[0040] 본 발명의 실시예에 따르면, 통계 분포의 특성에 기반하는 제안된 양자화 오류 기법을 적용한 뉴럴 네트워크를 효과적으로 운용할 수 있다.

도면의 간단한 설명

[0041] 도 1은 데이터 양자화 과정에서 발생하는 양자화 에러를 도시하는 그래프이다.

도 2는 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 기법이 적용된 뉴럴 네트워크 및 뉴럴 네트워크의 동작/운용 방법을 도시하는 개념도이다.

도 3은 본 발명의 다른 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 기법이 적용된 뉴럴 네트워크 및 뉴럴 네트워크의 동작/운용 방법을 도시하는 개념도이다.

도 4는 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치 및 방법을 도시하는 개념도이다.

도 5는 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치 및 방법의 세부적인 구성을 도시하는 개념도이다.

도 6은 본 발명의 다른 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치 및 방법의 세부적인 구성을 도시하는 개념도이다.

도 7은 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치의 가중치 양자화 이후 보정을 통한 정확도 손실 최소화 성능을 나타낸 그래프이다.

도 8은 본 발명의 다른 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치의 가중치 양자화 이후 보정을 통한 정확도 손실 최소화 성능을 나타낸 그래프이다.

도 9는 도 1 내지 도 8의 과정의 적어도 일부를 수행할 수 있는 일반화된 뉴럴 네트워크 양자화 오류 보정 장치, 또는 컴퓨팅 시스템의 예시를 도시하는 개념도이다.

발명을 실시하기 위한 구체적인 내용

[0042] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0043] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. '및/또는' 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0044] 본 출원의 실시예들에서, "A 및 B 중에서 적어도 하나"는 "A 또는 B 중에서 적어도 하나" 또는 "A 및 B 중 하나 이상의 조합들 중에서 적어도 하나"를 의미할 수 있다. 또한, 본 출원의 실시예들에서, "A 및 B 중에서 하나 이상"은 "A 또는 B 중에서 하나 이상" 또는 "A 및 B 중 하나 이상의 조합들 중에서 하나 이상"을 의미할 수 있다.

[0045] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0046] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0047] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가진 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0048] 한편 본 출원일 전에 공지된 기술이라 하더라도 필요 시 본 출원 발명의 구성의 일부로서 포함될 수 있으며, 이에 대해서는 본 발명의 취지를 흐리지 않는 범위 내에서 본 명세서에서 설명한다. 다만 본 출원 발명의 구성을 설명함에 있어, 본 출원일 전에 공지된 기술로서 당업자가 자명하게 이해할 수 있는 사항에 대한 자세한 설명은 본 발명의 취지를 흐릴 수 있으므로, 공지 기술에 대한 지나치게 자세한 사항의 설명은 생략한다.

[0049] 예를 들어, 뉴럴 네트워크의 파라미터(가중치 등)를 양자화하는 기술 등은 본 발명의 출원 전 공지 기술을 이용할 수 있으며, 이들 공지 기술들 중 적어도 일부는 본 발명을 실시하는 데에 필요한 요소 기술로서 적용될 수 있다.

[0050] 그러나 본 발명의 취지는 이들 공지 기술에 대한 권리를 주장하고자 하는 것이 아니며 공지 기술의 내용은 본 발명의 취지에 벗어나지 않는 범위 내에서 본 발명의 일부로서 포함될 수 있다.

[0051] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 본 발명을 설명함에 있어 전체적인 이해를 용이하게 하기 위하여 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.

[0052] 도 1은 데이터 양자화 과정에서 발생하는 양자화 에러를 도시하는 그래프이다.

[0053] 도 1을 참조하면, 양자화 과정에 의하여 연속적으로 분포하는 가중치 값들이 양자화 오류를 가진 채로 양자화되는 과정이 도시된다.

[0054] 양자화 구간 내의 가중치 값들은 양자화된 대표값으로 매핑되고, 이 과정에서 양자화 전의 원본 가중치와 양자

화 이후의 양자화된 가중치 사이에는 양자화 오류가 발생한다.

[0055] 양자화로 인하여 연속적으로 분포하는 가중치 값들의 평균에서 가장 먼 구간에 속하는 원본 값들은 Truncation 되고 이로 인한 Truncation error 또한 일종의 양자화 오류로 볼 수 있다.

[0056] 이러한 양자화 오류를 최소화하기 위한 종래 기술들의 한 종류는 학습 후 양자화 기법으로 발전하였다.

[0057] 종래의 학습 후 양자화 기법의 경우 양자화 오류를 최소화하기 위하여 주어진 가중치와 양자화된 가중치 사이의 오차를 최소화하는 양자화 알고리즘에 집중하여 설계된다. 양자화 후 가중치와 원본 가중치와의 차이로 인하여 정확도 손실이 발생할 수 있는데, 이러한 정확도 손실의 원인은 양자화된 가중치의 제한된 표현형으로 인한 손실과 양자화된 가중치가 원본 가중치와 다른 통계값 (평균/분산)을 가짐으로 인하여 뉴럴 네트워크를 거치면서 오차가 누적되어 발생하는 손실로 나눌 수 있다. 종래 기술들은 후자의 오차를 batch normalization layer의 running mean/variance 등을 보정하여 양자화를 위한 오차를 최소화하고자 시도되었다.

[0058] 본 발명의 일 실시예에서는 해당 오류를 직접적으로 보정함으로써 normalization layer등의 보정 없이 에러를 최소화할 수 있는 알고리즘을 제안한다.

[0059] 도 2는 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 기법이 적용된 뉴럴 네트워크 및 뉴럴 네트워크의 동작/운용 방법을 도시하는 개념도이다.

[0060] 이하 본 발명의 명세서에서 뉴럴 네트워크의 "파라미터"는 가중치를 의미할 수 있다. 본 발명의 변형된 실시예에서는 뉴럴 네트워크의 파라미터는 가중치 및 액티베이션 파라미터를 의미할 수 있다. 본 발명의 실시예들에서 파라미터의 양자화 과정은 가중치의 양자화를 의미할 수도 있고, 가중치 및 액티베이션 파라미터의 양자화를 의미할 수 있다.

[0061] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 뉴럴 네트워크의 양자화되기 전 제1 파라미터를 수신하는 단계(S330); 오류가 보정된 양자화기(200)를 통하여 양자화 오류가 보정된 제3 파라미터를 생성하는 단계(S340); 및 제3 파라미터에 기반하여 입력 데이터에 대한 추론 결과를 생성하는 단계(S360)를 포함한다.

[0062] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 제3 파라미터에 기반하여 새로운 뉴럴 네트워크(120)를 생성하는 단계를 더 포함할 수 있다.

[0063] 추론 결과를 생성하는 단계(S360)는 새로운 뉴럴 네트워크(120)에 입력 데이터를 입력하는 단계(S350); 및 새로운 뉴럴 네트워크(120)의 출력을 추론 결과로서 생성하는 단계(S360)를 포함할 수 있다.

[0064] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 양자화 전의 뉴럴 네트워크(100)에 학습 데이터를 입력하는 단계(S310); 및 학습 데이터에 기반하여 뉴럴 네트워크(100)가 학습하도록 뉴럴 네트워크(100)를 제어하는 단계(S320)를 더 포함할 수 있다.

[0065] 도 3은 본 발명의 다른 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 기법이 적용된 뉴럴 네트워크 및 뉴럴 네트워크의 동작/운용 방법을 도시하는 개념도이다.

[0066] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크(100)의 동작 방법은 오류가 보정된 양자화기(200)에서 뉴럴 네트워크(100)의 양자화되기 전 제1 파라미터를 수신하는 단계(S330); 오류가 보정된 양자화기(200)에서 양자화 오류가 보정된 제3 파라미터를 생성하는 단계(S340); 및 제3 파라미터에 기반하여 입력 데이터에 대한 추론 결과를 생성하는 단계(S360)를 포함한다.

[0067] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크(100)의 동작 방법은 제3 파라미터를 뉴럴 네트워크(100)에 전달하는 단계(S340); 및 제3 파라미터에 기반하여 뉴럴 네트워크(100)의 모든 파라미터를 갱신하는 단계를 더 포함할 수 있다.

[0068] 추론 결과를 생성하는 단계(S360)는 모든 파라미터가 갱신된 뉴럴 네트워크(100)에 입력 데이터를 입력하는 단계(S350); 및 뉴럴 네트워크(100)의 출력을 추론 결과로서 생성하는 단계(S360)를 포함할 수 있다.

[0069] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크의 동작 방법은 양자화 전의 뉴럴 네트워크(100)에 학습 데이터를 입력하는 단계(S310); 및 학습 데이터에 기반하여 뉴럴 네트워크(100)가 학습하도록 뉴럴 네트워크(100)를 제어하는 단계(S320)를 더 포함할 수 있다.

[0070] 도 4는 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220) 및 방법을 도시하는

개념도이다.

[0071] 본 발명의 목적을 달성하기 위한 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220)는, 프로세서(processor); 및 프로세서를 통해 실행되는 적어도 하나의 명령이 저장된 메모리(memory)를 포함하고, 프로세서가 적어도 하나의 명령을 실행함으로써, 뉴럴 네트워크(100)의 양자화되기 전 제1 파라미터를 수신하고(S330), 뉴럴 네트워크(100)의 제1 파라미터가 양자화기(210)에서 양자화된 이후의 제2 파라미터를 수신하고(S212), 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정하고(220), 보정된 제2 파라미터를 제3 파라미터로서 출력한다(250).

[0072] 본 발명의 뉴럴 네트워크 양자화 오류 보정 방법은 딥러닝 최적화, 특히 양자화(quantization)와 관련하여 뉴럴 네트워크를 학습시킨 후(S320) 가중치(weight)에 양자화(quantization)를 적용하여 최적화를 적용한 후(210), 에러의 누적으로 인해 발생하는 정확도 손실을 최소화하도록 보정함으로써(220) 정확도를 유지시킬 수 있는 알고리즘을 제안한다.

[0073] 본 발명의 뉴럴 네트워크 양자화 오류 보정 장치(220)는 convolution 및 fully-connected layer 모두에 적용 가능하다. 일반적으로 convolution layer의 가중치는 4차원, fully-connected layer의 가중치는 2차원 데이터를 가지지만 본 발명의 경우 fan-in과 fan-out 차원으로 나누어서 보정을 적용하므로 편의를 위해 2차원 가중치를 가정한다.

[0074] 원본 가중치를 $W_{o,i}$, 양자화 이후의 가중치를 $Q(W_{o,i})$ 라 할 때, 해당 가중치의 통계값 보정을 위해 2단계의 보정을 적용한다.

[0075] 도 5는 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220) 및 양자화 오류 방법의 세부적인 구성을 도시하는 개념도이다.

[0076] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하도록 제2 파라미터 각각을 보정할 수 있다(230).

[0077] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다(230).

[0078] 2단계 보정 중 첫 번째로, 양자화된 가중치와 원본 가중치의 출력 차원 별 평균의 차이를 최소화하기 위하여 다음과 같은 수학식 1의 보정을 적용한다.

[0079] [수학식 1]

[0080] $$Q_{mean}(W_{i,j}) = Q(W_{i,j}) + E_j[W_{i,j} - Q(W_{i,j})]$$

[0081] 이때 평균에 기반하여 보정된 가중치값 Q_{mean} 은 출력 차원에 대한 평균을 원본 가중치와 동일하게 유지할 수 있으므로 하기 수학식 2에 의하여 나타내어지는 양자화 후 정확도 손실을 줄일 수 있다.

[0082] [수학식 2]

[0083] $$(E_j[Q_{mean}(W_{i,j})] = E_j[W_{i,j}])$$

[0084] 도 6은 본 발명의 다른 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220) 및 양자화 오류 보정 방법의 세부적인 구성을 도시하는 개념도이다.

[0085] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하고, 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 차이를 최소화하도록 제2 파라미터 각각을 보정할 수 있다(240).

[0086] 프로세서가 적어도 하나의 명령을 실행함으로써, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 제1 차이; 및 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 제2 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다(240).

[0087] 2단계 보정 중 두 번째로 양자화된 가중치와 원본 가중치의 출력 차원 별 평균 및 분산의 차이를 모두 최소화하기 위하여 다음과 같은 수학식 3의 보정을 적용한다.

[0088] [수학식 3]

[0089] $$Q_{std,mean}(W_{i,j}) = \frac{\sigma_W}{\sigma_Q} Q(W_{i,j}) + E_j[W_{i,j} - \frac{\sigma_W}{\sigma_Q} Q(W_{i,j})]$$

[0090] 이때 σ_W, σ_Q 는 원본 가중치 및 양자화된 가중치의 표준 편차이다. 평균과 표준편차가 모두 고려되어 보정된 가중치값 $Q_{std,mean}$ 은 원본 가중치의 출력 차원에 대한 평균 및 분산을 동일하게 유지할 수 있으므로 양자화 후 정확도 손실을 최소화할 수 있다.

[0091] 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220)의 뉴럴 네트워크 양자화 오류 보정 방법의 작동을 설명하면 다음과 같다.

[0092] 본 발명은 가중치에 양자화를 적용한 이후 양자화된 가중치가 원본 가중치와 다른 통계값(평균/분산)을 가짐으로 인하여 오차가 누적되어 정확도 손실이 발생할 수 있다.

[0093] 본 발명은 양자화 전후의 통계값 차이를 보정해주는 알고리즘을 통해 양자화 정확도 손실을 최소화하였다.

[0094] 도 7은 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220)의 가중치 양자화 이후 보정을 통한 정확도 손실 최소화 성능을 나타낸 그래프이다.

[0095] 도 7에 도시된 일 실시예는 MobileNet-v2 네트워크를 이용하여 양자화 및 오류 보정 과정을 수행하였다.

[0096] 도 7의 그래프의 x축은 양자화 레벨, y축은 정확도를 나타내며, 왼쪽 그래프는 CIFAR-100 데이터셋을, 오른쪽 그래프는 CIFAR-10 데이터셋을 이용한 실험 결과가 도시된다.

[0097] 도 7을 참조하면, MobileNet-v2 네트워크를 이미지 분류 작업에 대하여 학습시킨 후 가중치 양자화를 적용했을 때 정확도 손실을 측정한 결과가 도시된다. 다양한 가중치 양자화 레벨에 대하여 정확도 트렌드가 도시된다. 양자화 레벨이 낮을수록 양자화로 인한 손실이 크고 정확도가 낮음을 확인할 수 있다.

[0098] 도 7에서 도시된 것처럼 mean 보정은 baseline보다 정확도가 높고, mean&std 보정은 mean 보정보다도 정확도가 더 높아 순차적으로 정확도 손실이 최소화되는 것을 확인할 수 있다.

[0099] 도 8은 본 발명의 다른 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 장치(220)의 가중치 양자화 이후 보정을 통한 정확도 손실 최소화 성능을 나타낸 그래프이다.

[0100] 도 8에 도시된 일 실시예는 ResNet-18 네트워크를 이용하여 양자화 및 오류 보정 과정을 수행하였다.

[0101] 도 8의 그래프의 x축은 양자화 레벨, y축은 정확도를 나타내며, 왼쪽 그래프는 CIFAR-100 데이터셋을, 오른쪽 그래프는 CIFAR-10 데이터셋을 이용한 실험 결과가 도시된다.

[0102] 도 8을 참조하면, ResNet-18 네트워크를 이미지 분류 작업에 대하여 학습시킨 후 가중치 양자화를 적용했을 때 정확도 손실을 측정한 결과가 도시된다. 다양한 가중치 양자화 레벨에 대하여 정확도 트렌드가 도시된다. 양자화 레벨이 낮을수록 양자화로 인한 손실이 크고 정확도가 낮음을 확인할 수 있다.

[0103] 도 8에서 도시된 것처럼 mean 보정은 baseline보다 정확도가 높고, mean&std 보정은 mean 보정보다도 정확도가 더 높아 순차적으로 정확도 손실이 최소화되는 것을 확인할 수 있다.

[0104] 도 9는 도 1 내지 도 8의 과정의 적어도 일부를 수행할 수 있는 일반화된 뉴럴 네트워크 양자화 오류 보정 장치(220), 또는 장치(220)를 구성하는 컴퓨팅 시스템의 예시를 도시하는 개념도이다.

[0105] 도 1 내지 도 8의 실시예에서도 도면 상으로는 생략되었으나 프로세서, 및 메모리가 전자적으로 각 구성 요소와 연결되고, 프로세서에 의하여 각 구성 요소의 동작이 제어되거나 관리될 수 있다.

[0106] 본 발명의 일 실시예에 따른 방법의 적어도 일부의 과정은 도 9의 컴퓨팅 시스템(1000)에 의하여 실행될 수 있다.

[0107] 도 9를 참조하면, 본 발명의 일 실시예에 따른 컴퓨팅 시스템(1000)은, 프로세서(1100), 메모리(1200), 통신 인

터페이스(1300), 저장 장치(1400), 입력 인터페이스(1500), 출력 인터페이스(1600) 및 버스(bus)(1700)를 포함하여 구성될 수 있다.

[0108] 본 발명의 일 실시예에 따른 컴퓨팅 시스템(1000)은, 적어도 하나의 프로세서(processor)(1100) 및 상기 적어도 하나의 프로세서(1100)가 적어도 하나의 단계를 수행하도록 지시하는 명령어들(instructions)을 저장하는 메모리(memory)(1200)를 포함할 수 있다. 본 발명의 일 실시예에 따른 방법의 적어도 일부의 단계는 상기 적어도 하나의 프로세서(1100)가 상기 메모리(1200)로부터 명령어들을 로드하여 실행함으로써 수행될 수 있다.

[0109] 프로세서(1100)는 중앙 처리 장치(central processing unit, CPU), 그래픽 처리 장치(graphics processing unit, GPU), 또는 본 발명의 실시예들에 따른 방법들이 수행되는 전용의 프로세서를 의미할 수 있다.

[0110] 메모리(1200) 및 저장 장치(1400) 각각은 휘발성 저장 매체 및 비휘발성 저장 매체 중에서 적어도 하나로 구성될 수 있다. 예를 들어, 메모리(1200)는 읽기 전용 메모리(read only memory, ROM) 및 랜덤 액세스 메모리(random access memory, RAM) 중에서 적어도 하나로 구성될 수 있다.

[0111] 또한, 컴퓨팅 시스템(1000)은, 무선 네트워크를 통해 통신을 수행하는 통신 인터페이스(1300)를 포함할 수 있다.

[0112] 또한, 컴퓨팅 시스템(1000)은, 저장 장치(1400), 입력 인터페이스(1500), 출력 인터페이스(1600) 등을 더 포함할 수 있다.

[0113] 또한, 컴퓨팅 시스템(1000)에 포함된 각각의 구성 요소들은 버스(bus)(1700)에 의해 연결되어 서로 통신을 수행할 수 있다.

[0114] 본 발명의 컴퓨팅 시스템(1000)의 예를 들면, 통신 가능한 데스크탑 컴퓨터(desktop computer), 랩탑 컴퓨터(laptop computer), 노트북(notebook), 스마트폰(smart phone), 태블릿 PC(tablet PC), 모바일폰(mobile phone), 스마트 워치(smart watch), 스마트 글래스(smart glass), e-book 리더기, PMP(portable multimedia player), 휴대용 게임기, 네비게이션(navigation) 장치, 디지털 카메라(digital camera), DMB(digital multimedia broadcasting) 재생기, 디지털 음성 녹음기(digital audio recorder), 디지털 음성 재생기(digital audio player), 디지털 동영상 녹화기(digital video recorder), 디지털 동영상 재생기(digital video player), PDA(Personal Digital Assistant) 등일 수 있다.

[0115] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크(100, 120)의 동작 방법은 메모리(memory)(1200)에 저장되는 적어도 하나의 명령을 실행하는 프로세서(processor)(1100)에 의하여 수행되는 방법으로서, 프로세서(1100)가 적어도 하나의 명령을 실행함으로써, 뉴럴 네트워크(100)의 양자화되기 전 제1 파라미터를 수신하는 단계(S330); 뉴럴 네트워크(100)의 양자화된 이후의 제2 파라미터를 수신하는 단계(S340); 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정함으로써 제3 파라미터를 생성하는 단계(220); 및 제3 파라미터에 기반하여 입력 데이터에 대한 추론 결과를 생성하는 단계(S360)를 포함한다.

[0116] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크(100, 120)의 동작 방법은 제3 파라미터에 기반하여 새로운 뉴럴 네트워크(120)를 생성하는 단계를 더 포함할 수 있다.

[0117] 추론 결과를 생성하는 단계(S360)는 새로운 뉴럴 네트워크(120)에 입력 데이터를 입력하는 단계(S350); 및 새로운 뉴럴 네트워크(120)의 출력을 추론 결과로서 생성하는 단계(S360)를 포함할 수 있다.

[0118] 본 발명의 일 실시예에 따른 양자화 오류가 보정된 뉴럴 네트워크(100)의 동작 방법은 제3 파라미터를 뉴럴 네트워크(100)에 전달하는 단계(S340); 및 제3 파라미터에 기반하여 뉴럴 네트워크(100)의 모든 파라미터를 갱신하는 단계를 더 포함할 수 있다.

[0119] 추론 결과를 생성하는 단계(S360)는 모든 파라미터가 갱신된 뉴럴 네트워크(100)에 입력 데이터를 입력하는 단계(S350); 및 뉴럴 네트워크(100)의 출력을 추론 결과로서 생성하는 단계(S360)를 포함할 수 있다.

[0120] 제3 파라미터를 생성하는 단계(220, 250)는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하도록 제2 파라미터 각각을 보정하는 단계(230)를 포함할 수 있다.

[0121] 제2 파라미터 각각을 보정하는 단계(220)는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다(230).

[0122] 제3 파라미터를 생성하는 단계(220, 250)는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 차이를 최소화하고, 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 차이를 최소화하도록 제2 파라미터 각각을 보정하는 단계(240)를 포함할 수 있다.

[0123] 제2 파라미터 각각을 보정하는 단계(220)는, 제1 파라미터의 평균과 제2 파라미터의 평균 간의 제1 차이; 및 제1 파라미터의 표준편차와 제2 파라미터의 표준편차 간의 제2 차이에 기반하여 제2 파라미터 각각을 보정할 수 있다(240).

[0124] 본 발명의 일 실시예에 따른 뉴럴 네트워크 양자화 오류 보정 방법은 메모리(memory)(1200)에 저장되는 적어도 하나의 명령을 실행하는 프로세서(processor)(1100)에 의하여 수행되는 방법으로서, 프로세서(1100)가 적어도 하나의 명령을 실행함으로써, 뉴럴 네트워크(1000의 양자화되기 전 제1 파라미터를 수신하는 단계(S330); 뉴럴 네트워크(100)의 양자화된 이후의 제2 파라미터를 수신하는 단계(S212); 제1 파라미터의 통계적 정보와 제2 파라미터의 통계적 정보에 기반하여 제2 파라미터를 보정하는 단계(220); 및 보정된 제2 파라미터를 제3 파라미터로서 출력하는 단계(250)를 포함한다.

[0125] 본 발명의 실시예에 따른 방법의 동작은 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 프로그램 또는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의해 읽힐 수 있는 정보가 저장되는 모든 종류의 기록장치를 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산 방식으로 컴퓨터로 읽을 수 있는 프로그램 또는 코드가 저장되고 실행될 수 있다.

[0126] 또한, 컴퓨터가 읽을 수 있는 기록매체는 롬(rom), 램(ram), 플래시 메모리(flash memory) 등과 같이 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치를 포함할 수 있다. 프로그램 명령은 컴파일러(compiler)에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터(interpreter) 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

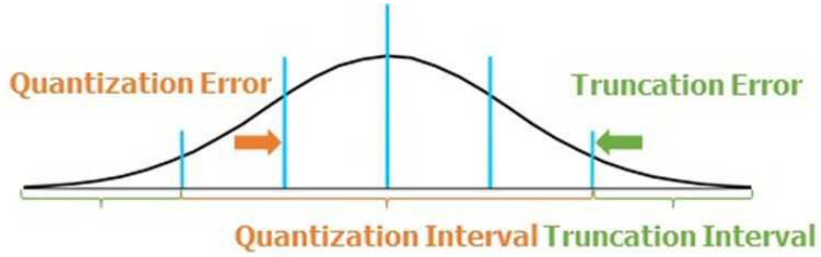
[0127] 본 발명의 일부 측면들은 장치의 문맥에서 설명되었으나, 그것은 상응하는 방법에 따른 설명 또한 나타낼 수 있고, 여기서 블록 또는 장치는 방법 단계 또는 방법 단계의 특징에 상응한다. 유사하게, 방법의 문맥에서 설명된 측면들은 또한 상응하는 블록 또는 아이템 또는 상응하는 장치의 특징으로 나타낼 수 있다. 방법 단계들의 몇몇 또는 전부는 예를 들어, 마이크로프로세서, 프로그램 가능한 컴퓨터 또는 전자 회로와 같은 하드웨어 장치에 의해(또는 이용하여) 수행될 수 있다. 몇몇의 실시 예에서, 가장 중요한 방법 단계들의 적어도 하나 이상은 이와 같은 장치에 의해 수행될 수 있다.

[0128] 실시예들에서, 프로그램 가능한 로직 장치(예를 들어, 필드 프로그래머블 게이트 어레이)가 여기서 설명된 방법들의 기능의 일부 또는 전부를 수행하기 위해 사용될 수 있다. 실시예들에서, 필드 프로그래머블 게이트 어레이(field-programmable gate array)는 여기서 설명된 방법들 중 하나를 수행하기 위한 마이크로프로세서(microprocessor)와 함께 작동할 수 있다. 일반적으로, 방법들은 어떤 하드웨어 장치에 의해 수행되는 것이 바람직하다.

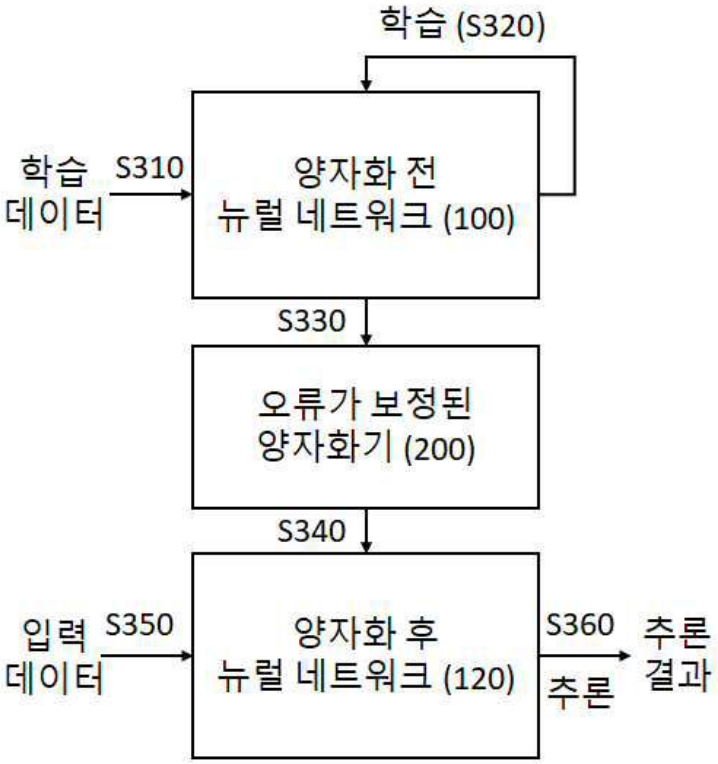
[0129] 이상 본 발명의 바람직한 실시 예를 참조하여 설명하였지만, 해당 기술 분야의 숙련된 당업자는 하기의 특허 청구의 범위에 기재된 본 발명의 사상 및 영역으로부터 벗어나지 않는 범위 내에서 본 발명을 다양하게 수정 및 변경시킬 수 있음을 이해할 수 있을 것이다.

도면

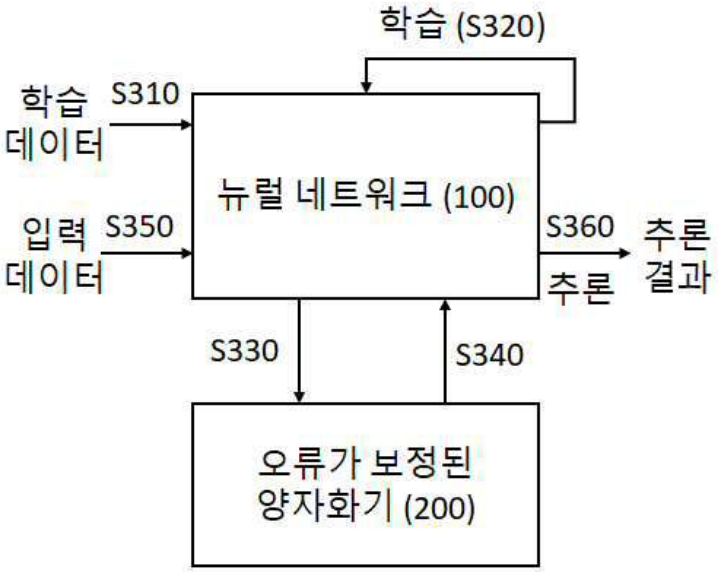
도면1



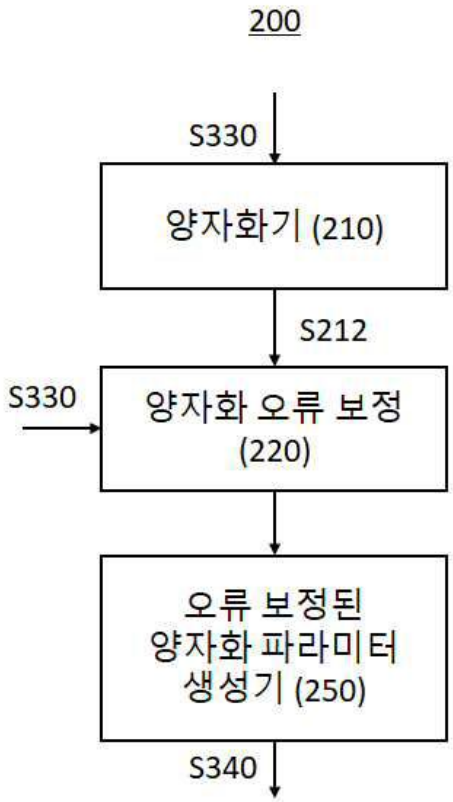
도면2


학습 (S320)
학습 데이터
S310
양자화 전 뉴럴 네트워크 (100)
S330
오류가 보정된 양자화기 (200)
S340
입력 데이터
S350
양자화 후 뉴럴 네트워크 (120)
S360
추론
추론 결과

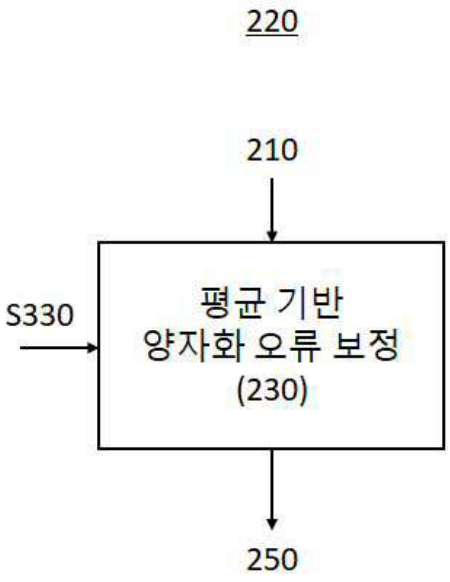
도면3


학습 (S320)
학습 데이터
S310
뉴럴 네트워크 (100)
입력 데이터
S350
S360
추론
추론 결과
S330
S340
오류가 보정된 양자화기 (200)

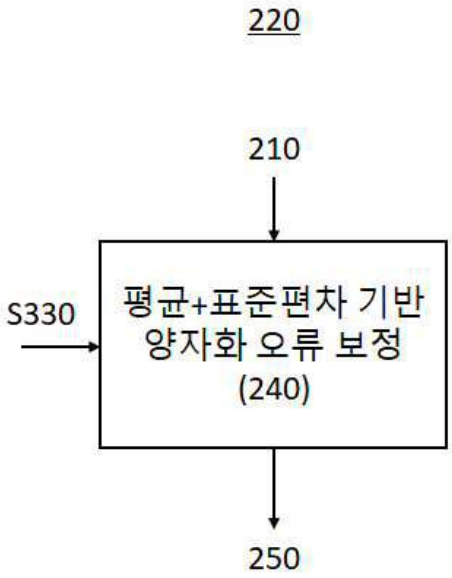
도면4


200
S330
양자화기 (210)
S212
S330
양자화 오류 보정
(220)
오류 보정된
양자화 파라미터
생성기 (250)
S340

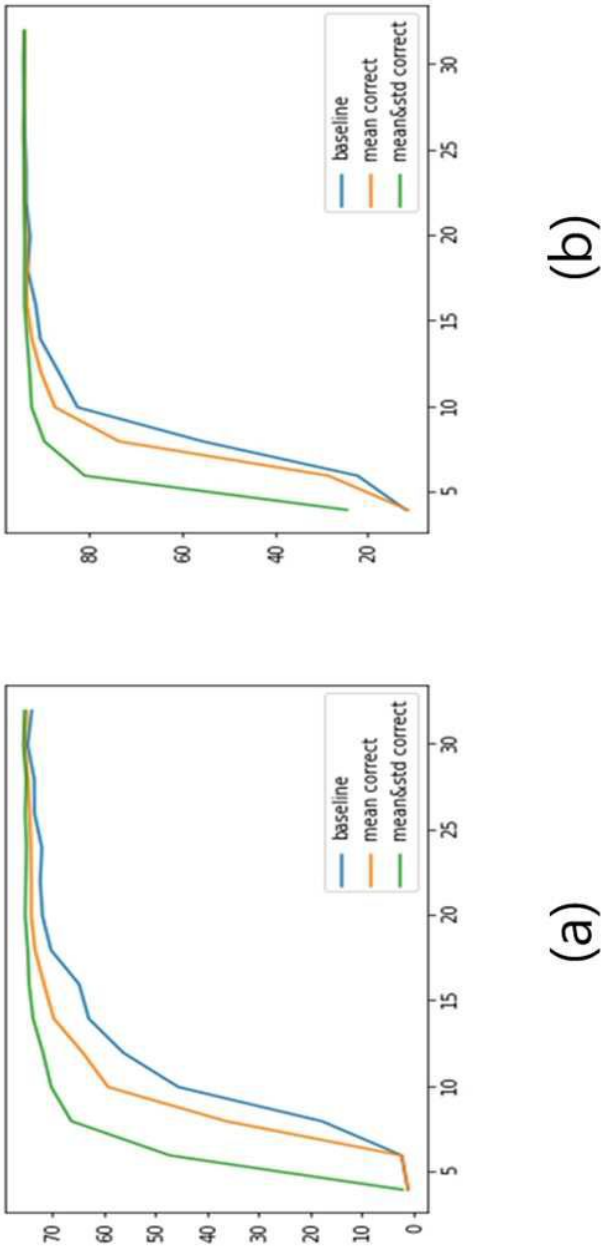
도면5


220
210
S330
평균 기반
양자화 오류 보정
(230)
250

도면6



도면7

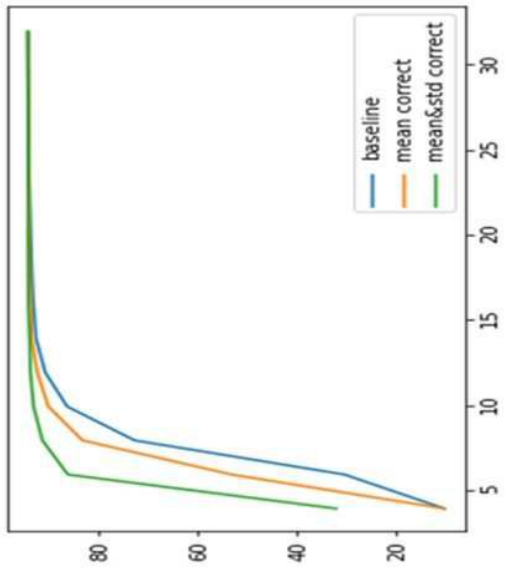

baseline
mean correct
mean&std correct
0
10
20
30
40
50
60
70
5
10
15
20
25
30

(a)

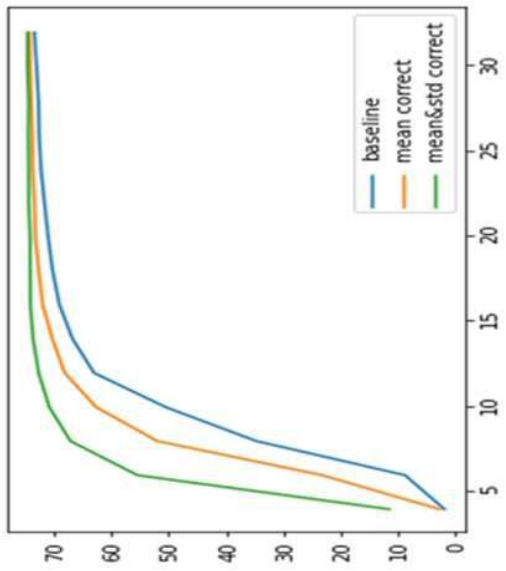
baseline
mean correct
mean&std correct
20
40
60
80
5
10
15
20
25
30

(b)

도면8

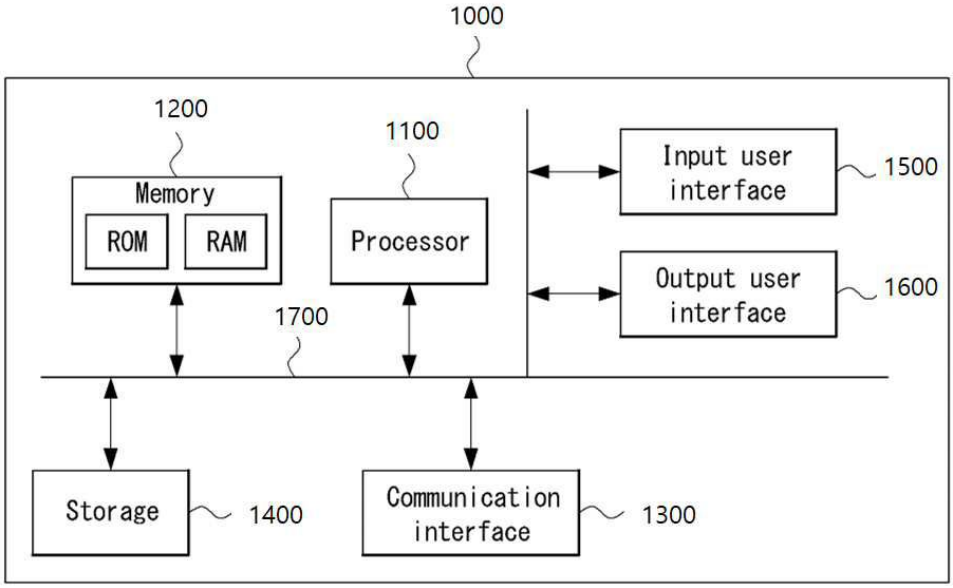

baseline
mean correct
mean&std correct
80
60
40
20
5
10
15
20
25
30

(b)


baseline
mean correct
mean&std correct
70
60
50
40
30
20
10
0
5
10
15
20
25
30

(a)

도면9


1000
1200
Memory
ROM
RAM
1100
Processor
Input user interface
1500
Output user interface
1600
1700
Storage
1400
Communication interface
1300