(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0134877
(43) 공개일자 2023년09월22일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 5/04* (2023.01)
(52) CPC특허분류
G06N 3/084 (2023.01)
G06N 3/086 (2023.01)
(21) 출원번호 10-2022-0032221
(22) 출원일자 2022년03월15일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
이호르바실조프
경기도 수원시 영통구 삼성로 129 (매탄동)
(74) 대리인
특허법인 무한

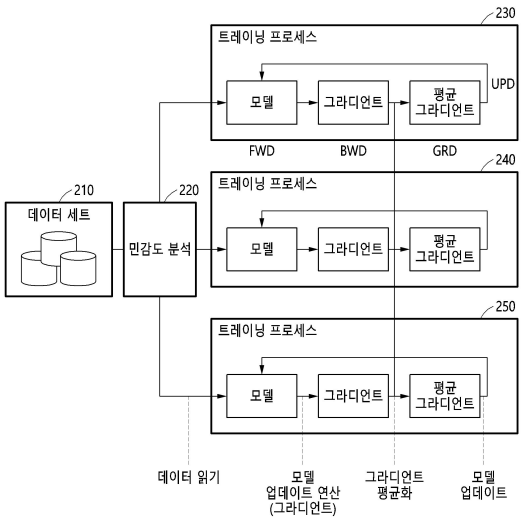
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **민감도-기반 양자화된 트레이닝을 수행하는 전자 장치 및 그 동작 방법**

(57) 요 약

민감도-기반 양자화된 트레이닝을 수행하는 전자 장치 및 그 동작 방법이 개시된다. 전자 장치는 프로세서 및 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리를 포함하고, 적어도 하나의 명령어가 프로세서에서 실행되면, 프로세서는 트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정하고, 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 모델을 트레이닝한다.

대 표 도 - 도2



(52) CPC특허분류
G06N 5/04 (2023.01)

명 세 서

청구범위

청구항 1

프로세서; 및

상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리

를 포함하고,

상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는

트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정하고,

미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝하는,

전자 장치.

청구항 2

제1항에 있어서,

상기 프로세서는

상기 임계치보다 낮은 민감도를 가진 레이어를 양자화하여 제1 정밀도(first precision)로 처리하고,

상기 임계치보다 높거나 같은 민감도를 가진 레이어를 양자화 없이 상기 제1 정밀도보다 높은 제2 정밀도로 처리하는,

전자 장치.

청구항 3

제1항에 있어서,

상기 프로세서는

상기 모델의 첫 번째 레이어로부터 마지막 레이어로의 순방향 전파 단계(forward propagation step);

상기 모델의 상기 마지막 레이어로부터 상기 첫 번째 레이어로의 역방향 전파 단계(backward propagation step);

상기 모델의 분산 트레이닝에 이용되는 복수의 노드들 각각에서 계산된 그라디언트들(gradients)의 평균 값을 결정하는 단계; 및

상기 평균 값에 따라 상기 모델의 가중치를 업데이트하는 단계

를 포함하는 분산 트레이닝(distributed training)을 상기 모델에 대해 수행하는,

전자 장치.

청구항 4

제1항에 있어서,

상기 프로세서는

상기 모델의 트레이닝마다 또는 상기 모델의 트레이닝에서 수행되는 에포크(epoch) 또는 하나 이상의 이터레이션(iteration)마다 주기적으로 상기 레이어들의 민감도를 결정하는,

전자 장치.

청구항 5

제1항에 있어서,

상기 프로세서는

상기 모델에 이용되는 텐서(tensor)의 채널별 민감도(channel-wise sensitivity)를 결정하고,

미리 정해진 제2 임계치보다 낮은 민감도를 가진 채널에 양자화를 적용하여 제1 정밀도로 처리하고,

상기 제2 임계치보다 높거나 같은 민감도를 가진 채널을 양자화 없이 상기 제1 정밀도보다 높은 제2 정밀도로 처리하는,

전자 장치.

청구항 6

제1항에 있어서,

상기 프로세서는

상기 모델에 포함된 레이어들의 민감도를 복수의 레벨들로 결정하고,

상기 레이어들을 대응하는 레벨의 정밀도로 양자화를 적용하여 상기 모델을 트레이닝하는,

전자 장치.

청구항 7

제3항에 있어서,

상기 프로세서는

상기 분산 트레이닝에 포함된 복수의 단계들 중 하나 이상의 단계에서 상기 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝하는,

전자 장치.

청구항 8

제7항에 있어서,

상기 프로세서는

상기 복수의 단계들 중 하나 이상의 단계에서 이용되는 데이터를 압축하는,

전자 장치.

청구항 9

제1항에 있어서,

상기 프로세서는

상기 모델을 트레이닝할 때 계산되는 그라디언트를 스케일링하여 상기 모델을 트레이닝하는,

전자 장치.

청구항 10

제3항에 있어서,

상기 프로세서는

상기 복수의 노드들 각각에서 계산된 그라디언트들 중 가장 큰 k개를 이용하여 상기 평균 값을 결정하거나, 상기 그라디언트들에 유진 알고리즘(Genetic Algorithms)을 적용함으로써 상기 평균 값을 결정하는,

전자 장치.

청구항 11

제1항에 있어서,

상기 트레이닝 대상인 모델은 양자화가 적용되지 않은 정밀도로 사전 트레이닝된 모델인,

전자 장치.

청구항 12

전자 장치의 동작 방법에 있어서,

트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정하는 동작; 및

미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝하는 동작

을 포함하는

전자 장치의 동작 방법.

청구항 13

제12항에 있어서,

상기 모델을 트레이닝하는 동작은

상기 임계치보다 낮은 민감도를 가진 레이어를 양자화하여 제1 정밀도(first precision)로 처리하고,

상기 임계치보다 높거나 같은 민감도를 가진 레이어를 양자화 없이 상기 제1 정밀도보다 높은 제2 정밀도로 처리하는,

전자 장치의 동작 방법.

청구항 14

제12항에 있어서,

상기 모델을 트레이닝하는 동작은

상기 모델의 첫 번째 레이어로부터 마지막 레이어로의 순방향 전파 단계;

상기 모델의 상기 마지막 레이어로부터 상기 첫 번째 레이어로의 역방향 전파 단계;

상기 모델의 분산 트레이닝에 이용되는 복수의 노드들 각각에서 계산된 그라디언트들의 평균 값을 결정하는 단계; 및

상기 평균 값에 따라 상기 모델의 가중치를 업데이트하는 단계

를 포함하는 분산 트레이닝을 상기 모델에 대해 수행하는,

전자 장치의 동작 방법.

청구항 15

제12항에 있어서,

상기 민감도를 결정하는 동작은

상기 모델의 트레이닝마다 또는 상기 모델의 트레이닝에서 수행되는 에포크 또는 하나 이상의 이터레이션마다 주기적으로 상기 레이어들의 민감도를 결정하는,

전자 장치의 동작 방법.

청구항 16

제12항에 있어서,

상기 민감도를 결정하는 동작은

상기 모델에 이용되는 텐서(tensor)의 채널별 민감도(channel-wise sensitivity)를 결정하고,

상기 모델을 트레이닝하는 동작은

미리 정해진 제2 임계치보다 낮은 민감도를 가진 채널에 양자화를 적용하여 상기 모델을 트레이닝하는,

전자 장치의 동작 방법.

청구항 17

제12항에 있어서,

상기 민감도를 결정하는 동작은

상기 모델에 포함된 레이어들의 민감도를 복수의 레벨들로 결정하고,

상기 모델을 트레이닝하는 동작은

상기 레이어들을 대응하는 레벨의 정밀도로 양자화를 적용하여 상기 모델을 트레이닝하는,

전자 장치의 동작 방법.

청구항 18

제14항에 있어서,

상기 모델을 트레이닝하는 동작은

상기 분산 트레이닝에 포함된 복수의 단계들 중 하나 이상의 단계에서 상기 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝하는,

전자 장치의 동작 방법.

청구항 19

제12항에 있어서,

상기 트레이닝 대상인 모델은 양자화가 적용되지 않은 정밀도로 사전 트레이닝된 모델인,

전자 장치의 동작 방법.

청구항 20

제12항 내지 제19항 중에서 어느 한 항의 방법을 실행하는 컴퓨터 프로그램을 저장하는 컴퓨터 판독가능 기록매체.

발명의 설명

기 술 분 야

[0001] 아래의 개시는 민감도-기반 양자화된 트레이닝을 수행하는 전자 장치 및 그 동작 방법에 관한 것이다.

배 경 기 술

[0003] 최첨단 DNN 모델은 너무 커서 모바일 장치 등 제한된 사용 환경에서 실행하기에 비효율적일 수 있다. DNN 모델의 양자화는 특정 하드웨어에서 실행되도록 특정 모델을 최적화하는 접근 방식 중 하나이며, 특히 혼합-정밀도 양자화(mixed precision quantization)는 DNN 최적화의 매우 유망한 접근 방식일 수 있다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0005] 일 실시예에 따른 전자 장치는 프로세서 및 상기 프로세서에 의해 실행 가능한 적어도 하나의 명령어를 포함하는 메모리를 포함하고, 상기 적어도 하나의 명령어가 상기 프로세서에서 실행되면, 상기 프로세서는 트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정하고, 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝한다.

[0006] 상기 프로세서는 상기 임계치보다 낮은 민감도를 가진 레이어를 양자화하여 제1 정밀도(first precision)로 처리하고, 상기 임계치보다 높거나 같은 민감도를 가진 레이어를 양자화 없이 상기 제1 정밀도보다 높은 제2 정밀도로 처리할 수 있다.

[0007] 상기 프로세서는 상기 모델의 첫 번째 레이어로부터 마지막 레이어로의 순방향 전파 단계(forward propagation step), 상기 모델의 상기 마지막 레이어로부터 상기 첫 번째 레이어로의 역방향 전파 단계(backward propagation step), 상기 모델의 분산 트레이닝에 이용되는 복수의 노드들 각각에서 계산된 그라디언트들(gradients)의 평균 값을 결정하는 단계 및 상기 평균 값에 따라 상기 모델의 가중치를 업데이트하는 단계를 포함하는 분산 트레이닝(distributed training)을 상기 모델에 대해 수행할 수 있다.

[0008] 상기 프로세서는 상기 모델의 트레이닝마다 또는 상기 모델의 트레이닝에서 수행되는 에포크(epoch) 또는 하나 이상의 이터레이션(iteration)마다 주기적으로 상기 레이어들의 민감도를 결정할 수 있다.

[0009] 상기 프로세서는 상기 모델에 이용되는 텐서(tensor)의 채널별 민감도(channel-wise sensitivity)를 결정하고,

미리 정해진 제2 임계치보다 낮은 민감도를 가진 채널에 양자화를 적용하여 제1 정밀도로 처리하고, 상기 제2 임계치보다 높거나 같은 민감도를 가진 채널을 양자화 없이 상기 제1 정밀도보다 높은 제2 정밀도로 처리할 수 있다.

[0010] 상기 프로세서는 상기 모델에 포함된 레이어들의 민감도를 복수의 레벨들로 결정하고, 상기 레이어들을 대응하는 레벨의 정밀도로 양자화를 적용하여 상기 모델을 트레이닝할 수 있다.

[0011] 상기 프로세서는 상기 분산 트레이닝에 포함된 복수의 단계들 중 하나 이상의 단계에서 상기 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝할 수 있다.

[0012] 상기 프로세서는 상기 복수의 단계들 중 하나 이상의 단계에서 이용되는 데이터를 압축할 수 있다.

[0013] 상기 프로세서는 상기 모델을 트레이닝할 때 계산되는 그라디언트를 스케일링하여 상기 모델을 트레이닝할 수 있다.

[0014] 상기 프로세서는 상기 복수의 노드들 각각에서 계산된 그라디언트들 중 가장 큰 k개를 이용하여 상기 평균 값을 결정하거나, 상기 그라디언트들에 유진 알고리즘(Genetic Algorithms)을 적용함으로써 상기 평균 값을 결정할 수 있다.

[0015] 상기 트레이닝 대상인 모델은 양자화가 적용되지 않은 정밀도로 사전 트레이닝된 모델일 수 있다.

[0016] 전자 장치의 동작 방법은 트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정하는 동작 및 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 상기 모델을 트레이닝하는 동작을 포함한다.

도면의 간단한 설명

[0018] 도 1은 일 실시예에 따른 뉴럴 네트워크를 설명하기 위한 도면이다.

도 2는 일 실시예에 따른 분산 병렬 모델 트레이닝(distributed parallel model training)을 설명하기 위한 도면이다.

도 3은 일 실시예에 따라 레이어 민감도 분석에 따른 모델 트레이닝을 설명하기 위한 도면이다.

도 4는 일 실시예에 따라 채널별 민감도 분석에 따른 모델 트레이닝을 설명하기 위한 도면이다.

도 5는 일 실시예에 따라 복수의 레이어 민감도 리스트들에 따른 모델 트레이닝을 설명하기 위한 도면이다.

도 6은 일 실시예에 따라 부분 양자화 및 추가 압축에 따른 모델 트레이닝을 설명하기 위한 도면이다.

도 7은 일 실시예에 따라 그라디언트 로스에 따른 모델 트레이닝을 설명하기 위한 도면이다.

도 8은 일 실시예에 따라 그라디언트들의 평균 값을 계산하는 동작을 설명하기 위한 도면이다.

도 9는 일 실시예에 따른 전자 장치의 동작 방법을 나타낸 도면이다.

도 10은 일 실시예에 따른 전자 장치를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0019] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0020] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0021] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0022] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다"

또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0023] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0024] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.

[0026] 도 1은 일 실시예에 따른 뉴럴 네트워크를 설명하기 위한 도면이다.

[0027] 도 1을 참조하면, 뉴럴 네트워크(100)는 복수의 레이어들을 포함한다. 뉴럴 네트워크(100)는 입력 레이어(110), 복수의 히든 레이어들(120, 130) 및 출력 레이어(140)를 포함할 수 있다. 뉴럴 네트워크(100)를 통해 데이터 추론(data inference)이 수행될 수 있다. 데이터 추론은 예를 들어, 패턴 인식(예: 객체 인식, 얼굴 식별 등), 시퀀스 인식(예: 음성, 제스처, 필기 텍스트 인식, 기계 번역, 기계 통역 등), 제어(예: 차량 제어, 프로세스 제어 등), 추천 서비스, 의사 결정, 의료 진단, 금융 어플리케이션, 데이터 마이닝을 포함할 수 있으나, 데이터 추론의 예시가 이에 한정되는 것은 아니다. 본 명세서에서 뉴럴 네트워크(100)는 설명의 편의를 위해 모델로도 지칭될 수 있다.

[0028] 각각의 레이어들은 인공 뉴런이라고도 불리는 복수의 노드들을 포함할 수 있다. 각 노드는 하나 이상의 입력 및 출력을 가지는 계산 단위를 나타내고, 노드들은 상호 연결될 수 있다.

[0029] 입력 레이어(110)는 다른 노드와의 연결을 거치지 않고, 데이터가 직접 입력되는 하나 이상의 노드들을 포함할 수 있다. 출력 레이어(140)는 다른 노드와의 연결에서 출력 노드를 가지지 않는 하나 이상의 노드들을 포함할 수 있다. 히든 레이어들(120, 130)은 입력 레이어(110) 및 출력 레이어(140)를 제외한 뉴럴 네트워크(100)의 나머지 레이어들에 해당하며, 다른 노드와의 관계에서 입력 노드 또는 출력 노드에 해당하는 노드들을 포함할 수 있다. 도 1의 뉴럴 네트워크(100)는 설명의 편의를 위해 예시적으로 도시된 것으로, 해당 뉴럴 네트워크(100)의 구조에 의해 실시예의 범위가 한정되는 것으로 해석되어서는 안 된다. 실시예에서 이용되는 뉴럴 네트워크(100)의 구조는 다양할 수 있다. 실시예에 따라 뉴럴 네트워크(100)에 포함된 히든 레이어의 수, 각 레이어에 포함된 노드의 수 및/또는 노드들 간의 연결 관계는 상이할 수 있다. 여러 히든 레이어들이 포함된 뉴럴 네트워크(100)를 DNN(deep neural network)라고 지칭할 수 있다.

[0030] 노드들 간의 연결에는 가중치가 설정될 수 있다. 예를 들어, 입력 레이어(110)에 포함된 한 노드와 히든 레이어(120)에 포함된 다른 노드 간 연결에는 특정한 가중치가 설정될 수 있다. 이러한 가중치는 조정 또는 변경될 수 있다. 가중치는 연관된 데이터 값을 증폭, 감소 또는 유지시킴으로써 해당 데이터 값이 최종 결과에 미치는 영향도를 결정할 수 있다. 가중치는 뉴럴 네트워크(100)의 파라미터에 해당할 수 있다.

[0031] 한 레이어에 포함된 각각의 노드에는 이전 레이어에 포함된 노드들의 가중된 값들이 입력될 수 있다. 가중된 값은 이전 레이어에 포함된 노드의 값(예: 활성(activation))에 가중치가 곱해진 것일 수 있다. 가중된 데이터가 임의의 레이어로부터 다음 레이어로 입력되는 과정을 전파(propagation)라고 지칭할 수 있다.

[0032] 일반적으로, 가중치와 활성은 32비트로 데이터를 나타내는 FP32(32-bit floating point) 정밀도이나 16비트로 데이터를 나타내는 BFLOAT16(16-bit brain floating point) 정밀도로 표현될 수 있다. 해당 정밀도들을 통해 추론의 정확도를 향상시킬 수 있더라도, 뉴럴 네트워크(100)를 학습시키거나 뉴럴 네트워크(100)를 이용해 추론을 수행하는 데 많은 시간과 자원(예: 소비전력, 메모리 등)이 필요하며, 자원이 제한된 사용 환경(예: 모바일 디바이스 등)에서는 동작이 어려울 수 있다.

[0033] 모델 양자화를 통해 가중치와 활성이 상대적으로 적은 비트들로 표현됨으로써, 뉴럴 네트워크(100)의 추론이 압축되고, 가속화될 수 있다. 저-정밀도 가속기(low-precision accelerator)(예: INT2, INT4, INT8 정밀도의 가속기)를 이용하여 뉴럴 네트워크(100)를 실행할 수 있으므로, 추론하는 동안 대기 시간과 전력 소비를 효과적으

로 감소시킬 수 있다. 다만, 뉴럴 네트워크(100)에 포함된 모든 레이어들에 대해 동일한 정밀도(다시 말해, 비트 수)를 적용시킨다면, 양자화로 인한 추론의 정확도 저하가 발생할 수 있다.

[0034] 혼합-정밀도 양자화에서 뉴럴 네트워크(100)에 포함된 복수의 레이어들이 상이한 정밀도를 가질 수 있다. 혼합-정밀도 양자화를 통해 복수의 레이어들 중 민감한 레이어에 높은 정밀도를 적용시키고, 강인한 레이어(robust layer)에 낮은 정밀도를 적용시킴으로써 양자화에 따른 성능 저하를 최소화할 수 있더라도, 최적의 혼합-정밀도 양자화를 검색하는 복잡성이 커질 수 있다. 예를 들어, 뉴럴 네트워크(100)가 50개의 레이어들을 포함하고, 해당 뉴럴 네트워크(100)가 3개의 정밀도(예: INT4, INT8, INT16)를 이용할 수 있다면, 전체 검색 공간은 3^{50}에 해당하는 상당한 크기를 가질 수 있다. 또한, 낮은 정밀도를 통해 모델 트레이닝의 계산 복잡성을 줄일 수 있더라도, 상당한 정확도 저하가 발생하여 모델 수렴이 느려질 수 있으며, 모델의 정확도를 위해서는 추가 이터레이션(additional iteration)이 필요할 수 있다.

[0035] 본 명세서에서 설명되는 일 실시예에 따르면, 모델에 포함된 복수의 레이어들에 민감도 분석을 적용하고, 민감도에 따라 적어도 일부 레이어에 양자화를 적용하여 모델을 트레이닝하는 동작이 수행될 수 있다. 이처럼, 민감도가 낮은 일부 레이어를 양자화하여 모델 트레이닝을 수행함으로써, 트레이닝 시간을 단축하고, 데이터 통신의 대역폭과 전력 소모를 효과적으로 감소시킬 수 있으며, 양자화된 추론을 위한 보다 효율적인 솔루션을 자동으로 생성할 수 있다. 이하, 실시예들을 보다 자세히 설명한다.

[0037] 도 2는 일 실시예에 따른 분산 병렬 모델 트레이닝(distributed parallel model training)을 설명하기 위한 도면이다.

[0038] 도 2를 참조하면, 모델 트레이닝이 복수의 노드들(230, 240, 250)에서 수행될 수 있다. 설명의 명확과 편의를 위해, 본 명세서에서 모델 트레이닝에는 데이터 병렬 시나리오(data parallelism scenario)가 적용될 수 있다.

[0039] 복수의 노드들(230, 240, 250)에서 분산 병렬 모델 트레이닝이 수행될 수 있다. 복수의 노드들(230, 240, 250) 각각은 데이터 세트(210)로부터 전달된 학습 데이터에 기반하여 모델 트레이닝을 수행할 수 있으며, 이때 도 2에 도시된 FWD(forward), BWD(backward), GRD(gradient), UPD(update) 단계들이 반복적으로 수행될 수 있다.

[0040] FWD 단계에서는, 활성이 모델의 첫 번째 레이어부터 마지막 레이어까지 순차적으로 계산되는 순방향 전파가 수행될 수 있다. BWD 단계에서는, 로스(loss)가 모델의 마지막 레이어로부터 첫 번째 레이어로의 역방향으로 전파됨으로써 그라디언트(gradient)가 계산될 수 있다. 여기서, 로스는 FWD 단계에서 모델의 출력 레이어에서 출력되는 추론 결과와 학습 데이터에 포함된 레이블(label) 간 차이를 나타낼 수 있다. GRD 단계에서는, 각 노드들(230, 240, 250)에서 계산된 그라디언트들의 평균 값이 결정될 수 있다. 각 노드들(230, 240, 250)에서 계산된 그라디언트들은 이용된 학습 데이터 차이 등에 의해 서로 상이할 수 있으며, GRD 단계에서 그라디언트들의 평균 값이 계산될 수 있다. 실시예에 따라서는, 전체 프로세스 속도를 높이기 위해 BWD 단계와 GRD 단계가 함께 수행될 수도 있으나, 본 명세서에서는 설명의 편의를 위해 이러한 경우를 고려하지 않는다. 다만, 실시예가 이에 제한되지 않으며, 본 명세서에서 설명한 사항들이 BWD 단계와 GRD 단계가 함께 수행되는 경우에도 제한 없이 적용될 수 있다. UPD 단계에서는, 그라디언트들의 평균 값에 따라 모델의 가중치가 업데이트될 수 있다. 업데이트된 가중치는 각 노드들(230, 240, 250)로 전달되어 다음 트레이닝(예: 이터레이션, 에포크 등)에 반영될 수 있다.

[0041] 앞서 설명한 FWD, BWD, GRD, UPD 단계들은 모델이 수렴할 때까지 반복될 수 있다. 만약 이러한 단계들에서 모델 계산 및 노드들 간 데이터 통신이 높은 정밀도(예: FP32, FP16(16-bit floating point) 등)를 사용한다면, 종단 간 트레이닝(end-to-end training)이 느려지고 더 많은 전력이 소모될 수 있다. 반면, 종단 간 트레이닝 속도를 높이고, 전력 소비를 줄이기 위해, 낮은 정밀도의 가중치, 활성 및 그라디언트를 일률적으로 사용하면, 상당한 정확도 저하가 발생하여 모델 수렴이 느려지고, 모델의 정확도를 일정 수준으로 높이기 위해서는 더 많은 이터레이션이 필요할 수 있다. 따라서, 낮은 정밀도를 일정한 조건을 만족하는 적어도 일부의 레이어에 선택적으로 적용함으로써, 종단 간 트레이닝 속도를 높이고, 전력 소비를 줄이면서도 정확도 저하를 최소화할 수 있다. 이를 위해 민감도 분석(220)이 수행될 수 있다.

[0042] 민감도 분석(220)은 모델에 포함된 레이어들에 대해 수행될 수 있다. 각 레이어들의 민감도가 미리 정해진 임계치 낮은지 여부가 판단되고, 임계치보다 낮은 민감도를 가진 레이어에 대해서는 가중치 및/또는 입력 데이터(예: 입력 텐서(input tensor)가 양자화되어 낮은 정밀도(예: INT2, INT4, INT8 등)로 처리될 수 있다. 예를

들어, 민감도 분석(220)과 양자화는 아래의 수학식 1로 표현될 수 있다.

수학식 1

[0043] $$\textbf{if } s_i < thr \textbf{ then: } \begin{matrix} \langle w_i \rangle = \text{quantize}(\langle w_i \rangle) \\ \langle x_i \rangle = \text{quantize}(\langle x_i \rangle) \end{matrix} \textbf{ with } \begin{matrix} quantize\ (*) = int8(*) \\ quantize\ (*) = int4(*) \end{matrix}$$

[0044] 위의 수학식 1에서, S_i는 i번째 레이어의 민감도를 나타내고, thr는 미리 정해진 임계치를 나타내고, w_i는 i번째 레이어의 가중치 벡터를 나타내며, x_i는 i번째 레이어로 입력되는 텐서를 나타낼 수 있다.

[0045] 수학식 1에서 $S_i < thr$로 표현된 민감도 분석(220)은 주기적으로 수행될 수 있다. 예를 들어, 민감도 분석(220)은 모델의 트레이닝마다 또는 모델 트레이닝에서 수행되는 에포크 또는 하나 이상의 이터레이션마다 주기적으로 수행될 수 있다.

[0046] 민감도 분석(220)을 기반으로 덜 민감한(또는, 더 강인한(more robust)) 레이어에 양자화를 선택적으로 적용함으로써, 양자화로 인한 모델의 정확도 저하를 줄여서 모델 정확도 확보를 위한 추가 이터레이션을 최소화하면서, 종단 간 트레이닝 속도를 높이고, 전력 소비를 효과적으로 감소시킬 수 있다. 또한, 트레이닝 시간이 감소되고, 레이어들 간 및/또는 노드들 간 데이터 통신의 대역폭이 감소되며, 양자화된 추론을 위한 보다 효율적인 솔루션이 자동으로 생성될 수 있다.

[0048] 도 3은 일 실시예에 따라 레이어 민감도 분석에 따른 모델 트레이닝을 설명하기 위한 도면이다.

[0049] 도 3을 참조하면, 혼합 정밀도를 갖는 민감도-기반 양자화 트레이닝 동작이 예시적으로 도시된다.

[0050] 동작(310)에서, 모델은 사전 트레이닝될 수 있다. 본 명세서에서 설명하는 분산 병렬 모델 트레이닝을 시작하기 전에 모델은 부분적으로 높은 정밀도(예: FP32, FP16 등)로 트레이닝될 수 있다.

[0051] 예를 들어, FP32 기반으로 모델이 짧은 문장(예: 길이가 128words와 동일하거나 작음)에 대해 트레이닝되고, 긴 문장(예: 길이가 128words보다 크고, 512words와 동일하거나 작음)에 대한 추가 트레이닝 전 상태인 MLPerf 시나리오에 해당하는 모델이, 동작(310)에서 사전 트레이닝된 모델일 수 있다.

[0052] 또한, 일부 워밍업(warming-up) 기간 동안 FP32 및/또는 FP16 정밀도로 모델을 부분적으로 사전 트레이닝된 모델이, 동작(310)에서 사전 트레이닝된 모델일 수 있다.

[0053] 동작(320)에서, 모델에 포함된 레이어들의 민감도가 분석될 수 있다. 동작(320)은 트레이닝마다 또는 모델의 트레이닝에서 수행되는 에포크 또는 하나 이상의 이터레이션마다 주기적으로 수행됨으로써, 레이어 민감도 분석에 따른 연산 오버헤드는 작아서 무시될 수 있다. 레이어 민감도 분석을 통해 레이어 민감도 리스트(330)가 생성될 수 있다. 레이어 민감도 리스트(330)는 모델에 포함된 복수의 레이어들의 민감도 정보를 포함할 수 있으며, 레이어 민감도 리스트(330)에 기반하여 FWD 단계(340), BWD 단계(350), GRD 단계(360) 및 UPD 단계(370)에서 미리 정해진 임계치보다 낮은 민감도를 가진 하나 이상의 레이어들이 양자화되어 낮은 정밀도로 처리될 수 있다.

[0054] 이처럼, 낮은 민감도를 가진 레이어를 선택적으로 양자화하여 낮은 정밀도로 처리하고, 높은 민감도를 가진 레이어는 양자화 없이 높은 정밀도로 처리해서 트레이닝을 수행함으로써, 모든 레이어들을 양자화하여 낮은 정밀도로 처리하는 경우보다는 성능이 우수하고, 때로는 더 나은 정규화로 인해 모든 레이어들을 항상 높은 정밀도로 처리하는 경우보다 성능이 우수할 수 있다. 정리하면, 선택적인 양자화를 적용한 혼합 정밀도 기반 트레이닝을 통해, 합리적인 모델 성능을 기대하면서도, 모델의 전체 트레이닝 시간을 단축하고, 모델 처리에서 수행되는 데이터 통신의 대역폭을 감소시키고, 하드웨어 기능을 활용하여 양자화된 데이터의 계산을 효율적으로 가속할 수 있으며, 양자화된 추론을 위한 보다 효율적인 솔루션을 자동으로 획득할 수 있다.

[0056] 도 4는 일 실시예에 따라 채널별 민감도 분석에 따른 모델 트레이닝을 설명하기 위한 도면이다.

[0057] 도 4를 참조하면, 모델에 이용되는 텐서의 채널별 민감도에 따라 선택적으로 양자화를 적용하여 모델을 트레이

닝하는 예시가 도시된다. 도 3에서는 민감도를 레이어별로 분석한 반면, 도 4에 도시된 예시에서는 민감도를 더 세부적인 단위인 채널별로 분석하고, 미리 정해진 제2 임계치보다 낮은 민감도를 가진 채널에 양자화를 적용하여 모델을 트레이닝할 수 있다. 제2 임계치보다 높거나 같은 민감도를 가진 채널은 양자화가 적용되지 않고, 높은 정밀도로 처리될 수 있다. 도 4에 도시된 동작(410)에서, 채널별 민감도(channel-wise sensitivity)이 분석될 수 있으며, 그 결과 채널별 민감도 리스트(420)가 생성될 수 있다. 채널별 민감도 리스트(420)는 채널별 민감도 정보를 포함할 수 있으며, 이러한 정보가 활용되어 FWD 단계 내지 UPD 단계에서 채널별 민감도에 따라 선택적으로 양자화가 적용될 수 있다. 나머지 동작에 대해서는 도 3를 통해 전술한 사항들이 마찬가지로 적용될 수 있으므로, 보다 상세한 설명은 생략한다.

[0058] 이처럼 채널별 민감도에 따라 선택적으로 양자화를 채널에 적용하여 모델을 트레이닝함으로써, 모델 근사의 특이성을 더욱 높일 수 있으며, 결과적으로 양자화로 인한 정확도 저하를 더욱 줄일 수 있다. 또한, 종단 간 분산 트레이닝의 속도를 효과적으로 향상시킬 수 있다.

[0060] 도 5는 일 실시예에 따라 복수의 레이어 민감도 리스트들에 따른 모델 트레이닝을 설명하기 위한 도면이다.

[0061] 도 5를 참조하면, 레이어 민감도 분석으로 생성된 복수의 레이어 민감도 리스트들(520)에 기초하여 모델을 트레이닝하는 예시가 도시된다. 도 3에서는 하나의 레이어 민감도 리스트(330)에 따라 양자화가 선택적으로 적용되는 예시가 도시되어 있으나, 전술한 예에 한정되지 않으며, 실시예에 따라서는 레이어 민감도 리스트가 복수일 수 있다.

[0062] 동작(510)에서 수행되는 레이어 민감도 분석으로, 모델에 포함된 레이어들의 민감도가 복수의 레벨들로 결정될 수 있으며, 그 결과 복수의 레이어 민감도 리스트들(520)이 생성될 수 있다. 예를 들어, 제1 레이어 민감도 리스트는 제1 임계치보다 낮은 민감도를 가지는 레이어에 대한 정보를 포함하고, 제2 레이어 민감도 리스트는 제2 임계치보다 낮은 민감도를 가지는 레이어에 대한 정보를 포함하며, 제3 레이어 민감도 리스트는 제3 임계치보다 낮은 민감도를 가지는 레이어에 대한 정보를 포함할 수 있다. 이때, 제1 임계치, 제2 임계치, 제3 임계치 순서로 그 크기가 커질 수 있으나, 전술한 예에 한정되지 않는다. 복수의 레이어 민감도 리스트들(520)을 통해 레이어의 민감도를 더 미세하게 세분화하여 분석하고, 모델 트레이닝에 활용할 수 있다. 이러한 복수의 레이어 민감도 리스트들(520)은 FWD 단계, BWD 단계, GRD 단계, UPD 단계 중 적어도 하나에 적용될 수 있으며, 각 단계에 다른 레이어 민감도 리스트가 적용될 수 있다.

[0063] 이를 통해, 모델 근사의 특이성을 더욱 향상시킬 수 있으며, 결과적으로 양자화로 인한 정확도 저하를 효과적으로 억제할 수 있다. 또한, 종단 간 분산 트레이닝의 속도도 더 향상시킬 수 있다.

[0065] 도 6은 일 실시예에 따라 부분 양자화 및 추가 압축에 따른 모델 트레이닝을 설명하기 위한 도면이다.

[0066] 도 6을 참조하면, 분산 병렬 모델 트레이닝에 포함된 네 단계들 중 일부에 양자화가 적용될 수 있다.

[0067] 동작(610)에서 레이어들의 민감도가 분석될 수 있다. 동작(620)에서, 레이어별 정밀도가 할당될 수 있다. 예를 들어, 높은 민감도를 가진 레이어에는 높은 정밀도(예: FP32, FP16 등)가 할당되고, 낮은 민감도를 가진 레이어에는 낮은 정밀도(예: INT2, INT4, INT8 등)이 할당될 수 있다. 이러한 과정을 통해 레이어 민감도 리스트(630)가 생성될 수 있다.

[0068] 도 6의 예시에서, FWD 단계(640)에는 양자화가 적용되어 낮은 민감도를 가진 레이어는 낮은 정밀도로 처리될 수 있다. BWD 단계는 실시예에 따라 양자화가 적용되는 경우(650) 및 양자화가 적용되지 않는 경우(660) 중 어느 하나의 경우로 수행될 수 있다. BWD 단계에 양자화가 적용되지 않는 경우(660)에는 GRD 단계에 압축(670)이 수행될 수 있다. 이때 압축 동작에는 양자화가 적용되어 낮은 민감도를 가진 레이어는 낮은 정밀도로 처리될 수 있다. 또한, 필요시 추가적인 GRD 압축(680)이 수행될 수 있으며, 이때에도 양자화가 적용되어 낮은 민감도를 가진 레이어는 낮은 정밀도로 처리될 수 있다. 이후 GRD 단계나 UPD 단계는 양자화 적용 없이 높은 정밀도로 처리될 수 있으며, 양자화가 적용된 압축 및 가중치 업데이트(690)가 추가적으로 수행될 수 있다. 다만, 도 6에 도시된 예시는 설명의 편의를 위한 것으로, 양자화가 적용되는 단계가 전술한 예에 한정되지 않는다.

[0069] 이처럼 분산 병렬 모델 트레이닝에 포함된 일부 단계들에 선택적으로 양자화를 적용함으로써, 모델 트레이닝 중 설계의 복잡성, 성능 속도 향상 및 정확도 저하 간 균형을 유연하게 제어할 수 있다.

[0071] 도 7은 일 실시예에 따라 그라디언트 로스에 따른 모델 트레이닝을 설명하기 위한 도면이다.

[0072] 도 7을 참조하면, 혼합 정밀도가 모델 트레이닝에 적용될 때 정확도 저하를 줄이는 알려진 방법인 그라디언트/로스 스케일링(gradient/loss scaling)이 본 명세서에서 설명하는 분산 병렬 모델 트레이닝에 적용될 수 있다. 도 7에 도시된 그라디언트/로스 스케일링 기법 중 그라디언트 스케일링(710)이 분산 병렬 모델 트레이닝에 적용될 수 있으며, 이때 양자화가 적용될 수 있다. 이를 통해, GRD 단계에서 다른 노드로 이동할 데이터 크기를 효과적으로 감소시킬 수 있고, 트레이닝 프로세스의 속도를 더욱 향상시킬 수 있다.

[0074] 도 8은 일 실시예에 따라 그라디언트들의 평균 값을 계산하는 동작을 설명하기 위한 도면이다.

[0075] 도 8을 참조하면, 다른 트레이닝 목표(예: 속도, 전력, 대역폭)가 동시에 고려된다면, 양자화 목적 기반 모델의 최적화는 다변량(multi-variant)일 수 있다. 레이어당 정밀도(per-layer-precisions)도 다양해질 수 있다. 이러한 경우, 사용된 다양한 레이어별 정밀도에 대해 효율적이고 진보된 평균화 방법(efficient/advance averaging methods)을 적용할 수 있다.

[0076] 분산 병렬 모델 트레이닝에 사용되는 복수의 노드들(820, 850)은 동일한 기본 모델(base model)을 가지지만, 레이어별 정밀도는 모델 민감도 분석(810)에 따른 다른 민감도 리스트들(830, 860)에 기초하여 노드 간에 다를 수 있다. 예를 들어, 제1 노드(820)에 적용되는 제1 민감도 리스트(830)에 따라 각 레이어들에 적용되는 정밀도(840)는 제n 노드(850)에 적용되는 제n 민감도 리스트(860)에 따라 각 레이어들에 적용되는 정밀도(870)와 상이할 수 있다.

[0077] 효율적이고 진보된 평균화를 위해, TopK, 유진 알고리즘이 GRD 단계(880)에 사용될 수 있다. 예를 들어, 복수의 노드들(820, 850) 각각에서 계산된 그라디언트들 중 가장 큰 k개를 이용하여 그라디언트들의 평균 값이 계산될 수 있다. 또한, GRD 단계(880)는 민감도 분석에 피드백을 제공하여, 보다 최적의 양자화 방식을 찾을 수 있게끔 도울 수 있다. 특히, 큰 DNN 모델의 경우, 서로 다른 레이어별 정밀도 구성(different layer-per-layer-precision configurations)을 동시에 사용하여 모델의 수렴 속도를 효과적으로 높일 수 있다.

[0079] 도 9는 일 실시예에 따른 전자 장치의 동작 방법을 나타낸 도면이다.

[0080] 이하 실시예에서 각 동작들은 순차적으로 수행될 수도 있으나, 반드시 순차적으로 수행되는 것은 아니다. 예를 들어, 각 동작들의 순서가 변경될 수도 있으며, 적어도 두 동작들이 병렬적으로 수행될 수도 있다. 동작(910) 내지 동작(920)은 전자 장치의 적어도 하나의 구성요소(예: 프로세서, 가속기 등)에 의해 수행될 수 있다.

[0081] 동작(910)에서, 전자 장치는 트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정한다. 전자 장치는 모델의 트레이닝마다 또는 모델의 트레이닝에서 수행되는 에포크 또는 하나 이상의 이터레이션마다 주기적으로 레이어들의 민감도를 결정할 수 있다. 트레이닝 대상인 모델은 양자화가 적용되지 않은 정밀도로 사전 트레이닝된 모델일 수 있다.

[0082] 동작(920)에서, 전자 장치는 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 모델을 트레이닝한다.

[0083] 전자 장치는 임계치보다 낮은 민감도를 가진 레이어를 양자화하여 제1 정밀도(first precision)로 처리하고, 임계치보다 높거나 같은 민감도를 가진 레이어를 양자화 없이 제1 정밀도보다 높은 제2 정밀도로 처리할 수 있다.

[0084] 전자 장치는 모델의 첫 번째 레이어로부터 마지막 레이어로의 순방향 전파 단계, 모델의 마지막 레이어로부터 첫 번째 레이어로의 역방향 전파 단계, 모델의 분산 트레이닝에 이용되는 복수의 노드들 각각에서 계산된 그라디언트들의 평균 값을 결정하는 단계 및 평균 값에 따라 모델의 가중치를 업데이트하는 단계를 포함하는 분산 트레이닝을 모델에 대해 수행할 수 있다.

[0085] 또한, 전자 장치는 모델에 이용되는 텐서(tensor)의 채널별 민감도(channel-wise sensitivity)를 결정하고, 미리 정해진 제2 임계치보다 낮은 민감도를 가진 채널에 양자화를 적용하여 모델을 트레이닝할 수도 있다.

[0086] 또한, 전자 장치는 모델에 포함된 레이어들의 민감도를 복수의 레벨들로 결정하고, 레이어들을 대응하는 레벨의 정밀도로 양자화를 적용하여 모델을 트레이닝할 수 있다.

[0087] 또한, 전자 장치는 분산 트레이닝에 포함된 복수의 단계들 중 하나 이상의 단계에서 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 모델을 트레이닝할 수 있다.

[0089] 도 10은 일 실시예에 따른 전자 장치를 나타낸 도면이다.

[0090] 도 10을 참조하면, 일 실시예에 따른 전자 장치(1000)는 메모리(1010) 및 프로세서(1020)를 포함한다. 메모리(1010) 및 프로세서(1020)는 버스(bus), PCIe(Peripheral Component Interconnect Express), NoC(Network on a Chip) 등을 통하여 서로 통신할 수 있다.

[0091] 메모리(1010)는 컴퓨터에서 읽을 수 있는 명령어를 포함할 수 있다. 프로세서(1020)는 메모리(1010)에 저장된 명령어가 프로세서(1020)에서 실행됨에 따라 앞서 언급된 동작들을 수행할 수 있다. 메모리(1010)는 휘발성 메모리 또는 비휘발성 메모리일 수 있다.

[0092] 프로세서(1020)는 명령어들, 혹은 프로그램들을 실행하거나, 전자 장치(1000)를 제어하는 장치로서, 예를 들어, 전자 장치(1000)에 포함된 호스트 프로세서 및/또는 가속기를 포함할 수 있다. 호스트 프로세서는 전자 장치(1000)에 포함된 컴포넌트들의 동작을 제어하는 장치로, 예를 들어, 중앙 처리 장치(CPU; Central Processing Unit)를 포함할 수 있다. 가속기는 호스트 프로세서의 명령어에 따라 뉴럴 네트워크를 실행하여 입력되는 데이터를 추론하는 AI 가속기(Artificial Intelligence accelerator)로서, 예를 들어, NPU(neural processing unit), GPU(graphics processing unit), TPU(tensor processing unit), DSP(digital signal processor) 등을 포함할 수 있다.

[0093] 프로세서(1020)는 트레이닝 대상인 모델에 포함된 레이어들의 민감도(sensitivity)를 결정하고, 미리 정해진 임계치보다 낮은 민감도를 가진 레이어에 양자화를 적용하여 모델을 트레이닝한다.

[0094] 전자 장치(1000)는 서버나 특수 설계된 컴퓨팅 장치로 구현될 수 있으나, 실시예가 이로 한정되는 것은 아니며, 이외에도 스마트 폰, 태블릿, 랩탑, 퍼스널 컴퓨터 등 다양한 컴퓨팅 장치, 스마트 시계, 스마트 안경, 스마트 의류 등 다양한 웨어러블 기기, 스마트 스피커, 스마트 TV, 스마트 냉장고 등 다양한 가전장치, 스마트 자동차, 스마트 키오스크, IoT(Internet of Things) 기기, WAD(Walking Assist Device), 드론, 로봇 등 다양한 디바이스로 제한 없이 구현될 수 있다.

[0095] 그 밖에, 전자 장치(1000)에 관해서는 상술된 동작을 처리할 수 있다.

[0097] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 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, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 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 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0098] 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 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0099] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 저장할 수 있으며 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.

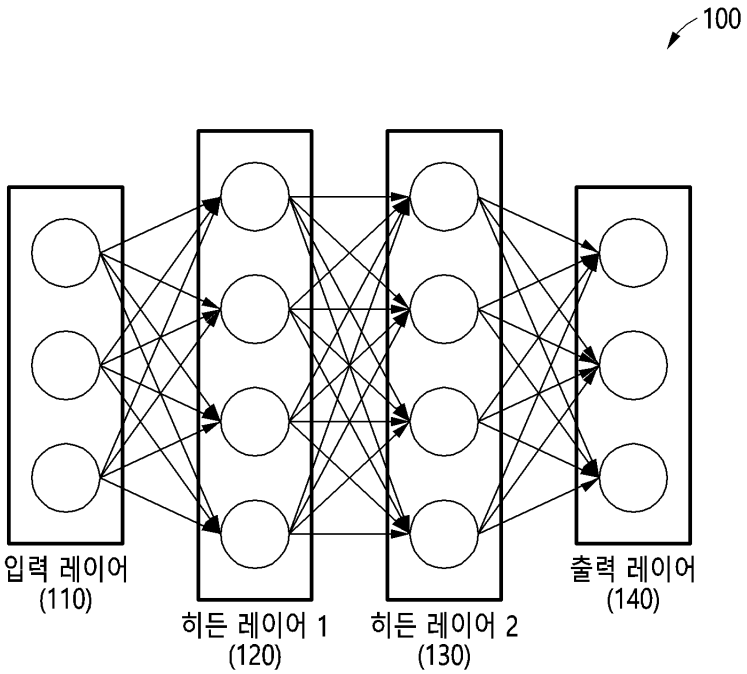
[0100] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0101] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 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.

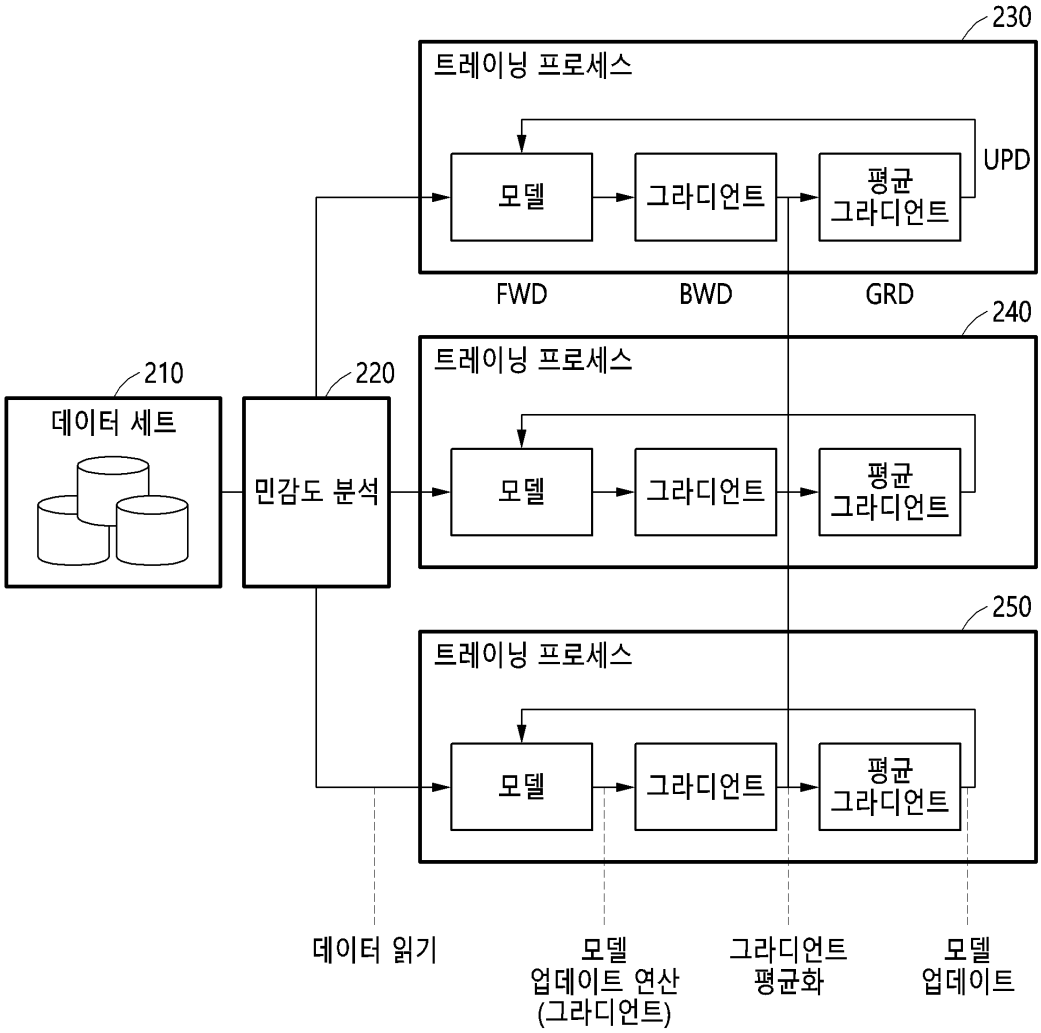
[0102] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

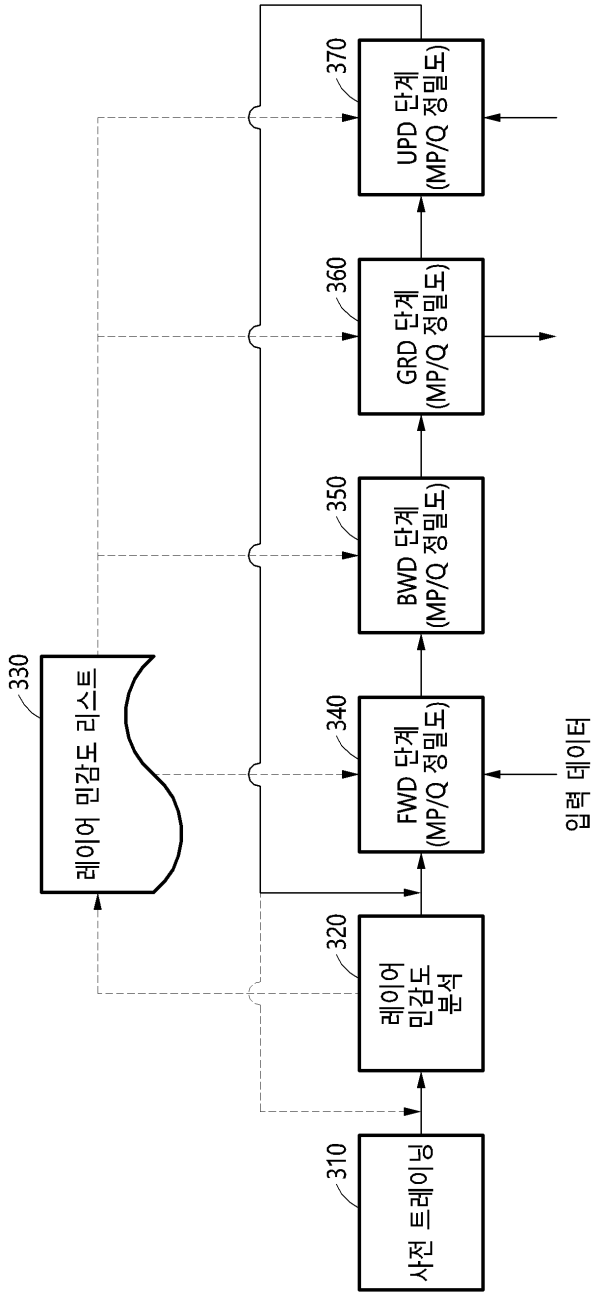
도면1



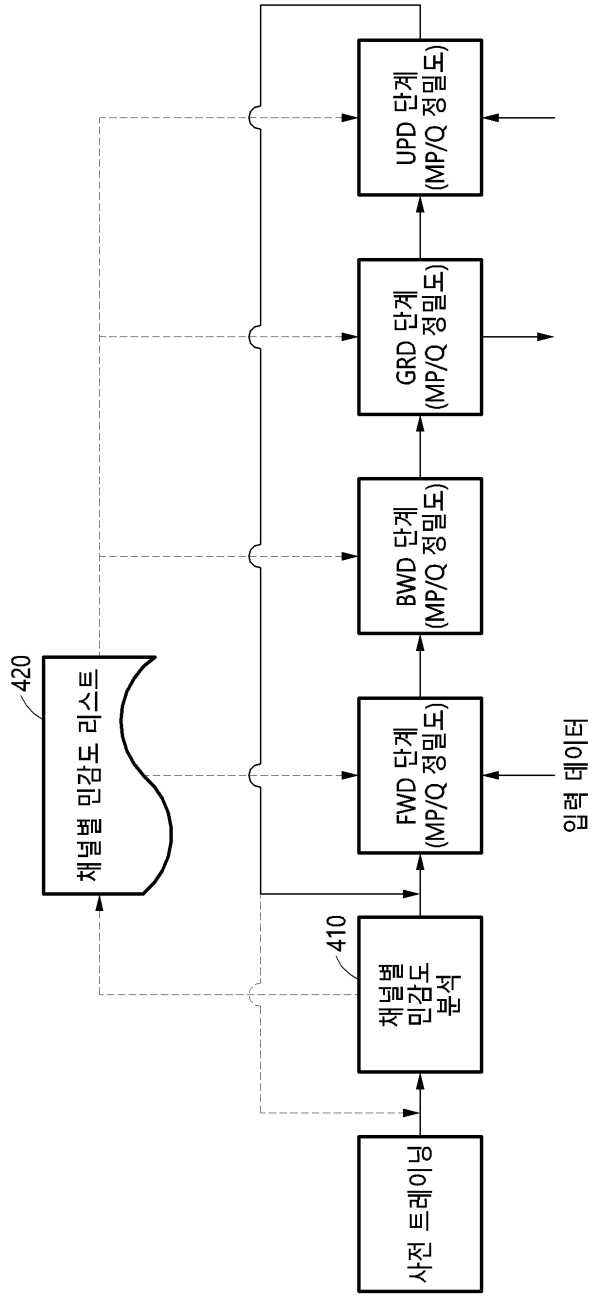
도면2


230
트레이닝 프로세스
모델
그라디언트
평균
그라디언트
UPD
FWD
BWD
GRD
240
트레이닝 프로세스
모델
그라디언트
평균
그라디언트
210
데이터 세트
220
민감도 분석
250
트레이닝 프로세스
모델
그라디언트
평균
그라디언트
데이터 읽기
모델
업데이트 연산
(그라디언트)
그라디언트
평균화
모델
업데이트

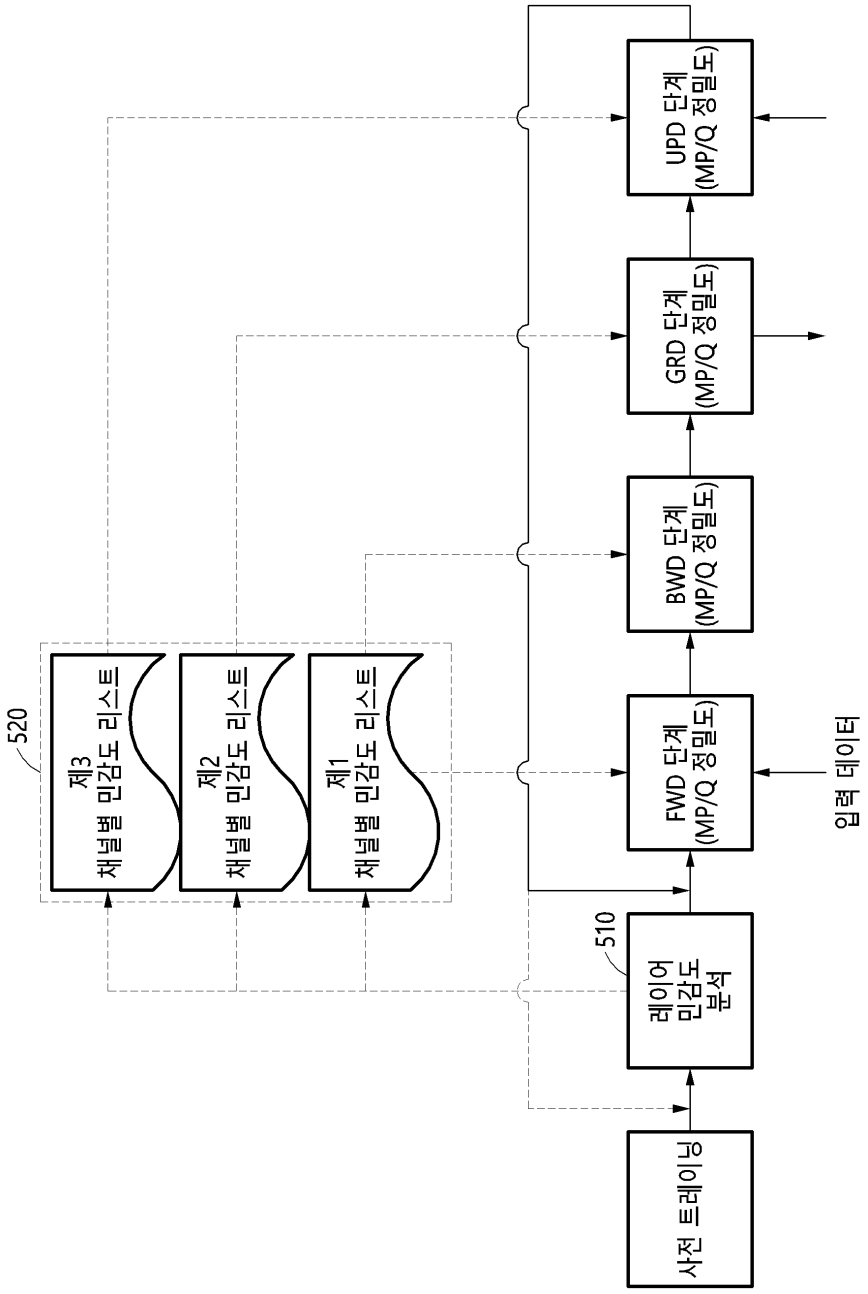
도면3


310
사전 트레이닝
320
레이어 민감도 분석
330
레이어 민감도 리스트
340
FWD 단계 (MP/Q 정밀도)
350
BWD 단계 (MP/Q 정밀도)
360
GRD 단계 (MP/Q 정밀도)
370
UPD 단계 (MP/Q 정밀도)
입력 데이터

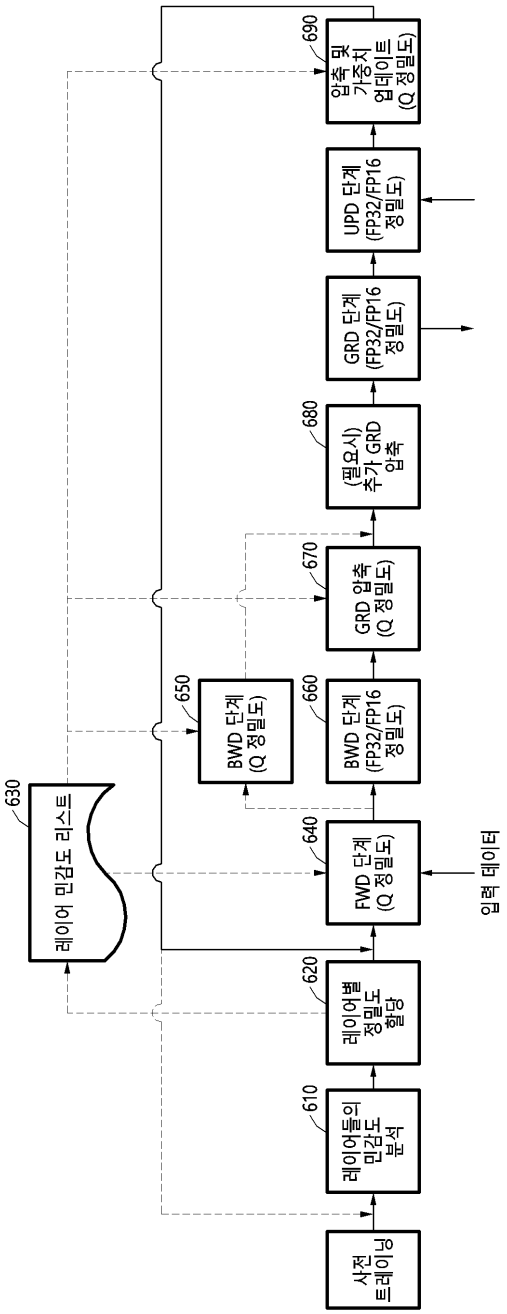
도면4


사전 트레이닝
410
채널별 민감도 분석
420
채널별 민감도 리스트
FWD 단계 (MP/Q 정밀도)
BWD 단계 (MP/Q 정밀도)
GRD 단계 (MP/Q 정밀도)
UPD 단계 (MP/Q 정밀도)
입력 데이터

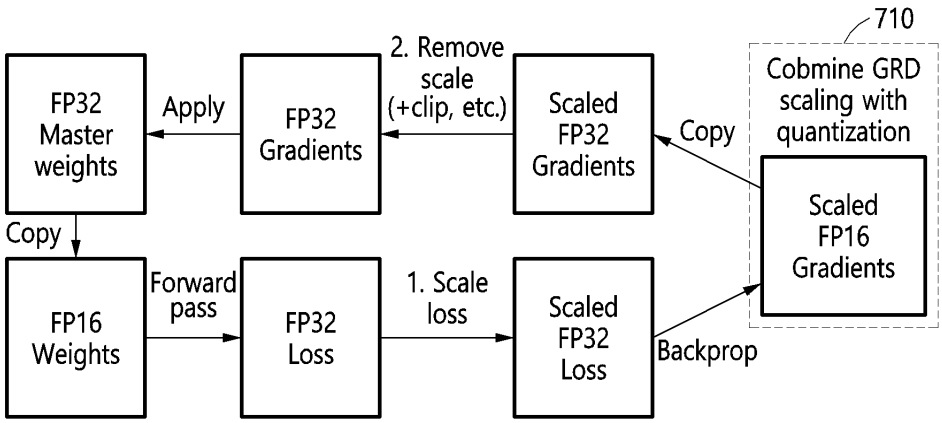
도면5


사전 트레이닝
레이어 민감도 분석
510
520
제3 채널별 민감도 리스트
제2 채널별 민감도 리스트
제1 채널별 민감도 리스트
FWD 단계 (MP/Q 정밀도)
BWD 단계 (MP/Q 정밀도)
GRD 단계 (MP/Q 정밀도)
UPD 단계 (MP/Q 정밀도)
입력 데이터

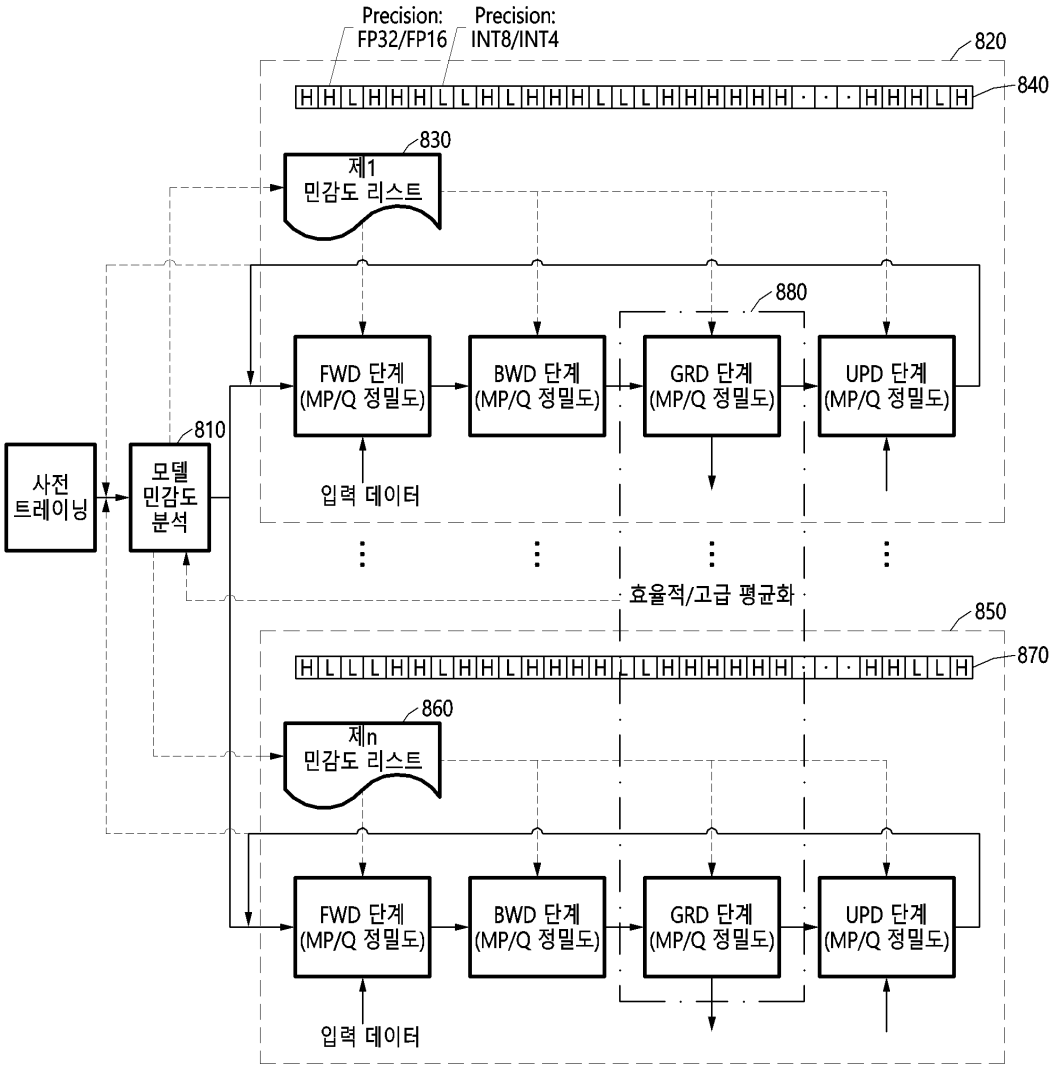
도면6


사전 트레이닝
610
레이어들의 민감도 분석
620
레이어별 정밀도 할당
630
레이어 민감도 리스트
640
FWD 단계 (Q 정밀도)
입력 데이터
650
BWD 단계 (Q 정밀도)
660
BWD 단계 (FP32/FP16 정밀도)
670
GRD 업축 (Q 정밀도)
680
(필요시) 추가 GRD 업축
GRD 단계 (FP32/FP16 정밀도)
UPD 단계 (FP32/FP16 정밀도)
690
업축 및 가중치 업데이트 (Q 정밀도)

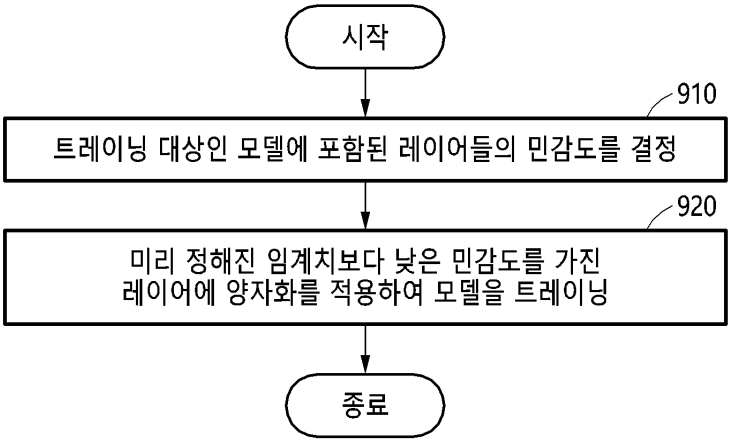
도면7


710
FP32 Master weights
Apply
FP32 Gradients
2. Remove scale (+clip, etc.)
Scaled FP32 Gradients
Copy
Cobmine GRD scaling with quantization
Scaled FP16 Gradients
Copy
FP16 Weights
Forward pass
FP32 Loss
1. Scale loss
Scaled FP32 Loss
Backprop

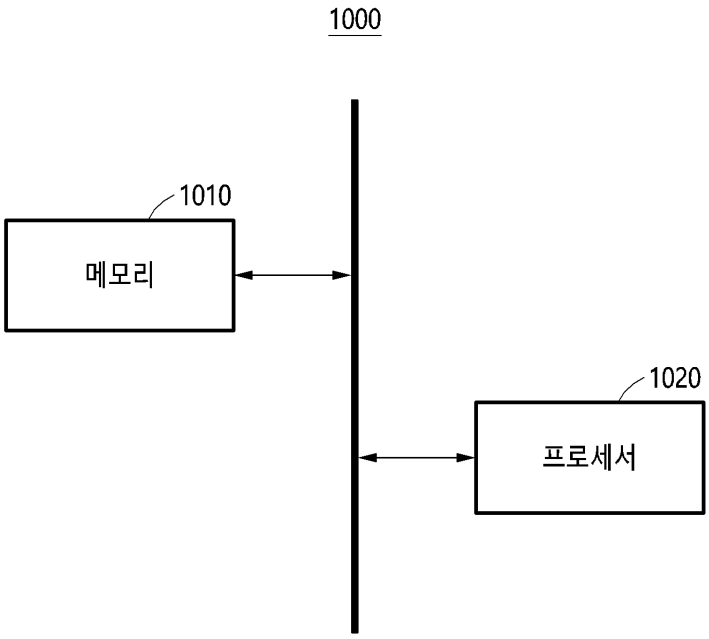
도면8


Precision: FP32/FP16
Precision: INT8/INT4
820
840
830
제1 민감도 리스트
880
FWD 단계 (MP/Q 정밀도)
BWD 단계 (MP/Q 정밀도)
GRD 단계 (MP/Q 정밀도)
UPD 단계 (MP/Q 정밀도)
810
사전 트레이닝
모델 민감도 분석
입력 데이터
효율적/고급 평균화
850
870
860
제n 민감도 리스트
FWD 단계 (MP/Q 정밀도)
BWD 단계 (MP/Q 정밀도)
GRD 단계 (MP/Q 정밀도)
UPD 단계 (MP/Q 정밀도)
입력 데이터

도면9


시작
910
트레이닝 대상인 모델에 포함된 레이어들의 민감도를 결정
920
미리 정해진 임계치보다 낮은 민감도를 가진
레이어에 양자화를 적용하여 모델을 트레이닝
종료

도면10


1000
1010
메모리
1020
프로세서