(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0030307
(43) 공개일자 2025년03월05일

(51) 국제특허분류(Int. Cl.)
G06N 3/0495 (2023.01) *G06F 17/10* (2006.01)
(52) CPC특허분류
G06N 3/0495 (2023.01)
G06F 17/10 (2013.01)
(21) 출원번호 10-2023-0111513
(22) 출원일자 2023년08월24일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
박상우
경기도 수원시 영통구 삼성로 129 (매탄동, 삼성전자)
서태원
경기도 남양주시 순화궁로 18, 4110동 902호 (별내동, 신안인스빌아파트)
(74) 대리인
리앤목특허법인

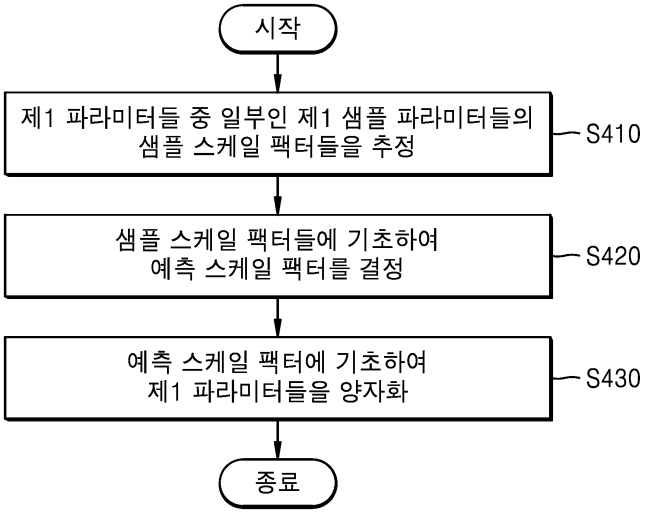
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **인공 신경망의 양자화 방법 및 장치**

(57) 요 약

인공 신경망의 양자화 방법 및 장치가 개시된다. 본 개시의 인공 신경망의 양자화 방법은, 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하는 단계, 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 단계, 및 예측 스케일 팩터에 기초하여 제1 파라미터들을 양자화하는 단계를 포함할 수 있다.

대 표 도 - 도4



명 세 서

청구범위

청구항 1

추론(Inference) 가속화를 위한 인공 신경망(Artificial Neural Network)의 양자화(Quantization) 방법에 있어서,

상기 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하는 단계;

상기 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 단계; 및

상기 예측 스케일 팩터에 기초하여 상기 제1 파라미터들을 양자화하는 단계를 포함하는, 인공 신경망의 양자화 방법.

청구항 2

청구항 1에 있어서,

상기 제1 파라미터들은, 상기 인공 신경망 내 제1 레이어로 입력되는 입력 변수들과 상기 제1 레이어의 가중치(Weight)들을 포함하는 것을 특징으로 하는, 인공 신경망의 양자화 방법.

청구항 3

청구항 1에 있어서,

상기 예측 스케일 팩터를 결정하는 단계는,

상기 샘플 스케일 팩터들의 지수 이동 평균(Exponential Moving Average)에 기초하여 결정되는 것을 특징으로 하는, 인공 신경망의 양자화 방법.

청구항 4

청구항 1에 있어서,

상기 예측 스케일 팩터를 결정하는 단계는,

상기 샘플 스케일 팩터들의 산술 평균(Arithmetic Mean)에 기초하여 결정되는 것을 특징으로 하는, 인공 신경망의 양자화 방법.

청구항 5

청구항 1에 있어서,

상기 예측 스케일 팩터를 결정하는 단계는,

상기 샘플 스케일 팩터들의 최솟값 또는 최댓값에 기초하여 결정되는 것을 특징으로 하는, 인공 신경망의 양자화 방법.

청구항 6

청구항 1에 있어서,

상기 제1 파라미터들을 양자화하는 단계는,

어파인 양자화(Affine Quantization) 또는 스케일 양자화(Scale Quantization)을 통해 양자화하는 것을 특징으로 하는, 인공 신경망의 양자화 방법.

청구항 7

청구항 6에 있어서,

상기 제1 파라미터들을 양자화하는 단계는,

각각의 상기 제1 파라미터들의 스케일 팩터들을 대신하여, 상기 예측 스케일 팩터를 활용하여 상기 제1 파라미터들을 양자화하는 것을 특징으로 하는, 인공 신경망의 양자화 방법.

청구항 8

추론(Inference) 가속화를 위한 인공 신경망(Artificial Neural Network)의 양자화(Quantization) 시스템으로서,

적어도 하나의 프로세서; 및

상기 적어도 하나의 프로세서에 의해서 실행시, 상기 적어도 하나의 프로세서로 하여금 상기 인공 신경망의 양자화 과정을 실행하도록 명령어들을 저장하는 비일시적 저장매체를 포함하고,

상기 인공 신경망의 양자화 과정은,

상기 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하는 단계;

상기 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 단계; 및

상기 예측 스케일 팩터에 기초하여 상기 제1 파라미터들을 양자화하는 단계를 포함하는, 인공 신경망의 양자화 시스템.

청구항 9

청구항 8에 있어서,

상기 인공 신경망의 양자화 과정은,

초기 데이터들을 입력 받음에 따라 상기 적어도 하나의 프로세서에 의해서 실시간으로 실행되는 것을 특징으로 하는, 신경망의 양자화 시스템.

청구항 10

추론(Inference) 가속화를 위한 인공 신경망(Artificial Neural Network)의 양자화(Quantization) 장치로서,

상기 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하고, 상기 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 스케일 팩터 추정 모듈; 및

상기 예측 스케일 팩터에 기초하여 상기 제1 파라미터들을 양자화하는 양자화 모듈을 포함하는, 인공 신경망의 양자화 장치.

발명의 설명

기 술 분 야

[0001] 본 개시의 기술적 사상은 인공 신경망에 관한 것으로서, 자세하게는 인공 신경망의 양자화 방법, 장치 및 시스템에 관한 것이다.

배 경 기 술

[0002] 인공 신경망(Artificial Neural Network)은, 인공 뉴런(Neuron)(또는 뉴런 모델)들의 상호연결된 집합들을 구현하기 위하여 컴퓨팅 기기 또는 컴퓨팅 기기에 의해서 수행되는 방법을 지칭할 수 있다. 인공 뉴런은 입력 데이터에 대한 단순 연산들을 수행함으로써 출력 데이터를 생성할 수 있고, 출력 데이터는 다른 인공 뉴런에 전달될 수 있다. 인공 신경망의 예시로서 심층 신경망(Deep Neural Network) 또는 딥 러닝(Deep Learning)은 멀티 레이어 구조를 가질 수 있다.

[0003] 딥 러닝 추론 연산에 동적 양자화 기술 (Dynamic Quantization) 적용 시, 모델 각 레이어의 입력에 대해서 스케일 팩터가 요구될 수 있다. 딥 러닝 추론에서 모든 레이어마다 모든 입력에 대해 스케일 팩터를 구해 양자화 연산을 수행하게 되는 경우, 상당한 연산량이 요구될 수 있다.

발명의 내용

해결하려는 과제

[0004] 본 개시의 기술적 사상이 해결하려는 과제는, 인공 신경망에서 높은 정확도 및 낮은 계산 복잡도를 동시에 제공하는 인공 신경망의 양자화를 위한 방법 및 장치를 제공하는 데에 있다.

과제의 해결 수단

[0005] 상기와 같은 목적을 달성하기 위하여, 본 개시의 기술적 사상의 일측면에 따른 인공 신경망의 양자화 방법은, 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하는 단계, 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 단계, 및 예측 스케일 팩터에 기초하여 제1 파라미터들을 양자화하는 단계를 포함할 수 있다.

[0006] 본 개시의 기술적 사상의 일측면에 따른 인공 신경망의 양자화 시스템은, 적어도 하나의 프로세서, 및 적어도 하나의 프로세서에 의해서 실행시, 적어도 하나의 프로세서로 하여금 인공 신경망의 양자화 과정을 실행하도록 명령어들을 저장하는 비일시적 저장매체를 포함할 수 있고, 인공 신경망의 양자화 과정은, 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하는 단계, 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 단계, 및 예측 스케일 팩터에 기초하여 제1 파라미터들을 양자화하는 단계를 포함할 수 있다.

[0007] 본 개시의 기술적 사상의 일측면에 따른 인공 신경망의 양자화 장치는, 인공 신경망 내 제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터(Scale Factor)들을 추정하고 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 스케일 팩터 추정 모듈, 및 예측 스케일 팩터에 기초하여 제1 파라미터들을 양자화하는 양자화 모듈을 포함할 수 있다.

발명의 효과

[0008] 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화 방법 및 장치에 의하면, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다.

[0009] 또한, 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화 방법 및 장치에 의하면, 높은 정확도를 가지는 양자화된 인공 신경망에 기인하여 인공 신경망을 구현하기 위한 컴퓨팅 리소스를 감소시킬 수 있고, 인공 신경망의 어플리케이션 범위를 확장할 수 있다.

[0010] 본 개시의 예시적 실시예들에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 아니하며, 언급되지 아니한 다른 효과들은 이하의 설명으로부터 본 개시의 예시적 실시예들이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 도출되고 이해될 수 있다. 즉, 본 개시의 예시적 실시예들을 실시함에 따른 의도하지 아니한 효과들 역시 본 개시의 예시적 실시예들로부터 당해 기술분야의 통상의 지식을 가진 자에 의해 도출될 수 있다.

도면의 간단한 설명

[0011] 도 1은 본 개시의 예시적 실시예에 따른 양자화 시스템을 나타내는 블록도이다.

도 2는 본 개시의 예시적 실시예에 따른 인공 신경망의 예시를 나타내는 도면이다.

도 3a 및 도 3b는 본 개시의 예시적 실시예에 따른 양자화 방법의 종류를 나타내는 도면이다.

도 4는 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화 방법을 나타내는 순서도이다.

도 5는 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화 장치를 나타내는 블록도이다.

도 6은 본 개시의 예시적 실시예에 따른 도 5의 장치 내부의 처리 소자를 나타내는 블록도이다.

도 7은 본 개시의 예시적 실시예에 따른 도 5의 장치 내부의 QSE 유닛을 나타내는 블록도이다.

도 8은 본 개시의 예시적 실시예에 따른 컴퓨팅 시스템을 나타내는 블록도이다.

도 9는 본 개시의 예시적 실시예에 따른 휴대용 컴퓨팅 장치를 나타내는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0012] 이하, 첨부한 도면을 참조하여 본 개시의 실시예에 대해 상세히 설명한다.

[0013] 도 1은 본 개시의 예시적 실시예에 따른 양자화 시스템을 나타내는 블록도이다.

[0014] 도 1은 본 개시의 예시적 실시예에 따른 양자화 시스템(100)을 나타내는 블록도이다. 인공 신경망(Artificial Neural Network; ANN)은, 동물의 두뇌를 구성하는 생물학적 신경망에 착안된 컴퓨팅 시스템을 지칭할 수 있다. 인공 신경망(ANN)은, 규칙-기반(Rule-Based) 프로그래밍과 같이 미리 정의된 조건에 따라 작업을 수행하는 고전적인 알고리즘과 달리, 다수의 샘플들(또는 예시들)을 고려함으로써 작업을 수행하는 것을 학습할 수 있다. 인공 신경망(ANN)은 인공 뉴런(Neuron)(또는 뉴런)들이 연결된 구조를 가질 수 있고, 뉴런들간 연결은 시냅스(Synapse)로 지칭될 수 있다. 뉴런은 수신된 신호를 처리할 수 있고, 처리된 신호를 시냅스를 통해서 다른 뉴런에 전송할 수 있다. 뉴런의 출력은 액티베이션(Activation)으로 지칭될 수 있다. 뉴런 및/또는 시냅스는 변동할 수 있는 가중치(Weight)를 가질 수 있고, 가중치에 따라 뉴런에 의해서 처리된 신호의 영향력이 증가하거나 감소할 수 있다. 특히 개개의 뉴런에 관련된 가중치는 바이어스(Bias)로서 지칭될 수 있다.

[0015] 심층 신경망(Deep Neural Network; DNN) 또는 딥 러닝 구조(Deep Learning Architecture)는 레이어 구조를 가질 수 있고, 특정 레이어의 출력은 후속하는 레이어의 입력이 될 수 있다. 이와 같은 멀티-레이어드(Multi-Layered) 구조에서, 레이어들 각각은 다수의 샘플들에 따라 훈련될 수 있다. 심층 신경망과 같은 인공 신경망은 인공 뉴런들에 각각 대응하는 다수의 프로세싱 노드들에 의해서 구현될 수 있는데, 양호한 결과들, 예컨대 높은 정확도의 결과들을 획득하기 위해서 높은 계산 복잡도(Computational Complexity)가 요구될 수 있고, 이에 따라 많은 컴퓨팅 리소스들이 요구될 수 있다.

[0016] 계산 복잡도를 감소시키기 위하여, 인공 신경망은 양자화(Quantization)될 수 있다. 양자화는, 예컨대 반올림을 통해서 실수(Real Number)를 정수(Integer)로 맵핑하는 것과 같이, 입력 값들을 입력 값들의 개수보다 작은 개수의 값들로 맵핑하는 과정을 지칭할 수 있다. 인공 신경망에서 양자화는, 부동 소수점(Floating Point) 신경망을 정수(Integer) 신경망으로 변환하는 과정을 포함할 수 있다. 예를 들면, 인공 신경망에서 양자화는, 레이어의 액티베이션들, 가중치들 등에 적용될 수 있다. 부동 소수점 숫자들은 부호(Sign), 지수(Exponent) 및 유효숫자(Significant)로 구성될 수 있는 한편, 정수 숫자들은 정수 부분(Integer Part)으로 구성될 수 있다. 일부 예시들에서, 정수 숫자들은 정수 부분은 부호 비트를 포함할 수 있다. 도 1을 참조하면, 입력 데이터 세트(IN)는 부동 소수점 숫자로 표현되고 있으나, 양자화 과정을 거친 출력 데이터 세트(OUT)는 정수 숫자로 표현될 수 있다. 부동 소수점 숫자들을 사용하는 인공 신경망은 높은 정확도 및 계산 복잡도를 가질 수 있는 한편, 정수 숫자들을 사용하는 인공 신경망은 감소된 정확도 및 계산 복잡도를 가질 수 있다.

[0017] 인공 신경망에서 결과의 정확도 및 계산 복잡도 사이 트레이드-오프 관계에 기인하여, 인공 신경망의 양자화는 정확도의 감소를 초래할 수 있고, 정확도의 감소 특성은 양자화의 방식에 의존할 수 있다. 이하에서 도면들을 참조하여 후술되는 바와 같이, 본 개시의 예시적 실시예에 따른 양자화 시스템(100)은 정확도의 감소를 최소화하면서 요건에 따른 양자화를 제공할 수 있고, 이에 따라 양호한 성능을 가지면서도 감소된 복잡도를 가지는 양자화된 신경망이 제공될 수 있다.

[0018] 양자화 시스템(100)은 본 개시의 예시적 실시예들에 따른 양자화를 수행하는 임의의 시스템일 수 있고, 양자화 장치로서 지칭될 수도 있다. 예를 들면, 양자화 시스템(100)은 적어도 하나의 프로세서 및 메모리를 포함하는 컴퓨팅 시스템일 수 있다. 비제한적인 예시로서, 양자화 시스템(100)은, 데스크탑 컴퓨터, 서버 등과 같은 고정형(Stationary) 컴퓨팅 시스템뿐만 아니라, 랩탑 컴퓨터, 스마트 폰 등과 같은 휴대용(Mobile) 컴퓨팅 시스템일 수도 있다. 도 1에 도시된 바와 같이, 양자화 시스템(100)은 추정기(120) 및 양자화기(140)를 포함할 수 있고, 근사화기(120) 및 양자화기(140) 각각은 논리 합성을 통해서 구현되는 로직 블록, 프로세서에 의해서 수행되는 소프트웨어 블록 또는 이들의 조합으로 구현될 수 있다. 일부 실시예들에서, 추정기(120) 및 양자화기(140) 각각은, 프로세서에 의해서 실행되는 복수의 명령어들의 집합으로서 프로시저일 수 있고, 프로세서에 의해서 엑세스 가능한 메모리에 저장될 수 있다.

[0019] 도 1을 참조하면, 양자화 시스템(100)은 입력 데이터 세트(IN)를 수신할 수 있고, 양자화된 출력 데이터 세트

(OUT)를 생성할 수 있다. 입력 데이터 세트(IN)는 복수의 입력 데이터들을 포함할 수 있고, 양자화된 출력 데이터 세트(OUT)는 복수의 출력 데이터들을 포함할 수 있다. 도 1에 도시된 바와 같이, 예시적 실시예에 따른 양자화 시스템(100)은 입력 데이터 세트(IN)로서 매트리스 형태의 데이터들을 입력받을 수 있고, 양자화된 출력 데이터 세트(OUT)로서 매트리스 형태의 데이터들을 출력할 수 있다. 일예시에 따른 데이터들은 이미지로부터 추출된 각 픽셀의 데이터들을 의미할 수 있다. 입력 데이터 세트(IN)을 구성하는 각 입력 데이터들은 부동 소수점 숫자들로 구성될 수 있다. 양자화 시스템을 거쳐 출력되는 출력 데이터 세트(OUT)를 구성하는 각 출력 데이터들은 정수 숫자들로 구성될 수 있다. 후술되는 바와 같이, 본 개시의 예시적 실시예에 따른 양자화 시스템(100)은 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다.

[0020] 추정기(120)는 입력 데이터 세트(IN)를 수신할 수 있고, 액티베이션 또는 가중치 등의 예측 스케일 팩터를 결정하여 양자화기(140)에 제공할 수 있다. 부동 소수점 숫자 형식의 변수들을 양자화하기 위해서는 스케일 팩터가 요구될 수 있다. 스케일 팩터는 양자화 과정에서 값의 범위를 맞추기 위해 최댓값/최솟값을 사용하는 비트의 최대숫자에 맞추기 위한 값을 의미할 수 있다. 스케일 팩터에 관한 구체적인 설명은 도 3a 및 도 3b에서 상세히 후술될 것이다. 변수들을 양자화하는 과정은 동적 양자화(Dynamic Quantization) 방법과 정적 양자화(Static Quantization) 방법 등으로 분류될 수 있다. 본 개시의 예시적 실시예에 따른 양자화 시스템(100)은 학습 이후 추론 단계에서 인공 신경망 내 가중치와 액티베이션 등이 실시간으로 양자화되는, 동적 양자화 방법에 의해 데이터들을 양자화할 수 있다. 인공 신경망 내 가중치와 액티베이션 등의 양자화 대상이 되는 데이터들은 본 명세서에서 편의상 파라미터로 통칭하기로 한다. 양자화 대상이 되는 파라미터들 모두에 대해 스케일 팩터를 계산하거나 추정하게 되는 경우, 높은 정확도를 가질 수 있다는 장점이 있으나, 과도한 연산량이 요구될 수 있다. 후술될 바와 같이, 본 개시의 예시적 실시예에 따른 추정기(120)는 양자화 대상이 되는 파라미터들 중 일부에 대해 스케일 팩터를 추정하고, 추정된 스케일 팩터들로부터 예측 스케일 팩터를 결정하여 양자화기(140)에 제공할 수 있다. 양자화 시스템(100)은 파라미터들 전체가 아닌 일부에 대해서만 스케일 팩터를 추정하여, 예측 스케일 팩터를 통해 양자화를 진행하므로, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다.

[0021] 양자화기(140)는 추정기(120)로부터 파라미터들의 예측 스케일 팩터를 수신할 수 있고, 예측 스케일 팩터를 기반으로 파라미터들을 양자화함으로써 양자화된 출력 데이터 세트(OUT)을 생성할 수 있다. 양자화기(140)의 구체적인 양자화 과정에 대해서는 도 7 등에서 상세히 후술될 것이다.

[0022] 도 2는 본 개시의 예시적 실시예에 따른 인공 신경망의 예시를 나타내는 도면이다.

[0023] 도 2를 참조하면, 인공 신경망(NN)은 입력 레이어(L1), 히든 레이어들(L1, L2, 쩐) 및 출력 레이어(Ln)(n은 3 이상의 자연수)를 포함하는 구조를 가지며, 수신되는 입력 데이터(예를 들어, I_1 및 I_2)를 기초로 연산을 수행하고, 수행 결과를 기초로 출력 데이터(예를 들어, O_1 및 O_2)를 생성할 수 있다.

[0024] 인공 신경망(NN)은 1개 이상의 히든 레이어를 포함하는 딥 인공 신경망(Deep Neural Network: DNN) 또는 n-계층 인공 신경망(n-Layers Neural Networks)일 수 있다. 예를 들어, 도 2에 도시된 바와 같이, 인공 신경망(NN)은 입력 레이어(L1), 히든 레이어들(L2 및 L3) 및 출력 레이어(Ln)를 포함하는 DNN일 수 있다. DNN은 Convolutional Neural Networks(CNN), Recurrent Neural Networks(RNN), Deep Belief Networks, Restricted Boltzman Machines 등을 포함할 수 있으나, 이에 제한되지 않는다.

[0025] 인공 신경망(NN)은 DNN 구조를 갖는 경우 유효한 정보를 추출할 수 있는 보다 많은 레이어들을 포함하므로, 인공 신경망(NN)은 종래의 인공 신경망보다 복잡한 데이터 집합들을 처리할 수 있다. 인공 신경망(NN)은 도 2에 도시된 것과는 다른 다양한 구조의 레이어들을 포함할 수도 있다.

[0026] 인공 신경망(NN)에 포함된 레이어들(L1~Ln) 각각은 뉴런, 유닛 또는 이와 유사한 용어들로 알려진, 복수의 인공 노드들을 포함할 수 있다. 예를 들어, 도 2에 도시된 바와 같이, 입력 레이어(L1)는 2개의 노드(ND)들, 히든 레이어들(L2, L3)는 3개의 노드(ND)들을 포함할 수 있다. 다만, 이는 예시에 불과할 뿐 인공 신경망(NN)에 포함된 레이어들 각각은 다양한 개수의 노드(ND)들을 포함할 수 있으며, 각각의 레이어에 포함되는 노드(ND)들의 개수들은 상이할 수 있다.

[0027] 인공 신경망(NN)에 포함된 레이어들 각각에 포함된 노드들은 서로 연결되어 데이터를 교환할 수 있다. 예를 들어, 하나의 노드(ND)는 다른 노드(ND)들로부터 데이터를 수신하여 연산할 수 있고, 연산 결과를 또 다른 노드

(ND)들로 출력할 수 있다.

[0028] 노드(ND)들 각각의 입력 및 출력은 액티베이션(Activation)이라고 지칭될 수 있다. 액티베이션은 한 노드(ND)의 출력 값이면서, 다음 레이어에 포함된 노드(ND)들의 입력 값일 수 있다. 한편, 노드(ND)들 각각은 이전 레이어에 포함된 노드(ND)들로부터 수신된 액티베이션들 및 가중치들에 기초하여 자신의 액티베이션을 결정할 수 있다. 가중치는 각 노드(ND)에서의 액티베이션을 계산하기 위해 이용되는 네트워크 파라미터로서, 노드(ND)들 간의 연결관계에 할당되는 값일 수 있다. 예를 들어, 제2 레이어(L2)에서, 노드(ND)들은 이전 레이어(L1)에서 수신된 액티베이션들(a_{11}, a_{12}), 가중치들(w_{21}, w_{22}, w_{23}, w_{24}, w_{25}, w_{26}), 및 바이어스들(b_{21}, b_{22}, b_{23})에 기초하여 자신의 액티베이션을 결정할 수 있다. 노드(ND)들 각각은 입력을 받아 액티베이션을 출력하는 연산 유닛(Computational Unit)일 수 있고, 입력-출력을 매핑할 수 있다.

[0029] 인공 신경망(NN)은 레이어들 사이에 활성화 함수(Activation Function)를 포함할 수 있다. 활성화 함수는 이전 레이어의 출력을 다음 레이어의 입력으로 변환할 수 있다. 예를 들면, 활성화 함수는 예컨대 ReLU(Rectified Linear Unit), PReLU(Parametric Rectified Linear Unit), 쌍곡탄젠트(hyperbolic tangent; tanh), 시그모이드(sigmoid) 함수와 같이 비선형 함수일 수 있고, 제2 레이어(L2) 및 제3 레이어(L3) 사이에서 제2 레이어(L2)의 출력을 비선형적으로 변환할 수 있다. 액티베이션은 이전 레이어로부터 수신된 액티베이션들의 가중 합(Weighted Sum)에 활성화 함수를 적용시킴으로써 획득된 값일 수도 있다.

[0030] 계속하여, 도 1을 참조하면, 도 1의 양자화 시스템(100)이 양자화하는 대상은 인공 신경망(NN)의 액티베이션 내지 가중치 등을 포함할 수 있다. 예를 들어, 제2 레이어(L2)로 입력되는 제1 레이어(L1)로부터 수신된 액티베이션들(a_{11}, a_{12}), 가중치들(w_{21}, w_{22}, w_{23}, w_{24}, w_{25}, w_{26})은 부동 소수점 숫자 형식의 데이터일 수 있다. 양자화 시스템(100)은 전술한 바와 같은 양자화 과정을 통해, 액티베이션들(a_{11}, a_{12}), 가중치들(w_{21}, w_{22}, w_{23}, w_{24}, w_{25}, w_{26})을 정수 숫자 형식의 데이터들로 양자화할 수 있다. 양자화된 엑티베이션들과 가중치들을 통해 가중 합을 계산함으로써 딥 러닝 모델의 사이즈를 축소시키고, 연산량을 감소시킬 수 있다. 이후 계산된 가중 합 값에 활성화 함수를 적용시킴에 따라, 제2 레이어(L2)에서 제3 레이어로 입력되는 액티베이션들(a_{21}, a_{22}, a_{23})은 다시 부동 소수점 숫자 형식의 데이터일 수 있다. 즉, 양자화 시스템이 양자화하는 대상은 특정 레이어의 파라미터들일 수도 있으며, 전체 레이어 각각에 대응되는 파라미터들일 수도 있다.

[0031] 도 3a 및 도 3b는 본 개시의 예시적 실시예에 따른 양자화 방법의 종류를 나타내는 도면이다.

[0032] 구체적으로, 도 3a는 어파인 양자화(Affine Quantization) 방법을, 도 3b는 스케일 양자화(Scale Quantization) 방법에 대해 나타낸다.

[0033] 도 3a를 참조하면, 어파인 양자화 방법은 양자화 전의 데이터들의 최댓값과 최솟값을 각각 양자화하고자 하는 데이터 형식의 최댓값과 최솟값에 선형적으로 대응시킬 수 있다. 예를 들어, 도 3a에 도시된 바와 같이 양자화되기 전의 데이터들의 최댓값(α)은 7, 양자화되기 전의 데이터들의 최솟값(β)은 -5일 수 있다. 가령 8 bit의 정수 숫자 형식으로 데이터를 양자화하는 경우, 양자화된 데이터는 -128에서 127의 범위의 값을 가질 수 있다. 어파인 양자화 방법에 의하면, 양자화되기 전의 데이터들의 최댓값(α)을 127에, 양자화되기 전의 데이터들의 최솟값(β)을 -128에 대응시킬 수 있다. 그 사이의 데이터 값들은 선형적으로 대응될 수 있다. 따라서 양자화되기 전의 데이터들의 최댓값(α)과 양자화되기 전의 데이터들의 최솟값(β)이 상이한 값을 가지는 경우, 양자화되기 전의 0 값은 양자화 후 0이 아닌 다른 값에 대응될 수 있다. 선형적으로 값을 대응시키는 과정에서 양자화 되기 전의 데이터들의 값의 범위와 양자화된 후의 데이터들의 값의 범위 간의 비가 스케일 팩터가 될 수 있다. 즉, 어파인 양자화 방법에서 스케일 팩터는 아래의 [수학식 1]과 같이 계산될 수 있다.

[0034] [수학식 1]

[0035] $$SF(Scale\ Factor) = \frac{2^b - 1}{\alpha - \beta}$$

[0036] [수학식 1]에서 b는 양자화되는 정수 숫자 데이터의 비트 수를 의미한다. 도 3에서 예시된 값을 참조하면, 도

3a의 어파인 양자화 방법의 스케일 팩터는 $\frac{255}{12}$ 일 수 있다.

[0037] 도 3b를 참조하면, 스케일 양자화 방법은 어파인 양자화 방법과 유사하게, 양자화 전의 데이터들의 최댓값과 최솟값을 각각 양자화하고자 하는 데이터 형식의 최댓값과 최솟값에 선형적으로 대응시킬 수 있다. 다만 스케일 양자화 방법은 양자화되기 전의 0 값을 양자화 후 0에 대응시키기 위해, $\alpha = -\beta$ 를 만족시키도록 할 수 있다. 즉, 스케일 양자화 방법은 양자화되기 전의 데이터들의 절댓값의 최댓값(α)을 구하고, α와 $-\alpha$를 각각 양자화하고자 하는 데이터 형식의 최댓값과 최솟값에 선형적으로 대응시킬 수 있다. 선형적으로 값을 대응시키는 과정에서 양자화 되기 전의 데이터들의 값의 범위와 양자화된 후의 데이터들의 값의 범위 간의 비가 스케일 팩터가 될 수 있다. 즉, 스케일 양자화 방법에서 스케일 팩터는 아래의 [수학식 2]와 같이 계산될 수 있다.

[0038] [수학식 2]

[0039] $$SF = \frac{2^{b-1} - 1}{\alpha}$$

[0040] [수학식 2]에서 b는 양자화되는 정수 숫자 데이터의 비트 수를 의미한다. 도 3에서 예시된 값을 참조하면, 도 3b의 스케일 양자화 방법의 스케일 팩터는 $\frac{127}{7}$ 일 수 있다.

[0041] 도 3a와 도 3b의 양자화 방법 모두, 양자화되기 전의 데이터에 스케일 팩터를 곱해줌으로써 양자화된 후의 데이터 범위 내에 값을 위치시키도록 할 수 있다. 양자화 방법은 스케일 팩터를 나눈 이후, 정수 형태로 표현하기 위한 반올림(Round) 연산과, 범위를 벗어난 숫자들을 보정하는 클립(Clip) 연산을 포함할 수 있으며, 이에 대해서는 도 7에서 상세히 후술될 것이다. 도 3a와 도 3b의 양자화 방법 모두, 스케일 팩터를 계산 내지 추정하기 위해 데이터들의 최댓값 내지 최솟값을 계산하는 과정이 요구될 수 있다. 여러 데이터 세트들이 입력되는 경우, 모든 데이터 세트들에 대하여 스케일 팩터들을 계산하게 되는 경우, 과도한 연산량이 요구될 수 있다. 후술될 바와 같이, 본 개시의 예시적 실시예에 의하면, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다.

[0042] 도 4는 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화 방법을 나타내는 순서도이다.

[0043] 예를 들면, 도 4의 양자화 방법은 도 1의 양자화 시스템(100)에 의해서 수행될 수 있다. 이하에서, 도 4는 도 1을 참조하여 설명될 것이다.

[0044] 도 4를 참조하면, 단계 S410에서 제1 샘플 파라미터들의 샘플 스케일 팩터들을 추정하는 동작이 수행될 수 있다. 제1 샘플 파라미터들은 제1 파라미터들의 일부일 수 있다. 제1 파라미터들은 양자화 대상이 되는 제1 레이어로 입력되는 입력 변수들, 내지는 제1 레이어의 가중치들을 포함할 수 있다. 제1 레이어로 입력되는 입력 변수들은 제1 레이어의 이전 레이어로부터 출력되는 액티베이션일 수 있다.

[0045] 제1 샘플 파라미터들은 제1 파라미터들의 일부로서, 후술되는 예측 스케일 팩터를 결정하기 위한 파라미터들일 수 있다. 제1 샘플 파라미터들은 사전에 정의된 샘플이 아닌, 입력되는 제1 파라미터들 내부에서 일부 샘플링하여 선택될 수 있다. 예를 들어, 제1 파라미터들이 총 A개의 데이터 세트들을 포함하는 경우, 제1 샘플 파라미터들은 입력되는 제1 파라미터들의 첫 B개의 데이터 세트들을 의미할 수 있다(B는 자연수 A보다 작은 자연수). 즉, 제1 샘플 파라미터들은 제1 파라미터들 중, 시간상 먼저 입력되는 일련의 파라미터들을 지칭할 수 있다. 입력되는 제1 파라미터들의 첫 B개의 데이터 세트들은 제1 샘플 파라미터들로서, 각각의 제1 샘플 파라미터들에 대한 스케일 팩터를 추정 내지 계산하는 동작이 수행될 수 있다. 제1 샘플 파라미터에 대응되는 스케일 팩터는 도 1 및 도 3에서 전술된 방법에 의해 추정될 수 있다. 예를 들어, 도 1에서와 같이 입력 데이터 세트(IN)가 제1 샘플 파라미터로 입력되는 경우, 도 3에서 전술된 바와 같이 데이터 세트 내 최댓값 내지 최솟값을 구하여 스케일 팩터를 계산 내지 추정할 수 있다. 데이터 세트 내 최댓값 내지 최솟값을 구하기 위해 비교기(Comparator) 활용 또는 비교 연산이 수행될 수 있으며, 이에 대한 내용은 도 7에서 상세히 후술될 것이다.

[0046] 도 4를 참조하면, 단계 S420에서 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 동작이 수행될 수 있다. 이전 단계 S410에서 샘플 스케일 팩터들이 추정된 이후, 샘플 스캐일 팩터들로부터 전체 제1 파라미터들을 대표할 수 있는 예측 스케일 팩터를 결정할 수 있다. 전술된 바와 같이, 전체 파라미터들에 대해 모두 스케일 팩터를 계산하는 경우, 과도한 연산량이 요구될 수 있으므로 본 개시의 양자화 방법은 예측 스케일 팩터를

양자화 계산에 활용할 수 있다. 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 방법은 다양한 수식을 활용할 수 있으며, 데이터 세트들의 특성이나 연관성 등에 따라 다른 수식을 선택할 수 있다. 일예시에 따른 예측 스케일 팩터를 결정하는 방법은 아래의 [수학식 3]과 같이 계산될 수 있다.

[0047] [수학식 3]

[0048] $$ESF(Estimated\ \ Scale\ \ Factor) = \frac{ESF + SF}{2}$$

[0049] [수학식 3]은 컴퓨터 또는 프로그램으로 수행될 수 있는 연산식으로서, 샘플 스케일 팩터(SF)들이 순차적으로 입력됨에 따른, 예측 스케일 팩터(ESF)의 연산 방법에 대해 나타내는 연산식이다. [수학식 3]은 구체적으로, 지수 이동 평균(Exponential Moving Average)을 반영하여 예측 스케일 팩터를 계산할 수 있다. [수학식 3]은 예측 스케일 팩터(ESF)의 연산 방법 중 비례계수가 1/2인 경우로서, [수학식 3]으로만 제한되지 아니한다. [수학식 3]의 조금 더 일반적인 수식은 아래의 수학식 4]와 같다.

[0050] [수학식 4]

[0051] $$ESF = (1-k)\cdot ESF + k\cdot SF$$

[0052] [수학식 4] 또한 컴퓨터 또는 프로그램으로 수행될 수 있는 연산식으로서, 샘플 스케일 팩터(SF)들이 순차적으로 입력됨에 따른, 예측 스케일 팩터(ESF)의 연산 방법에 대해 나타내는 연산식이다. k는 0과 1사이의 실수로서, k 값이 클수록 이전 관측치가 더 빨리 감소될 수 있다. k가 1/2인 경우, [수학식 3]이 된다.

[0054] 또다른 일예시에 따른 예측 스케일 팩터를 결정하는 방법은 아래의 [수학식 4]와 같이 계산될 수 있다.

[0055] [수학식 5]

[0056] $$ESF += \frac{SF}{(Number\ \ of\ \ Sample)}$$

[0057] [수학식 5]는 컴퓨터 또는 프로그램으로 수행될 수 있는 연산식으로서, 샘플 스케일 팩터(SF)들이 순차적으로 입력됨에 따른, 예측 스케일 팩터(ESF)의 연산 방법에 대해 나타내는 연산식이다. [수학식 5]는 구체적으로, 산술 평균(Arithmetic Mean)을 반영하여 예측 스케일 팩터를 계산할 수 있다.

[0058] 또다른 일예시에 따른 예측 스케일 팩터를 결정하는 방법은 아래의 [수학식 6]와 같이 계산될 수 있다.

[0059] [수학식 6]

[0060] $$ESF = \min(ESF, SF)$$

[0061] [수학식 6]은 컴퓨터 또는 프로그램으로 수행될 수 있는 연산식으로서, 샘플 스케일 팩터(SF)들이 순차적으로 입력됨에 따른, 예측 스케일 팩터(ESF)의 연산 방법에 대해 나타내는 연산식이다. [수학식 6]은 구체적으로, 샘플 스케일 팩터의 최솟값을 반영하여 예측 스케일 팩터를 계산할 수 있다.

[0062] 또다른 일예시에 따른 예측 스케일 팩터를 결정하는 방법은 아래의 [수학식 7]과 같이 계산될 수 있다.

[0063] [수학식 7]

[0064] $$ESF = \max(ESF, SF)$$

[0065] [수학식 7]은 컴퓨터 또는 프로그램으로 수행될 수 있는 연산식으로서, 샘플 스케일 팩터(SF)들이 순차적으로 입력됨에 따른, 예측 스케일 팩터(ESF)의 연산 방법에 대해 나타내는 연산식이다. [수학식 7]은 구체적으로, 샘플 스케일 팩터의 최댓값을 반영하여 예측 스케일 팩터를 계산할 수 있다.

[0066] [수학식 3] 내지 [수학식 7] 등의 수식은 시간상 순차적으로 입력되는 샘플 파라미터들에 대해 순차적으로 샘플 스케일 팩터를 계산하여, 예측 스케일 팩터 값을 누적반영 하는 방식으로 계산될 수 있다. 또한 샘플 파라미터들에 대응되는 샘플 스케일 팩터들을 전부 계산한 후, 전체 샘플 스케일 팩터들을 기초로 하여 예측 스케일 팩터 값을 산출할 수도 있다.

[0067] [수학식 3] 내지 [수학식 7] 등의 수식을 통해, 단계 S420에서 샘플 스케일 팩터들에 기초하여 예측 스케일 팩

터를 결정하는 동작이 수행될 수 있다. 다만 이는 예시적인 수식으로서, 샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정하는 방법은 이에 의해서만 제한되지 아니한다. 예를 들어, [수학식 3] 내지 [수학식 7] 등의 수식에 적절한 계수(coefficient)를 더하거나 곱하는 방식으로 샘플 스케일 팩터들로부터 산출되는 예측 스케일 팩터 값을 조절할 수 있다. 또한, 상기 수학식들을 포함한 수학식들을 복수 개 결합하여 예측 스케일 팩터 값을 산출할 수도 있다. 적절한 수식의 활용을 통해, 인공 신경망의 연산량을 감소시키면서도 정확도가 확보되는 양자화 동작이 수행될 수 있다.

[0068] 도 4를 참조하면, 단계 S430에서 예측 스케일 팩터에 기초하여 제1 파라미터들을 양자화하는 동작이 수행될 수 있다. 양자화 동작은 도 1의 양자화기(140)의 동작으로 수행될 수 있으며, 도 7의 장치에 의해서 수행될 수도 있다. 제1 파라미터들을 양자화하는 방법으로는 도 3에서 전술된 어파인 양자화 방법 내지는 스케일 양자화 방법 등이 활용될 수 있다. 구체적으로, 이전 단계에서 계산된 예측 스케일 팩터는 모든 파라미터들에 대해 스케일 팩터를 계산하는 과정을 대신하여 계산된 팩터로서, 제1 파라미터들의 양자화 과정에 이용될 수 있다. 제1 파라미터들 각각의 스케일 팩터를 계산하는 대신, 계산된 예측 스케일 팩터를 곱하고 계산하여 양자화할 수 있다. 전술된 실시예에 의하면, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다. 또한 높은 정확도를 가지는 양자화된 인공 신경망에 기인하여 인공 신경망을 구현하기 위한 컴퓨팅 리소스를 감소시킬 수 있고, 인공 신경망의 어플리케이션 범위를 확장할 수 있다.

[0069] 도 5는 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화 장치를 나타내는 블록도이다.

[0070] 도 5를 참조하면, 인공 신경망의 양자화 장치(500)는 메모리(510)와 프로세서(520)를 포함할 수 있다. 이하에서 도 5는 도 1 및 도 4 등을 참조하여 서술될 것이며, 이전 도면에서의 서술과 중복되는 서술은 생략한다.

[0071] 양자화 장치(500)의 메모리(510)는 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화를 위한 프로그램을 저장할 수 있고, 양자화 방법에 따라 양자화된 액티베이션 내지 데이터들을 저장할 수도 있다. 또한, 메모리(510)는 도 1의 양자화된 출력 데이터 세트(OUT)를 저장할 수도 있으며, 바이어스 또는 양자화 중에 생성된 데이터 등을 저장할 수도 있다. 메모리(510)는 양자화 실행과 관련된 일련의 명령어들을 저장할 수 있다. 메모리(510)는 프로세서(520)과 상호 통신하며 데이터를 주고 받을 수 있다.

[0072] 프로세서(520)는 CPU(Central Processing Unit), GPU(Graphic Processing Unit), 및 NPU(Neural Processing Unit) 중 어느 하나일 수 있다. 다만 이는 예시적인 것으로서, 프로세서(520)는 열거된 예시로만 제한되지 아니한다. 도 5에서 예시하는 프로세서(520)는 데이터 패쳐(522), 벡터 프로세싱 유닛(524) 및 내적 어레이(526)를 포함할 수 있다. 데이터 패쳐(522)는 메모리(51)로부터 가중치, 액티베이션, 바이어스 등의 데이터들을 전달받을 수 있고, 벡터 프로세싱 유닛(524)으로 데이터를 전달할 수 있다. 벡터 프로세싱 유닛(524)은 내부에 QSE(Quantization with Scale factor Estimation) 유닛(524_1)을 포함할 수 있다. 도 5의 도시는 예시적인 것으로, QSE 유닛(524_1)은 벡터 프로세싱 유닛(524) 외부에 독립적으로 기능할 수도 있다. QSE 유닛(524_1)은 입력 받은 가중치, 액티베이션 등의 파라미터들에 대응되는 예측 스케일 팩터들을 추정하고, 이에 기반하여 양자화를 수행할 수 있다. QSE 유닛(524_1)에 관한 구체적인 동작은 도 7에서 상세히 후술될 것이다.

[0073] 벡터 프로세싱 유닛(524)는 양자화된 파라미터들을 데이터 패쳐(522)로 전달할 수 있고, 데이터 패쳐(522)는 이를 내적 어레이(526)으로 전달할 수 있다. 내적 어레이(526)은 입력 받은 양자화된 가중치, 액티베이션, 바이어스 등의 데이터를 기초로 하여 가중 합을 계산할 수 있다. 도 5에는 N개의 프로세싱 유닛이 포함되는 것으로 도시하고 있으나, 이는 예시적인 것으로 다른 개수의 프로세싱 유닛(PE)을 포함할 수도 있다. 프로세싱 유닛(PE)은 양자화된 가중치와 양자화된 액티베이션 간의 가중 합을 계산할 수 있으며, 프로세싱 유닛(PE)에 관한 구체적인 동작은 도 6에서 상세히 후술될 것이다. 내적 어레이(526)는 양자화되지 않은 파라미터들을 기반으로 가중 합을 계산하는 경우, 많은 연산량이 요구되며, 딥 러닝의 추론 단계에서 많은 시간과 자원이 소요될 수 있다. 본 개시의 예시적 실시예에 의하면, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다.

[0074] 내적 어레이(526)를 거쳐 출력되는 액티베이션 등의 데이터들은 다시 벡터 프로세싱 유닛(524)으로 전달되고, 벡터 프로세싱 유닛(524)은 데이터 패쳐(522)로부터 추가적인 데이터를 전달 받아, 다음 레이어의 양자화 단계를 진행할 수 있다. 또한 벡터 프로세싱 유닛(524)은 내적 어레이(526)으로부터 전달받은 출력 엑티베이션 데이터 등을 저장할 수 있도록 메모리(510)로 전달할 수 있다.

[0075] 도 5에서 도시되지는 아니하였으나, 내부에 데이터를 저장하는 버퍼나 레지스터의 구성이 추가적으로 포함될 수

있으며, 데이터를 선택하거나 합치는 에더(Adder), 멀티플렉서(Multiplexer)등의 구성이 추가적으로 포함될 수 있다.

[0076] 도 6은 본 개시의 예시적 실시예에 따른 도 5의 장치 내부의 처리 소자를 나타내는 블록도이다.

[0077] 도 5 및 도 6을 참조하면, 장치(500) 내부의 내적 어레이(526)은 다수의 처리 소자(PE)(526_N)들을 포함할 수 있다. 처리 소자(526_N)는 다수의 멀티플라이어(Multiplier)(MPR) 와 에더 트리(Adder Tree)를 포함할 수 있다. 처리 소자(526_N)는 아래의 [수학식 8]과 같은 연산을 수행할 수 있다.

[0078] [수학식 8]

[0079] $$y = \sum_{k=1}^{X} I_k W_k + Bias$$

[0080] [수학식 8]에서 I는 입력되는 양자화된 액티베이션, W는 입력되는 양자화된 가중치, Bias는 바이어스, y는 출력을 의미한다. 처리 소자(526_N)는 X개의 양자화된 액티베이션과 양자화된 가중치, 그리고 바이어스 값을 입력 받을 수 있다. X는 연산이 수행되는 레이어의 채널 수를 나타낼 수 있으며, 레이어마다 상이한 값을 가질 수 있다. 또한 도 6에서는 X개의 멀티플라이어(MPR)가 있는 것으로 도시되었으나, 이는 예시적인 것으로, 처리 소자(526_N)는 하드웨어 스펙에 따라 상이한 수의 멀티플라이어(MPR)를 포함할 수도 있다. 멀티플라이어(MPR)를 통해 곱해진 양자화된 액티베이션과 양자화된 가중치는 바이어스와 함께 에더 트리에서 더해질 수 있다. 결과적으로, 처리 소자(526_N)는 X개의 양자화된 액티베이션과 양자화된 가중치, 그리고 바이어스 값을 입력 받아 가중합을 계산할 수 있다.

[0081] 도 6에 도시되지는 아니하였으나, 예시적인 처리 소자(526_N)는 경우에 따라 데이터를 저장하기 위한 레지스터나 메모리 등을 포함할 수도 있으며, 활성화 함수를 적용시키기 위한 소자들을 추가적으로 포함할 수도 있다.

[0082] 도 7은 본 개시의 예시적 실시예에 따른 도 5의 장치 내부의 QSE 유닛을 나타내는 블록도이다.

[0083] 도 5 및 도 7을 참조하면, 장치(500) 내부의 벡터 프로세싱 유닛(524)은 QSE 유닛(524_1)을 포함할 수 있다. QSE 유닛(524_1)은 스케일 팩터 추정기(710)와 양자화기(720)를 포함할 수 있다. 스케일 팩터 추정기(710)와 양자화기(720)는 도 1의 추정기(120)와 양자화기(140)의 일예시일 수도 있으며, 일부 공통된 기능을 수행할 수도 있다. 양자화 방법의 동작과 관련하여 도 1 및 도 4에서 서술된 내용과 중복되는 서술은 생략한다.

[0084] QSE 유닛(524_1)은 액티베이션, 가중치 등의 파라미터들의 스케일 팩터 추정 및 양자화를 수행할 수 있다. QSE 유닛(524_1)은 제1 파라미터들을 입력 받아, 양자화된 파라미터들을 출력할 수 있다. 이때, 각각의 파라미터는 데이터 세트들을 지칭할 수 있으며, 각각의 데이터 세트들은 N개의 데이터들을 포함할 수 있다. 예를 들어, 3×3의 픽셀을 가지는 이미지의 경우, 하나의 이미지 데이터 세트는 9개의 픽셀에 대응되는 9개의 데이터들을 포함할 수 있으며, 첫 레이어에서의 양자화 연산에서 N은 9일 수 있다. 이 경우, 제1 파라미터들 각각은 9개의 데이터들을 포함하며, QSE 유닛(524_1)은 파라미터를 구성하는 입력 데이터들(O_1, O_2, 쩐, O_N)을 입력 받아, 양자화된 출력 데이터들(Q_1, Q_2, 쩐, Q_N)을 출력할 수 있다. 양자화된 출력 데이터들(Q_1, Q_2, 쩐, Q_N)은 하나의 양자화된 출력 파라미터를 구성할 수 있다. 즉, 도 7에서 예시하는 입력 출력은 하나의 파라미터, 즉 하나의 데이터 세트에 대응되는 동작을 예시한 것이다. 예를 들어, 1000장의 3×3의 픽셀을 가지는 이미지를 처리하는 경우, 첫 레이어는 9개의 입력 데이터들(O_1, O_2, 쩐, O_N)을 입력 받아, 양자화된 출력 데이터들(Q_1, Q_2, 쩐, Q_N)을 출력할 수 있다. 또한 해당 데이터들의 입출력을 1000장의 이미지들에 대해 각각 수행할 수 있다.

[0085] QSE 유닛(524_1) 내부의 스케일 팩터 추정기(710)는 제1 샘플 파라미터들의 샘플 스케일 팩터들을 추정할 수 있다. 제1 샘플 파라미터들은 제1 파라미터들의 일부일 수 있다. 도 4에서 전술된 바와 같이, 예시적 실시예에 따른 제1 샘플 파라미터들은 제1 파라미터들 중 첫 일부 파라미터들을 지칭할 수 있다. 스케일 팩터 추정기(710)는 제1 샘플 파라미터들에 대해 각각 샘플 스케일 팩터들을 추정하기 위해, 다수의 비교기(Comparator)(CPR)들을 통해 샘플 파라미터 내부의 최댓값과 최솟값을 계산할 수 있다. 도 3a 및 도 3b에서 전술된 바와 같이, 스케일 팩터 추정기(710)는 다수의 비교기(CPR)들을 통해 계산된 최댓값과 최솟값에 기초하여, 샘플 스케일 팩터를 계산할 수 있다.

[0086] 스케일 팩터 추정기(710)는 내부에 샘플 스케일 팩터와 예측 스케일 팩터를 저장하기 위한 레지스터(711)를 포함할 수 있다. 스케일 팩터 추정기(710)는 B개의 제1 샘플 파라미터, 즉 B개의 데이터 세트들을 입력받을 때까

지는 샘플 스케일 팩터를 상술한 다수의 비교기(CPR)들을 통해 계산할 수 있다. 계산된 샘플 스케일 팩터는 양자화기(720)로 전달되며, 레지스터(711)에도 저장할 수 있다. 스케일 팩터 추정기(710)는 도 4에서 전술된 바와 같이, [수학식 3] 내지 [수학식 6] 등의 수식과 계산된 샘플 스케일 팩터에 기초하여, 예측 스케일 팩터를 계산할 수 있다.

[0087] QSE 유닛(524_1)에 제1 샘플 파라미터들이 모두 입력된 후, 즉, B개의 데이터 세트들을 입력받은 후 스케일 팩터 추정기(710)의 레지스터(711)는 최종적으로 예측 스케일 팩터가 계산하여 저장할 수 있다. 이후 QSE 유닛(524_1)이 제1 파라미터들을 입력받는 경우, 스캐일 펙터 추정기(710)는 파라미터 내부의 데이터들 간에 비교 계산을 통한 최댓값 내지 최솟값 계산하는 연산 과정을 수행하지 않을 수 있다. 즉, 스케일 팩터 추정기(710)는, 예측 스케일 팩터 계산이 완료된 이후, 제1 파라미터들의 입력에 따라 스케일 팩터를 각각 추정하지 않고, 레지스터(711)에 저장된 예측 스케일 팩터를 양자화기(720)로 출력할 수 있다. 모든 파라미터, 즉 모든 데이터 세트들에 대하여 스케일 팩터 연산을 수행하지 않고, 추정된 예측 스케일 팩터를 활용하여 양자화 연산을 수행할 수 있다. 따라서, 본 개시의 예시적 실시예에 의하면, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다.

[0088] 양자화기(720)는 입력 데이터들(O_1, O_2, 쩐, O_N)과, 스케일 팩터 추정기(710)로부터 스케일 팩터(SF)를 입력받아, 양자화된 출력 데이터들 양자화된 출력 데이터들(Q_1, Q_2, 쩐, Q_N)을 출력할 수 있다. 제1 샘플 파라미터들이 입력되는 구간에서는 스케일 팩터 추정기(710)로부터 양자화기(720)로 계산된 샘플 스케일 팩터가 입력될 수 있다. 도 7에는 양자화기(720)로 스케일 팩터(SF)가 입력되는 것으로 도시되어 있으나, 이는 포괄적인 표기로서, 예측 스케일 팩터의 계산이 완료된 이후에는 전술된 바와 같이, 레지스터(711)에 저장된 예측 스케일 팩터가 양자화기(720)로 입력될 수 있다. 즉, 처음 샘플 파라미터들이 입력되는 동안에는 양자화기(720)로 샘플 스케일 팩터가 스케일 팩터 추정기(710)의 일련의 계산을 거쳐 입력될 수 있다. 샘플 파라미터들이 입력된 이후에는 양자화기(720)로 레지스터(711)에 저장된 예측 스케일 팩터가 입력될 수 있다.

[0089] 양자화기(720)는 입력 데이터들(O_1, O_2, 쩐, O_N) 각각에 멀티플라이어(MPR)들을 통해 예측 스케일 팩터를 곱할 수 있다. 이어서 양자화기(720)는 곱해진 값을 반올림(Round)하고, 클립(Clip) 연산을 통해 최종적인 양자화된 출력 데이터들(Q_1, Q_2, 쩐, Q_N)을 출력할 수 있다. 양자화기(720) 내 반올림(Round) 연산부는 출력 데이터를 정수 숫자 형태로 나타낼 수 있다.

[0090] 클립 연산부의 연산은 [수학식 9]와 같이 표현될 수 있다.

[0091] [수학식 9]

[0092]
$$\mathrm{clip}(x,l,u)=\begin{cases} l & (if\ x<l) \\ x & (if\ l\le x<u) \\ u & (if\ u<x) \end{cases}$$

[0093] [수학식 9]에서 x는 입력되는 데이터 값, l은 표현하고자 하는 양자화된 데이터의 최솟값, u는 표현하고자 하는 양자화된 데이터의 최댓값을 나타낸다. 양자화기(720) 내 클립(Clip) 연산부는 출력 데이터가 최댓값 내지 최솟값 사이에 들어오도록 데이터 값을 조정할 수 있다.

[0094] 전술된 실시예에 의하면, 딥 러닝 추론 시 입력들에 대한 스케일 팩터들을 추정하여 양자화 연산에서 스케일 팩터들을 찾는 연산을 생략하여 총 연산량을 감소시킬 수 있다. 또한 높은 정확도를 가지는 양자화된 인공 신경망에 기인하여 인공 신경망을 구현하기 위한 컴퓨팅 리소스를 감소시킬 수 있고, 인공 신경망의 어플리케이션 범위를 확장할 수 있다.

[0095] 도 8은 본 개시의 예시적 실시예에 따른 컴퓨팅 시스템을 나타내는 블록도이다.

[0096] 일부 실시예들에서, 도 1의 양자화 시스템(100) 내지 도 5의 양자화 장치(500)는 도 8의 컴퓨팅 시스템(2000)으로 구현될 수 있다. 도 8에 도시된 바와 같이, 컴퓨팅 시스템(2000)은 시스템 메모리(2100), 프로세서(2300), 스토리지(2500), 입출력 장치들(2700) 및 통신 접속들(2900)을 포함할 수 있다. 컴퓨팅 시스템(2000)에 포함된 구성요소들은 상호 통신가능하게, 예컨대 버스를 통해서 연결될 수 있다.

[0097] 시스템 메모리(2100)는 프로그램(2120)을 포함할 수 있다. 프로그램(2120)은 프로세서(2300)로 하여금 본 개시의 예시적 실시예들에 따른 인공 신경망의 양자화를 수행하도록 할 수 있다. 예를 들면, 프로그램(2120)은 프로세서(2300)에 의해서 실행가능한(Executable) 복수의 명령어들을 포함할 수 있고, 프로그램(2120)에 포함된 복수의 명령어들이 프로세서(2300)에 의해서 실행됨으로써 인공 신경망의 양자화가 수행될 수 있다. 시스템 메모

리(2100)는, 비제한적인 예시로서, SRAM(Static Random Access Memory), DRAM(Dynamic Random Access Memory)와 같은 휘발성 메모리를 포함할 수도 있고, 플래시 메모리(Flash Memory) 등과 같은 비휘발성 메모리를 포함할 수도 있다.

[0098] 프로세서(2300)는 임의의 명령어 세트(예컨대, IA-32(Intel Architecture-32), 64 비트 확장 IA-32, x86-64, PowerPC, Sparc, MIPS, ARM, IA-64 등)을 실행할 수 있는 적어도 하나의 코어를 포함할 수 있다. 프로세서(2300)는 시스템 메모리(2100)에 저장된 명령어들을 실행할 수 있으며, 프로그램(2120)을 실행함으로써 인공 신경망의 양자화를 수행할 수 있다.

[0099] 스토리지(2500)는 컴퓨팅 시스템(2000)에 공급되는 전력이 차단되더라도 저장된 데이터를 소실하지 아니할 수 있다. 예를 들면, 스토리지(2500)는 EEPROM(Electrically Erasable Programmable Read-Only Memory), 플래시 메모리(Flash Memory), PRAM(Phase Change Random Access Memory), RRAM(Resistance Random Access Memory), NFGM(Nano Floating Gate Memory), PoRAM(Polymer Random Access Memory), MRAM(Magnetic Random Access Memory), FRAM(Ferroelectric Random Access Memory) 등과 같은 비휘발성 메모리를 포함할 수도 있고, 자기 테이프, 광학 디스크, 자기 디스크와 같은 저장 매체를 포함할 수도 있다. 일부 실시예들에서, 스토리지(2500)는 컴퓨팅 시스템(2000)으로부터 탈착 가능할 수도 있다.

[0100] 일부 실시예들에서, 스토리지(2500)는 본 개시의 예시적 실시예에 따른 인공 신경망의 양자화를 위한 프로그램(2120)을 저장할 수 있고, 프로그램(2120)이 프로세서(2300)에 의해서 실행되기 이전에 스토리지(2500)로부터 프로그램(2120) 또는 그것의 적어도 일부가 시스템 메모리(2100)로 로딩될 수 있다. 일부 실시예들에서, 스토리지(2500)는 프로그램 언어로 작성된 파일을 저장할 수 있고, 파일로부터 컴파일러 등에 의해서 생성된 프로그램(2120) 또는 그것의 적어도 일부가 시스템 메모리(2100)로 로딩될 수도 있다.

[0101] 일부 실시예들에서, 스토리지(2500)는 프로세서(2300)에 의해서 처리될 데이터 및/또는 프로세서(2300)에 의해서 처리된 데이터를 저장할 수 있다. 예를 들면, 스토리지(2500)는 전술된 양자화 방법에 따라 양자화된 액티베이션 내지 데이터들을 저장할 수도 있고, 도 1의 양자화된 출력 데이터 세트(OUT)를 저장할 수도 있으며, 바이어스 또는 양자화 중에 생성된 데이터 등을 저장할 수도 있다.

[0102] 입출력 장치들(2700)은 키보드, 포인팅 장치 등과 같은 입력 장치를 포함할 수 있고, 디스플레이 장치, 프린터 등과 같은 출력 장치를 포함할 수 있다. 예를 들면, 사용자는 입출력 장치들(2700)을 통해서, 프로세서(2300)에 의한 프로그램(2120)의 실행을 트리거할 수도 있고, 도 2의 입력 데이터(예를 들어, I_1 및 I_3)를 입력할 수도 있고, 도 1의 양자화된 출력 데이터 세트(OUT) 및/또는 오류 메시지 등을 확인할 수도 있다.

[0103] 통신 접속들(2900)은 컴퓨팅 시스템(2000) 외부의 네트워크에 대한 액세스를 제공할 수 있다. 예를 들면, 네트워크는 다수의 컴퓨팅 시스템들 및 통신 링크들을 포함할 수 있고, 통신 링크들은 유선 링크들, 광학 링크들, 무선 링크들 또는 임의의 다른 형태의 링크들을 포함할 수 있다.

[0104] 도 9는 본 개시의 예시적 실시예에 따른 휴대용 컴퓨팅 장치를 나타내는 블록도이다.

[0105] 일부 실시예들에서, 본 개시의 예시적 실시예에 따라 양자화된 인공 신경망은 휴대용 컴퓨팅 장치(3000)에서 구현될 수 있다. 휴대용 컴퓨팅 장치(3000)는 비제한적인 예시로서, 모바일 폰, 태블릿 PC, 웨어러블 기기, 사물인터넷 장치 등과 같이 배터리 또는 자가 발전을 통해서 전력이 공급되는 임의의 휴대용 전자 기기일 수 있다.

[0106] 도 9에 도시된 바와 같이, 휴대용 컴퓨팅 장치(3000)는 메모리 서브시스템(3100), 입출력 장치들(3300), 프로세싱 유닛(3500) 및 네트워크 인터페이스(3700)를 포함할 수 있고, 메모리 서브시스템(3100), 입출력 장치들(3300), 프로세싱 유닛(3500) 및 네트워크 인터페이스(3700)는 버스(390)를 통해서 상호 통신할 수 있다. 일부 실시예들에서, 메모리 서브시스템(3100), 입출력 장치들(3300), 프로세싱 유닛(3500) 및 네트워크 인터페이스(3700) 중 적어도 2개 이상은, 시스템-온-칩(System-on-a-Chip; SoC)으로서 하나의 패키지에 포함될 수 있다.

[0107] 메모리 서브시스템(3100)은 RAM(3120) 및 스토리지(3140)를 포함할 수 있다. RAM(3120) 및/또는 스토리지(3140)는 프로세싱 유닛(3500)에 의해서 실행되는 명령어들 및 처리되는 데이터를 저장할 수 있다. 예를 들면, RAM(3120) 및/또는 스토리지(3140)는 인공 신경망의 신호들, 가중치들, 바이어스들과 같은 변수들을 저장할 수 있고, 인공 신경망의 인공 뉴런(또는 계산 노드)의 파라미터들을 저장할 수도 있다. 일부 실시예들에서, 스토리지(3140)는 비휘발성 메모리를 포함할 수 있다.

[0108] 프로세싱 유닛(3500)은 CPU(Central Processing Unit)(3520), GPU(Graphic Processing Unit)(3540),

DSP(Digital Signal Processor)(3560) 및 NPU(Neural Processing Unit)(3580)를 포함할 수 있다. 도 9에 도시된 바와 상이하게, 일부 실시예들에서, 프로세싱 유닛(3500)은 CPU(Central Processing Unit)(3520), GPU(Graphic Processing Unit)(3540), DSP(Digital Signal Processor)(3560) 및 NPU(Neural Processing Unit)(3580) 중 적어도 일부만을 포함할 수도 있다.

[0109] CPU(3520)는 휴대용 컴퓨팅 장치(3000)의 전체적인 동작, 예컨대 입출력 장치들(3300)을 통해서 수신된 외부 입력에 응답하여 특정 작업을 직접 수행하거나, 프로세싱 유닛(3500)의 다른 구성요소들에게 수행을 지시할 수 있다. GPU(3540)는 입출력 장치들(3300)에 포함된 디스플레이 장치를 통해서 출력되는 영상을 위한 데이터를 생성할 수도 있고, 입출력 장치들(3300)에 포함된 카메라로부터 수신되는 데이터를 인코딩할 수도 있다. DSP(3560)는 디지털 신호, 예컨대 네트워크 인터페이스(3700)로부터 제공되는 디지털 신호를 처리함으로써 유용한 데이터를 생성할 수 있다.

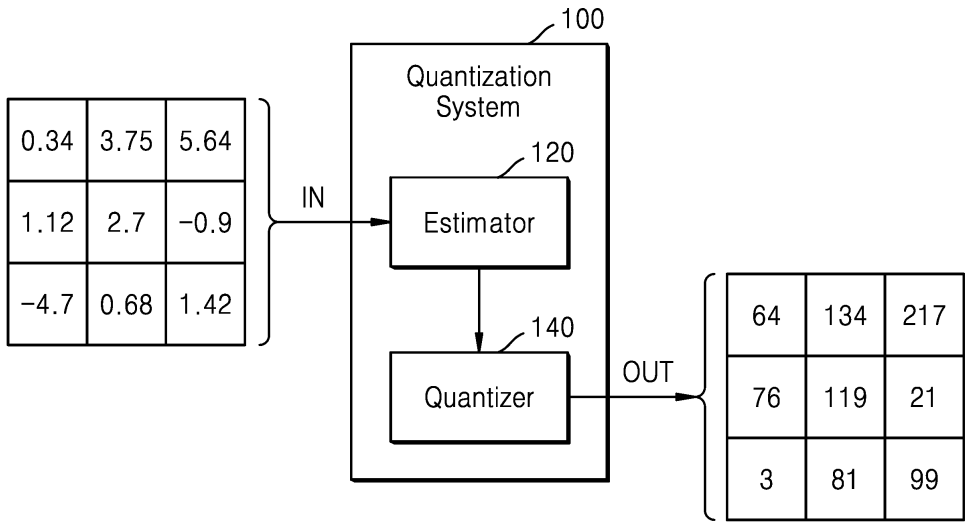
[0110] NPU(3580)는 인공 신경망을 위한 전용의 하드웨어로서, 인공 신경망을 구성하는 적어도 일부의 인공 뉴런에 대응하는 복수의 계산 노드들을 포함할 수 있고, 복수의 계산 노드들 중 적어도 일부는 병렬적으로 신호를 처리할 수 있다. 본 개시의 예시적 실시예에 따라 양자화된 인공 신경망, 예컨대 심층 신경망은 높은 정확도뿐만 아니라 낮은 계산 복잡도를 가지므로, 도 9의 휴대용 컴퓨팅 장치(3000)에 용이하게 구현될 수 있고, 빠른 처리 속도를 가질 수 있으며, 예컨대 단순하고 작은 규모의 NPU(3580)에 의해서도 구현될 수 있다.

[0111] 입출력 장치들(3300)은 터치 입력 장치, 사운드 입력 장치, 카메라 등과 같은 입력 장치들 및 디스플레이 장치 및 사운드 출력 장치 등과 같은 출력 장치들을 포함할 수 있다. 네트워크 인터페이스(3700)는 휴대용 컴퓨팅 장치(3000)에 LTE(Long Term Evolution), 5G 등과 같은 이동 통신 네트워크에 대한 액세스를 제공할 수도 있고, 와이파이와 같은 로컬 네트워크에 대한 액세스를 제공할 수도 있다.

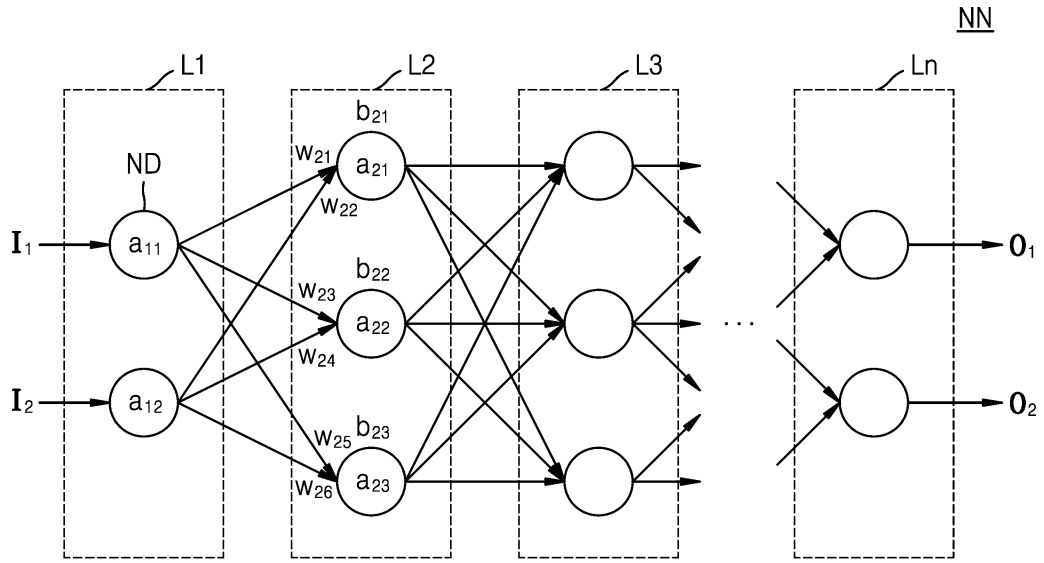
[0112] 이상에서와 같이 도면과 명세서에서 예시적인 실시예들이 개시되었다. 본 명세서에서 특정한 용어를 사용하여 실시예들을 설명되었으나, 이는 단지 본 개시의 기술적 사상을 설명하기 위한 목적에서 사용된 것이지 의미 한정이나 특허청구범위에 기재된 본 개시의 범위를 제한하기 위하여 사용된 것은 아니다. 그러므로 본 기술분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서, 본 개시의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

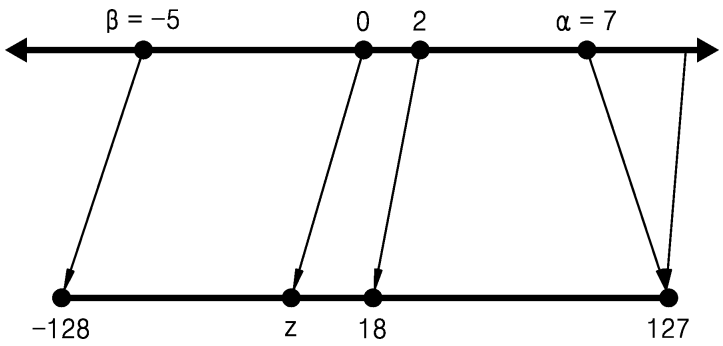
도면1



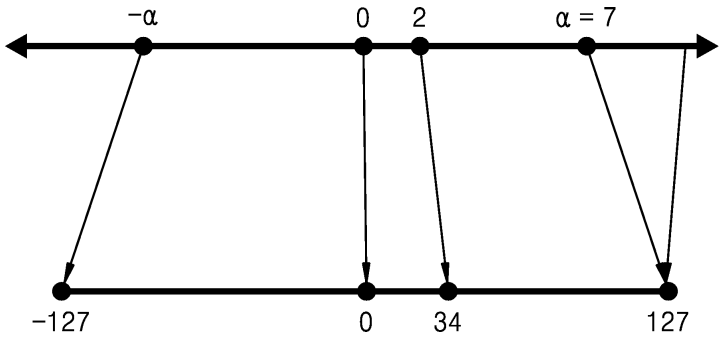
도면2



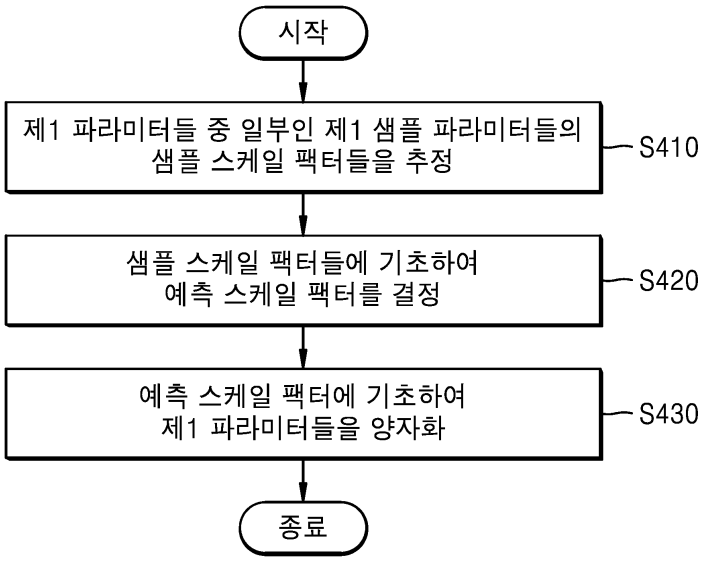
도면3a



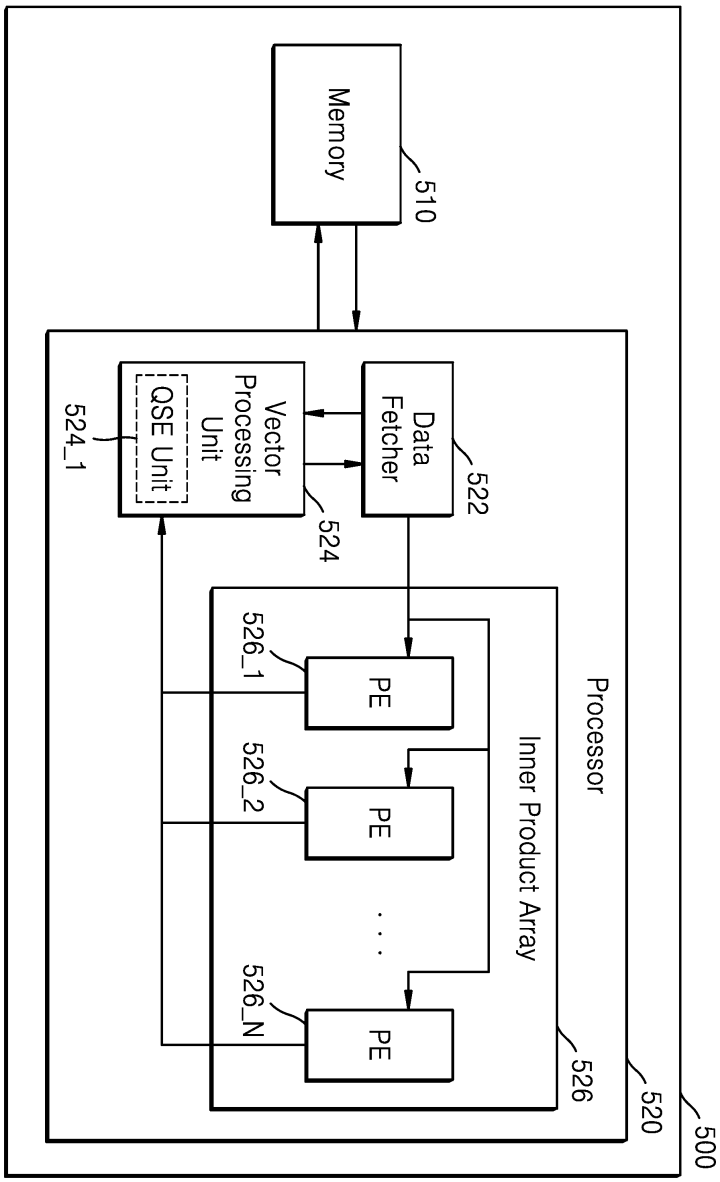
도면3b



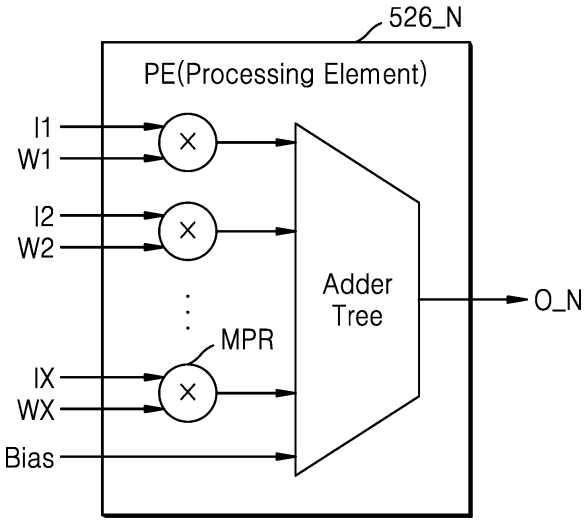
도면4


시작
제1 파라미터들 중 일부인 제1 샘플 파라미터들의 샘플 스케일 팩터들을 추정
S410
샘플 스케일 팩터들에 기초하여 예측 스케일 팩터를 결정
S420
예측 스케일 팩터에 기초하여 제1 파라미터들을 양자화
S430
종료

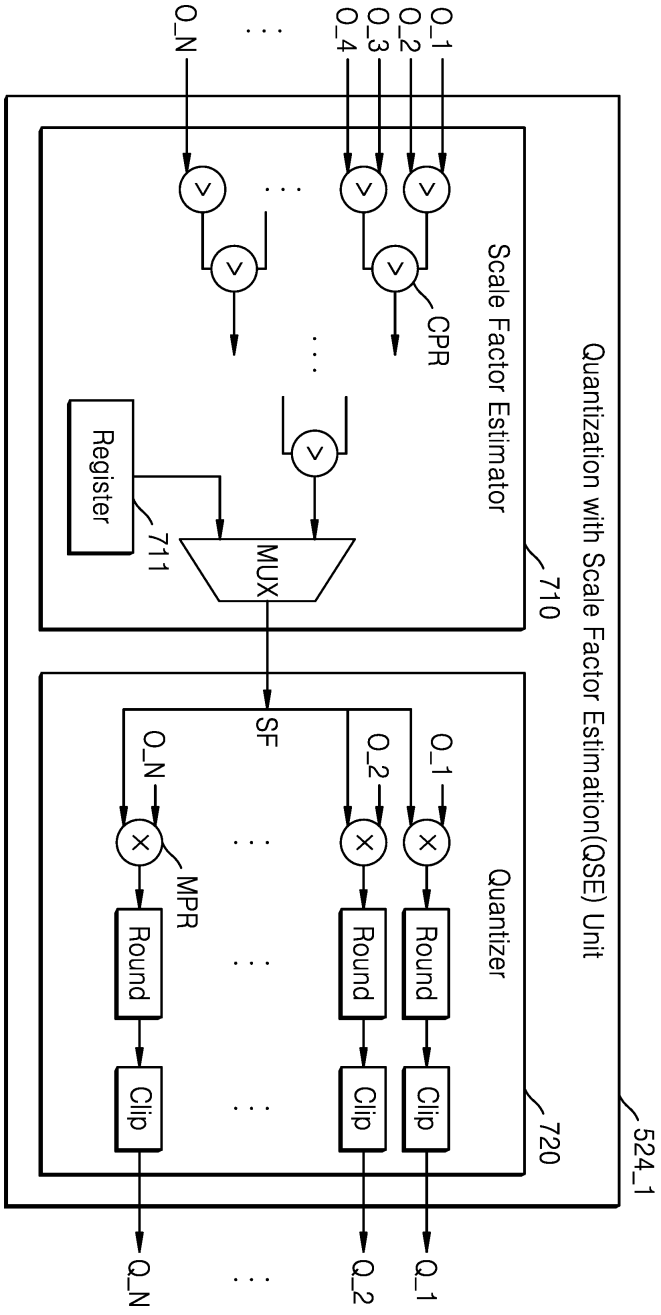
도면5


Memory
510
Vector Processing Unit
QSE Unit
524_1
524
Data Fetcher
522
526_1
PE
526_2
PE
526_N
PE
Inner Product Array
526
Processor
520
500

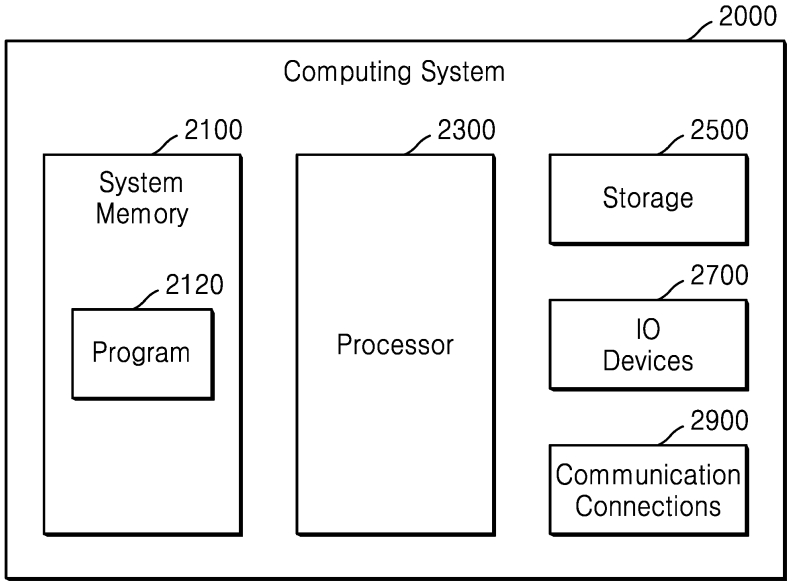
도면6


526_N
PE(Processing Element)
I1
W1
I2
W2
IX
WX
Bias
MPR
Adder Tree
O_N

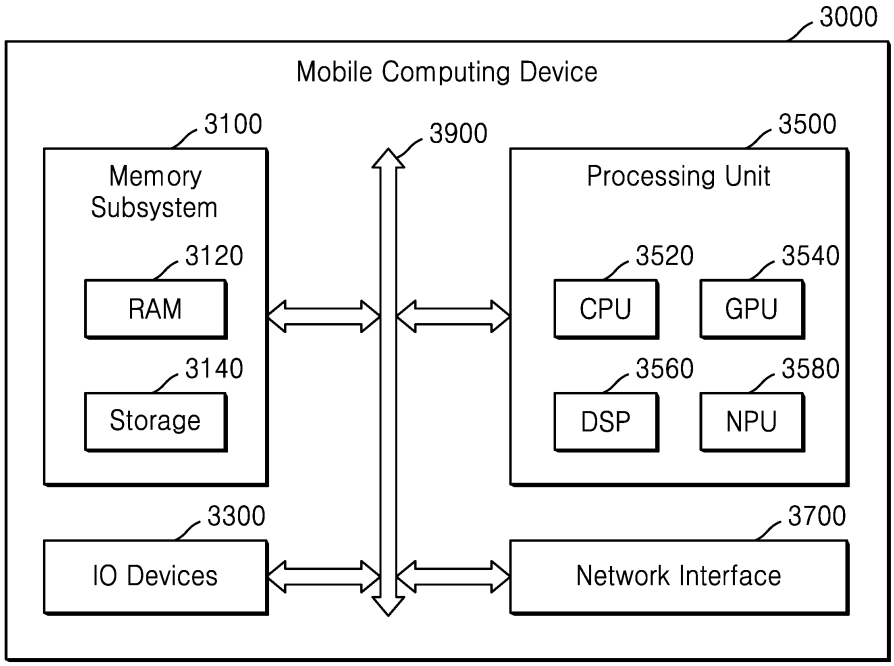
도면7


O_N
O_4
O_3
O_2
O_1
CPR
Register
711
MUX
Scale Factor Estimator
710
SF
MPR
Round
Clip
Quantizer
720
Quantization with Scale Factor Estimation(QSE) Unit
524_1
Q_N
Q_2
Q_1

도면8


2000
Computing System
2100
System Memory
2120
Program
2300
Processor
2500
Storage
2700
IO Devices
2900
Communication Connections

도면9


3000
Mobile Computing Device
3100
Memory Subsystem
3120
RAM
3140
Storage
3900
3500
Processing Unit
3520
CPU
3540
GPU
3560
DSP
3580
NPU
3300
IO Devices
3700
Network Interface