



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0139071
(43) 공개일자 2020년12월11일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01)

(52) CPC특허분류
G06N 3/04 (2013.01)

(21) 출원번호 10-2019-0104581

(22) 출원일자 2019년08월26일

심사청구일자 없음

(30) 우선권주장

62/856,212 2019년06월03일 미국(US)

(71) 출원인

삼성전자주식회사

경기도 수원시 영통구 삼성로 129 (매탄동)

울산과학기술원

울산광역시 울주군 언양읍 유니스트길 50

(72) 발명자

유형석

서울특별시 동작구 사당로23바길 9, 118동 1502호
(사당동, 동작삼성래미안아파트)

심현욱

울산광역시 울주군 언양읍 유니스트길 50

이중은

울산광역시 울주군 언양읍 유니스트길 50

(74) 대리인

리엔목특허법인

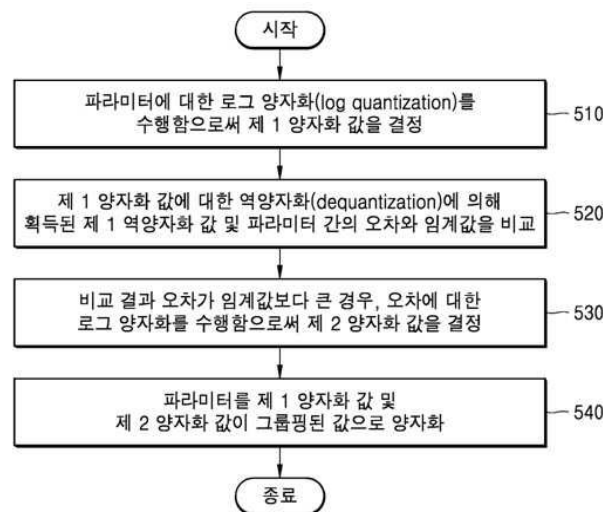
전체 청구항 수 : 총 19 항

(54) 발명의 명칭 뉴럴 네트워크에서 파라미터를 양자화하는 방법 및 장치

(57) 요약

뉴럴 네트워크에서 파라미터를 양자화하는 방법 및 장치는, 뉴럴 네트워크의 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정하고, 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 및 파라미터 간의 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정하고, 파라미터를 제 1 양자화 값 및 제 2 양자화 값이 그룹핑된 값으로 양자화한다.

대표도 - 도5



명세서

청구범위

청구항 1

뉴럴 네트워크에서 파라미터를 양자화하는 방법에 있어서,

뉴럴 네트워크의 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정하는 단계;

상기 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 및 상기 파라미터 간의 오차와 임계값을 비교하는 단계;

상기 비교 결과 상기 오차가 상기 임계값보다 큰 경우, 상기 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정하는 단계; 및

상기 파라미터를 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 그룹핑된 값으로 양자화하는 단계를 포함하는 방법.

청구항 2

제 1 항에 있어서,

상기 제 1 양자화 값을 결정하는 단계는,

상기 복수의 양자화 레벨들 중에서 상기 파라미터에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 1 양자화 값을 결정하고,

상기 제 2 양자화 값을 결정하는 단계는,

상기 복수의 양자화 레벨들 중에서 상기 오차에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 2 양자화 값을 결정하는 방법.

청구항 3

제 1 항에 있어서,

상기 제 2 양자화 값은,

상기 제 1 양자화 값을 나타내는 비트 수와 동일한 비트 수로 표현되는 방법.

청구항 4

제 1 항에 있어서,

상기 임계값은,

상기 뉴럴 네트워크의 인식률 및 상기 파라미터의 양자화에 따른 데이터의 크기 간의 트레이드-오프(trade off) 관계에 기초하여 결정되는 방법.

청구항 5

제 1 항에 있어서,

상기 양자화하는 단계는,

상기 제 1 양자화 값 및 상기 제 2 양자화 값 각각에 태그 비트(tag bit)를 추가하는 단계를 더 포함하는 방법.

청구항 6

제 5 항에 있어서,

상기 추가하는 단계는,

상기 제 1 양자화 값에 연속된 상기 제 2 양자화 값이 있음을 나타내는 제 1 태그 비트를 상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하고,

상기 제 2 양자화 값에 연속된 양자화 값이 없음을 나타내는 제 2 태그 비트를 상기 제 2 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하는 방법.

청구항 7

제 1 항에 있어서,

상기 양자화하는 단계는,

상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 및 상기 제 2 양자화 값을 구성하는 비트들 중 마지막 비트의 뒤 중 어느 하나에, 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 연속된 값을 나타내는 코드 값을 추가하는 단계를 더 포함하는 방법.

청구항 8

제 1 항에 있어서,

상기 그룹핑된 값을 역양자화하는 단계; 및

상기 그룹핑된 값을 역양자화한 값과 상기 입력 액티베이션 값들간의 컨볼루션(convolution) 연산을 수행하는 단계를 더 포함하는 방법.

청구항 9

제 8 항에 있어서,

상기 그룹핑된 값을 역 양자화하는 단계는,

상기 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 상기 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각을 산출하는 단계; 및

상기 제 1 역양자화 값 및 상기 제 2 역양자화 값을 더하여 상기 그룹핑된 값을 역양자화 값을 획득하는 단계를 더 포함하는 방법.

청구항 10

제 1항 내지 제 9항 중 어느 한 항의 방법을 실행하는 명령어들을 포함하는 하나 이상의 프로그램이 기록된 컴퓨터로 읽을 수 있는 기록매체.

청구항 11

뉴럴 네트워크를 실시(implement)하는 장치에 있어서,

메모리; 및

뉴럴 네트워크의 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 파라미터를 양자화하는 프로세서를 포함하고,

상기 프로세서는, 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정하고, 상기 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 및 상기 파라미터 간의 오차와 임계값을 비교하고, 상기 비교 결과 상기 오차가 상기 임계값보다 큰 경우, 상기 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정하고, 상기 파라미터를 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 그룹핑된 값으로 양자화하는 장치.

청구항 12

제 11 항에 있어서,

상기 프로세서는,

상기 복수의 양자화 레벨들 중에서 상기 파라미터에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 1 양자화 값을 결정하고,

상기 복수의 양자화 레벨들 중에서 상기 오차에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 2 양자화 값을 결정하는 장치.

청구항 13

제 11 항에 있어서,

상기 제 2 양자화 값은,

상기 제 1 양자화 값을 나타내는 비트 수와 동일한 비트 수로 표현되는 장치.

청구항 14

제 11 항에 있어서,

상기 임계값은,

상기 뉴럴 네트워크의 인식률 및 상기 파라미터의 양자화에 따른 데이터의 크기 간의 트레이드-오프(trade off) 관계에 기초하여 결정되는 장치.

청구항 15

제 11 항에 있어서,

상기 프로세서는,

상기 제 1 양자화 값 및 상기 제 2 양자화 값 각각에 태그 비트(tag bit)를 추가하는 장치.

청구항 16

제 15 항에 있어서,

상기 프로세서는,

상기 제 1 양자화 값에 연속된 상기 제 2 양자화 값이 있음을 나타내는 제 1 태그 비트를 상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하고,

상기 제 2 양자화 값에 연속된 양자화 값이 없음을 나타내는 제 2 태그 비트를 상기 제 2 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하는 장치.

청구항 17

제 11 항에 있어서,

상기 프로세서는,

상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 및 상기 제 2 양자화 값을 구성하는 비트들 중 마지막 비트의 뒤 중 어느 하나에, 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 연속된 값을 나타내는 코드 값을 추가하는 장치.

청구항 18

제 11 항에 있어서,

상기 프로세서는,

상기 그룹핑된 값을 역양자화하고,

상기 그룹핑된 값을 역양자화한 값과 상기 입력 액티베이션 값들간의 컨볼루션(convolution) 연산을 수행하는 장치.

청구항 19

제 18 항에 있어서,

상기 프로세서는,

상기 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 상기 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각을 산출하고,

상기 제 1 역양자화 값 및 상기 제 2 역양자화 값을 더하여 상기 그룹핑된 값을 역양자화 값을 획득하는 장치.

발명의 설명

기술 분야

[0001] 본 개시는 뉴럴 네트워크에서 파라미터를 양자화하는 방법 및 장치에 관한다.

배경 기술

[0002] 뉴럴 네트워크(neural network)는 생물학적 뇌를 모델링한 컴퓨터 과학적 아키텍처(computational architecture)를 참조한다. 최근 뉴럴 네트워크 기술이 발전함에 따라, 다양한 종류의 전자 시스템에서 뉴럴 네트워크 장치를 사용하여 입력 데이터를 분석하고 유효한 정보를 추출하는 연구가 활발히 진행되고 있다.

[0003] 뉴럴 네트워크 장치는 복잡한 입력 데이터에 대한 많은 양의 연산을 필요로 한다. 뉴럴 네트워크 장치가 입력을 실시간으로 분석하고, 정보를 추출하기 위해서 뉴럴 네트워크 연산을 효율적으로 처리할 수 있는 기술이 요구된다. 특히, 스마트폰과 같은, 저전력 고성능 임베디드 시스템은 제한된 리소스를 가지므로, 복잡한 입력 데이터를 처리하는데 필요한 연산량을 감소시키면서도 정확도 손실을 최소화할 수 있는 기술이 요구된다.

발명의 내용

해결하려는 과제

[0004] 뉴럴 네트워크에서 파라미터를 양자화하는 방법 및 장치를 제공하는데 있다. 또한, 상기 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공하는 데 있다.

[0005] 해결하고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 유추될 수 있다.

과제의 해결 수단

[0006] 일 측면에 따르면, 뉴럴 네트워크에서 파라미터를 양자화하는 방법은, 뉴럴 네트워크의 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 상기 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정하는 단계; 상기 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 상기 파라미터 간의 오차와 임계값을 비교하는 단계; 상기 비교 결과 상기 오차가 상기 임계값보다 큰 경우, 상기 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정하는 단계; 및 상기 파라미터를 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 그룹핑된 값으로 양자화하는 단계를 포함한다.

[0007] 또한, 상기 제 1 양자화 값을 결정하는 단계는, 상기 복수의 양자화 레벨들 중에서 상기 파라미터에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 1 양자화 값을 결정하고, 상기 제 2 양자화 값을 결정하는 단계는, 상기 복수의 양자화 레벨들 중에서 상기 오차에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 2 양자화 값을 결정한다.

[0008] 또한, 상기 제 2 양자화 값은, 상기 제 1 양자화 값을 나타내는 비트 수와 동일한 비트 수로 표현될 수 있다.

[0009] 또한, 상기 임계값은, 상기 뉴럴 네트워크의 인식률 및 상기 파라미터의 양자화에 따른 데이터의 크기 간의 트레이드-오프(trade off) 관계에 기초하여 결정될 수 있다.

[0010] 또한, 상기 양자화하는 단계는, 상기 제 1 양자화 값 및 상기 제 2 양자화 값 각각에 태그 비트(tag bit)를 추

가하는 단계를 더 포함할 수 있다.

- [0011] 또한, 상기 추가하는 단계는, 상기 제 1 양자화 값에 연속된 상기 제 2 양자화 값이 있음을 나타내는 제 1 태그 비트를 상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하고, 상기 제 2 양자화 값에 연속된 양자화 값이 없음을 나타내는 제 2 태그 비트를 상기 제 2 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가할 수 있다.
- [0012] 또한, 상기 양자화하는 단계는, 상기 제 1 양자화 값을 구성하는 비트들 중 첫 번째 비트의 앞 및 상기 제 2 양자화 값을 나타내는 비트들 중 마지막 비트의 뒤 중 어느 하나에, 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 연속된 값을 나타내는 코드 값을 추가하는 단계를 더 포함할 수 있다.
- [0013] 또한, 뉴럴 네트워크의 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 파라미터를 양자화하는 방법은, 상기 그룹핑된 값을 역양자화하는 단계; 및 상기 그룹핑된 값을 역양자화한 값과 상기 입력 액티베이션 값들간의 컨볼루션(convolution) 연산을 수행하는 단계를 더 포함할 수 있다.
- [0014] 또한, 상기 그룹핑된 값을 역 양자화하는 단계는, 상기 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 상기 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각을 산출하는 단계; 및 상기 제 1 역양자화 값 및 상기 제 2 역양자화 값을 더하여 상기 그룹핑된 값을 역양자화 값을 획득하는 단계를 더 포함할 수 있다.
- [0015] 다른 측면에 따르면, 뉴럴 네트워크를 실시(implement)하는 장치에 있어서, 메모리; 및 뉴럴 네트워크의 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 파라미터를 양자화하는 프로세서를 포함하고, 상기 프로세서는, 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정하고, 상기 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 및 상기 파라미터 간의 오차와 임계값을 비교하고, 상기 오차가 상기 임계값보다 큰 경우, 상기 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정하고, 상기 파라미터를 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 그룹핑된 값으로 양자화한다.
- [0016] 또한, 상기 프로세서는, 상기 복수의 양자화 레벨들 중에서 상기 파라미터에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 1 양자화 값을 결정하고, 상기 복수의 양자화 레벨들 중에서 상기 오차에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 상기 제 2 양자화 값을 결정할 수 있다.
- [0017] 또한, 상기 제 2 양자화 값은, 상기 제 1 양자화 값을 나타내는 비트 수와 동일한 비트 수로 표현될 수 있다.
- [0018] 또한, 상기 임계값은, 상기 뉴럴 네트워크의 인식률 및 상기 파라미터의 양자화에 따른 데이터의 크기 간의 트레이드-오프(trade off) 관계에 기초하여 결정될 수 있다.
- [0019] 또한, 상기 프로세서는, 상기 제 1 양자화 값 및 상기 제 2 양자화 값 각각에 태그 비트(tag bit)를 추가할 수 있다.
- [0020] 또한, 상기 프로세서는, 상기 제 1 양자화 값에 연속된 상기 제 2 양자화 값이 있음을 나타내는 제 1 태그 비트를 상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하고, 상기 제 2 양자화 값에 연속된 양자화 값이 없음을 나타내는 제 2 태그 비트를 상기 제 2 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가할 수 있다.
- [0021] 또한, 상기 프로세서는, 상기 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 및 상기 제 2 양자화 값을 구성하는 비트들 중 마지막 비트의 뒤 중 어느 하나에, 상기 제 1 양자화 값 및 상기 제 2 양자화 값이 연속된 값을 나타내는 코드 값을 추가할 수 있다.
- [0022] 또한, 상기 프로세서는, 상기 그룹핑된 값을 역양자화하고, 상기 그룹핑된 값을 역양자화한 값과 상기 입력 액티베이션 값들간의 컨볼루션(convolution) 연산을 수행할 수 있다.
- [0023] 또한, 상기 프로세서는, 상기 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 상기 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각을 산출하고, 상기 제 1 역양자화 값 및 상기 제 2 역양자화 값을 더하여 상기 그룹핑된 값을 역양자화 값을 획득할 수 있다.
- [0024] 또 다른 측면에 따르면, 컴퓨터로 읽을 수 있는 기록매체는 상술한 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 포함할 수 있다.

도면의 간단한 설명

- [0025] 도 1은 뉴럴 네트워크 아키텍처의 일 예를 설명하기 위한 도면이다.
- 도 2는 뉴럴 네트워크에서 수행되는 연산의 일 예를 설명하기 위한 도면이다.
- 도 3은 선형 양자화(linear quantization) 방식 및 로그 양자화(log quantization) 방식의 일 예를 나타내는 도면이다.
- 도 4는 선형 양자화 방식 및 로그 양자화 방식 각각에 대하여 양자화 에러의 일 예를 나타내는 도면이다.
- 도 5는 뉴럴 네트워크에서 파라미터를 양자화하는 방법의 일 예를 나타내는 흐름도이다.
- 도 6은 뉴럴 네트워크 장치가 제 1 양자화 값을 결정하는 일 예를 나타낸다.
- 도 7은 뉴럴 네트워크 장치가 제 1 역양자화 값과 파라미터의 차이가 임계값보다 큰 지 판단하는 일 예를 나타낸다.
- 도 8은 뉴럴 네트워크 장치가 제 2 양자화 값을 결정하는 일 예를 나타내는 도면이다.
- 도 9는 태그 비트(tag bit)를 이용하여 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내는 일 예를 나타내는 도면이다.
- 도 10은 코드 값을 이용하여 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내는 일 예를 나타내는 도면이다.
- 도 11은 뉴럴 네트워크 장치가 파라미터를 양자화하는 방법의 일 예를 설명하기 위한 흐름도이다.
- 도 12는 뉴럴 네트워크 장치가 파라미터를 역양자화하는 방법의 일 예를 설명하기 위한 흐름도이다.
- 도 13은 뉴럴 네트워크에서 파라미터를 양자화하는 방식에 따른 정확도를 비교한 도면이다.
- 도 14는 MAC(Multiplication and Accumulation) 연산을 수행하는 연산 장치의 일 예를 나타내는 도면이다.
- 도 15는 MAC(Multiplication and Accumulation) 연산을 수행하는 연산 장치의 다른 예를 나타내는 도면이다.
- 도 16은 일 실시예에 따른 전자 시스템을 나타내는 블록도이다.
- 도 17은 뉴럴 네트워크 장치의 구성의 일 예를 도시한 블록도이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 실시예들에서 사용되는 용어는 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 기술분야에 종사하는 기술자의 의도 또는 관례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 임의로 선정된 용어도 있으며, 이 경우 해당 실시예의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서, 명세서에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 실시예들의 전반에 걸친 내용을 토대로 정의되어야 한다.
- [0027] 명세서 전체에서, 어떤 부분이 다른 부분과 연결되어 있다고 할 때, 이는 직접적으로 연결되어 있는 경우뿐 아니라, 그 중간에 다른 구성요소를 사이에 두고 전기적으로 연결되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.
- [0028] 명세서에 기재된 "구성된다" 또는 "포함한다" 등의 용어는 명세서 상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다.
- [0029] 이하 첨부된 도면들을 참조하면서 실시예들을 상세히 설명하기로 한다. 그러나 실시 예는 여러 가지 상이한 형태로 구현될 수 있으며 이하에서 설명하는 예에 한정되지 않는다.
- [0030] 도 1은 뉴럴 네트워크 아키텍처의 일 예를 설명하기 위한 도면이다.
- [0031] 도 1에 도시된 뉴럴 네트워크(1)는 딥 뉴럴 네트워크(Deep Neural Network, 이하 'DNN'이라고 함) 아키텍처의 일 예일 수 있다. DNN은 컨볼루션 뉴럴 네트워크(Convolutional Neural Networks, CNN), 리커런트 뉴럴 네트워크

크(Recurrent Neural Networks, RNN), Deep Belief Networks, Restricted Boltzman Machines 등에 해당할 수 있으나 이에 한정되지 않는다. 뉴럴 네트워크(1)는 입력 레이어(Layer 1), 4개의 히든 레이어들(Layer 2, Layer 3, Layer 4, Layer 5) 및 출력 레이어(Layer 6)를 포함하는 DNN일 수 있다. 예를 들어, 뉴럴 네트워크(1)가 컨볼루션 뉴럴 네트워크(CNN)를 표현하는 것인 경우, Layer 1 내지 Layer 6는 컨볼루션 뉴럴 네트워크 중 일부의 레이어들에 해당될 수 있으며, 이는 컨볼루션 레이어, 풀링 레이어(pooling layer), 풀리 커넥티드(fully connected) 레이어 등에 해당될 수 있다.

[0032] 뉴럴 네트워크(1)에 포함된 레이어들 각각은 "뉴런(neuron)", "프로세싱 엘리먼트(Processing element: PE)", "유닛(unit)" 또는 이와 유사한 용어들로 알려진, 복수의 인공 노드(artificial node)들을 포함할 수 있다. 예를 들어, 도 1에 도시된 바와 같이, Layer 1은 5개의 노드들, Layer 2는 7개의 노드들을 포함할 수 있다. 다만, 이는 예시에 불과할 뿐 뉴럴 네트워크(1)에 포함된 레이어들 각각은 다양한 개수의 노드들을 포함할 수 있다.

[0033] 뉴럴 네트워크(1)에 포함된 레이어들 각각에 포함된 노드들은 서로 연결되어 데이터를 교환할 수 있다. 예를 들어, 하나의 노드는 다른 노드들로부터 데이터를 수신하여 연산할 수 있고, 연산 결과를 또 다른 노드들로 출력할 수 있다.

[0034] 한편, 뉴럴 네트워크(1)가 DNN 아키텍처로 구현된 경우 유효한 정보를 처리할 수 있는 많은 레이어들을 포함할 수 있다. 따라서, 뉴럴 네트워크(1)는 싱글 레이어를 포함하는 뉴럴 네트워크보다 복잡한 데이터 집합들을 처리할 수 있다. 한편, 도 1에는 뉴럴 네트워크(1)가 6개의 레이어들을 포함하는 것으로 도시되어 있으나, 이는 예시에 불과할 뿐 뉴럴 네트워크(1)는 더 적거나 많은 레이어들을 포함할 수 있다. 즉, 뉴럴 네트워크(1)는 도 1에 도시된 것과는 다른, 다양한 구조의 레이어들을 포함할 수 있다.

[0035] 도 2는 뉴럴네트워크에서 수행되는 연산의 일 예를 설명하기 위한 도면이다.

[0036] 도 2를 참조하면, 뉴럴 네트워크(2)는 입력 레이어, 히든 레이어들 및 출력 레이어를 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.

[0037] 뉴럴 네트워크(2)는 앞서 설명된 바와 같이, 2개 이상의 히든 레이어들을 포함하는 DNN 또는 n-계층 뉴럴 네트워크일 수 있다. 예를 들어, 도 2에 도시된 바와 같이, 뉴럴 네트워크(2)는 입력 레이어(Layer 1), 2개의 히든 레이어들(Layer 2 및 Layer 3) 및 출력 레이어(Layer 4)를 포함하는 DNN일 수 있다. 뉴럴 네트워크(2)가 DNN 아키텍처로 구현된 경우 유효한 정보를 처리할 수 있는 보다 많은 레이어들을 포함하므로, 뉴럴 네트워크(2)는 싱글 레이어를 갖는 뉴럴 네트워크보다 복잡한 데이터 집합들을 처리할 수 있다. 한편, 뉴럴 네트워크(2)는 4개의 레이어들을 포함하는 것으로 도시되어 있으나, 이는 예시에 불과할 뿐 뉴럴 네트워크(2)는 더 적거나 많은 레이어들을 포함하거나, 더 적거나 많은 채널들을 포함할 수 있다. 즉, 뉴럴 네트워크(2)는 도 2에 도시된 것과는 다른, 다양한 구조의 레이어들을 포함할 수 있다.

[0038] 뉴럴 네트워크(2)에 포함된 레이어들 각각은 복수의 채널들을 포함할 수 있다. 채널은 뉴런(neuron), 프로세싱 엘리먼트(Processing element, PE), 유닛(unit) 또는 이와 유사한 용어들로 알려진, 복수의 인공 노드(artificial node)들에 해당될 수 있다. 예를 들어, 도 2에 도시된 바와 같이, Layer 1은 2개의 채널들(노드들), Layer 2 및 Layer 3 각각은 3개의 채널들을 포함할 수 있다. 다만, 이는 예시에 불과할 뿐 뉴럴 네트워크(2)에 포함된 레이어들 각각은 다양한 개수의 채널들(노드들)을 포함할 수 있다.

[0039] 뉴럴 네트워크(2)의 레이어들 각각에 포함된 채널들은 서로 연결되어 데이터를 처리할 수 있다. 예를 들어, 하나의 채널은 다른 채널들로부터 데이터를 수신하여 연산할 수 있고, 연산 결과를 또 다른 채널들로 출력할 수 있다.

[0040] 채널들 각각의 입력 및 출력 각각은 입력 액티베이션 및 출력 액티베이션이라고 지칭될 수 있다. 즉, 액티베이션은 한 채널의 출력입과 동시에, 다음 레이어에 포함된 채널들의 입력에 해당되는 파라미터일 수 있다. 한편, 채널들 각각은 이전 레이어에 포함된 채널들로부터 수신된 액티베이션들 및 웨이트 및 바이어스에 기초하여 자신의 액티베이션을 결정할 수 있다. 웨이트는 각 채널에서의 출력 액티베이션을 계산하기 위해 이용되는 파라미터로서, 채널들 간의 연결관계에 할당되는 값일 수 있다.

[0041] 채널들 각각은 입력을 수신하여 출력 액티베이션을 출력하는 연산 유닛(computational unit) 또는 프로세싱 엘리먼트(processing element)에 의해 처리될 수 있고, 채널들 각각의 입력-출력은 매핑될 수 있다. 예를 들어, I_1

는 액티베이션 함수(activation function)이고, w_{jk}^i 는 (i-1) 번째 레이어에 포함된 k 번째 채널로부터 i 번째 레이어에 포함된 j 번째 채널로의 웨이트며, b_j^i 는 i 번째 레이어에 포함된 j 번째 채널의 바이어스(bias)이고, a_j^i 는 i 번째 레이어의 j 번째 채널의 액티베이션이라고 할 때, 액티베이션 a_j^i 는 다음과 같은 수학적 식 1을 이용하여 계산될 수 있다.

수학적 식 1

$$a_j^i = \sigma\left(\sum_k (w_{jk}^i \times a_k^{i-1}) + b_j^i\right)$$

[0042]

[0043] 도 2에 도시된 바와 같이, 2번째 레이어(Layer 2)의 첫 번째 채널(CH 1)의 액티베이션은 a_1^2 로 표현될 수 있다. 또한, a_1^2 은 수학적 식 1에 따라 $a_1^2 = \sigma(w_{1,1}^2 \times a_1^1 + w_{1,2}^2 \times a_2^1 + b_1^2)$ 의 값을 가질 수 있다. 다만, 앞서 설명한 수학적 식 1은 뉴럴 네트워크(2)에서 데이터를 처리하기 위해 이용되는 액티베이션 및 웨이트 및 바이어스를 설명하기 위한 예시일 뿐, 이에 제한되지 않는다. 액티베이션은 이전 레이어로부터 수신된 액티베이션들의 가중치 합(weighted sum)을 sigmoid 함수나 Rectified Linear Unit (ReLU) 함수 등의 액티베이션 함수에 통과시킴으로써 획득된 값일 수 있다.

[0044] 일반적으로 레이어들의 입력으로 부동 소수점 포맷 및 고정 소수점 포맷의 파라미터들이 이용될 수 있다. 또한 웨이트와 바이어스 값으로 부동 소수점 포맷 및 고정 소수점 포맷의 파라미터들이 이용될 수 있다.

[0045] 앞서 설명한 것과 같이, 뉴럴 네트워크(2)에서는 수많은 데이터 집합들이 상호 연결된 복수의 채널들 간에 교환되고, 레이어를 지나면서 수많은 연산 과정을 거친다. 따라서, 복잡한 입력 데이터를 처리하는데 필요한 연산량을 감소시키면서도 정확도 손실을 최소화할 수 있는 기술이 요구된다.

[0046] 도 3은 선형 양자화(linear quantization) 방식 및 로그 양자화(log quantization) 방식의 일 예를 나타내는 도면이다.

[0047] 뉴럴 네트워크에서 발생하는 많은 연산량을 줄이기 위해 뉴럴 네트워크가 포함하는 파라미터들의 다양한 양자화 방식에 대한 연구가 진행되었다. 예를 들어, 웨이트 파라미터들을 원래의 비트 길이 보다 작은 길이의 비트들로 양자화함으로써, 정확도가 다소 감소되더라도 웨이트 파라미터들의 처리에 필요한 연산량을 줄일 수 있다. 이러한 양자화 방식으로는 선형 양자화(linear quantization) 방식 및 로그 양자화(log quantization) 방식 등이 있다.

[0048] 선형 양자화 방식은 양자화의 대상이 되는 뉴럴 네트워크의 파라미터들에 일정한 간격의 양자화 레벨들을 할당하는 양자화 방식이다. 반면, 로그 양자화 방식은 로그 데이터 표현(logarithm data representation)을 기반으로 양자화 레벨들을 할당하므로, 큰 값들 근처의 파라미터들에 상대적으로 적은 양자화 레벨을 할당하고, 0 근처의 파라미터들에 상대적으로 많은 양자화 레벨들을 할당하는 양자화 방식이다.

[0049] 예를 들어, 도 3을 참조하면, -1 내지 1 구간 내에 존재하는 파라미터의 실제값(real value) x는 양자화된 정수값 q로 양자화될 수 있다. q는 -M 부터 M-1 사이의 정수값일 수 있으며, M은 예를 들어 2N-1일 수 있다.

[0050] 선형 양자화 방식에 의하면, -1 내지 $-(1-1/M)$ 구간 내의 x는 정수값 -M으로 양자화될 수 있고, $-(1-1/M)$ 내지 $-(1-2/M)$ 구간 내의 x는 정수값 $-(M-1)$ 으로 양자화될 수 있다. 마찬가지로, $1-1/M$ 내지 1 구간 내의 x는 정수값 M-1로 양자화될 수 있다.

[0051] 한편, 로그 양자화 방식에 의하면, -2^{-M} 내지 $-2^{-(M-1)}$ 구간 내의 x는 정수값 -M으로 양자화될 수 있고, $-2^{-(M-1)}$ 내지 $-2^{-(M-2)}$ 구간 내의 x는 정수값 $-(M-1)$ 으로 양자화될 수 있다. 마찬가지로, $2^{-(M-1)}$ 내지 2^{-M} 구간 내의 x는 정수값 M-1로 양자화될 수 있다.

[0052] 도 4는 선형 양자화 방식 및 로그 양자화 방식 각각에 대하여 양자화 오차의 일 예를 나타내는 도면이다.

[0053] 양자화 오차는, 파라미터의 실제값 x 및 양자화된 정수값 q간의 차이를 나타낼 수 있다. 도 4는 파라미터의 실제값 x에 따른 양자화 오차값을 나타내는 그래프이다.

- [0054] 파라미터의 실제값 x 가 0에 근접하는 값인 경우에, 로그 양자화 방식에 의한 양자화 오차값은 선형 양자화 방식에 의한 양자화 오차값보다 작거나 같을 수 있다. 그러나, 파라미터의 실제값 x 가 1에 가까워짐에 따라, 로그 양자화 방식에 의한 양자화 오차값은 선형 양자화 방식에 의한 양자화 오차값보다 커질 수 있다. 한편, 양자화 오차값이 커짐에 따라 뉴럴 네트워크의 인식률이 낮아질 수 있다. 따라서 파라미터의 실제값 x 가 1에 가까워지는 경우에 양자화 오차값이 커지는 현상을 해결하기 위한 양자화 방식이 요구된다.
- [0055] 이에 기존의 로그 양자화 방식에서는 파라미터를 하나의 양자화 값으로 표현하는데 반해, 본 실시예는 양자화 오차값이 임계값 이상인 경우에는 두 개 이상의 양자화 값으로 표현하는 로그 양자화 방식을 제한한다. 또한, 본 실시예는 제한하는 양자화 방식을 이용하여 뉴럴 네트워크 내 파라미터들을 양자화하여 저장하고, 이를 다시 역양자화하여 연산을 수행하는 등의 프로세싱 기능들을 갖는 뉴럴 네트워크의 생성을 목적으로 할 수 있다. 이하 도 5 내지 도 17을 참조하여 상세히 설명한다.
- [0056] 도 5는 뉴럴 네트워크에서 파라미터를 양자화하는 방법의 일 예를 나타내는 흐름도이다.
- [0057] 도 5를 참조하면, 뉴럴 네트워크에서 파라미터를 양자화하는 방법은 도 17에 도시된 뉴럴 네트워크 장치에서 시계열적으로 처리되는 단계들로 구성된다. 따라서, 이하에서 생략된 내용이라 하더라도 도 17에 도시된 뉴럴 네트워크 장치에 관하여 후술할 내용은 도 5의 방법에도 적용됨을 알 수 있다.
- [0058] 510 단계에서, 뉴럴 네트워크 장치는 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정할 수 있다.
- [0059] 파라미터는 뉴럴 네트워크의 레이어에서 처리되는 액티베이션 값들 및 레이어에서 처리되는 웨이트 값들 중 적어도 하나를 포함할 수 있으나, 이에 제한되는 것은 아니다.
- [0060] 도 6은 뉴럴 네트워크 장치가 제 1 양자화 값을 결정하는 일 예를 나타낸다.
- [0061] 도 6을 참조하면, 뉴럴 네트워크 장치는 밑이 2인 로그 양자화에 기초하여 제 1 양자화 값을 결정하고 있으나, 이에 제한되는 것은 아니며 밑은 2 이외에 다양한 값에 해당할 수 있다. 뉴럴 네트워크 장치는 복수의 양자화 레벨들 중에서 파라미터에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 제 1 양자화 값을 결정할 수 있다.
- [0062] 예를 들어 파라미터의 실제값이 $0.5(=2^{-1})$ 인 경우에, 뉴럴 네트워크 장치는 2의 지수에 해당하는 -1의 절대값에 해당하는 1을 이진법으로 표현하고 이를 제 1 양자화 값으로 결정할 수 있다. 뉴럴 네트워크 장치는 제 1 양자화 값을 5 비트(bit)로 나타낼 수 있으며, 제 1 양자화 값은 00001이 된다. 다만, 이에 제한되는 것은 아니며 제 1 양자화 값은 다양한 비트 수로 표현될 수 있다.
- [0063] 또는 파라미터의 실제값이 $-0.25(=-2^{-2})$ 인 경우에, 뉴럴 네트워크 장치는 2의 지수에 해당하는 -2의 절대값에 해당하는 2를 이진법으로 표현하고 이를 제 1 양자화 값으로 결정할 수 있다. 이 때, 파라미터의 실제값이 음수이므로 뉴럴 네트워크 장치는 2의 보수법을 이용하여, 제 1 양자화 값을 11110으로 결정할 수 있다.
- [0064] 또는 파라미터의 실제값이 0.4375인 경우에, 뉴럴 네트워크 장치는 0.4375에 근접한 $0.5(=2^{-1})$ 에 대응되는 00001을 제 1 양자화 값으로 결정할 수 있다. 다만, 이에 제한되는 것은 아니며 뉴럴 네트워크의 파라미터에 할당되는 양자화 레벨들의 간격에 따라, 제 1 양자화 값은 다양하게 결정될 수 있다. 예를 들어, 뉴럴 네트워크의 파라미터에 할당되는 복수의 양자화 레벨들에 $0.5(=2^{-1})$ 에 해당하는 양자화 레벨이 포함되지 않을 수 있다. 이러한 경우, 할당되는 복수의 양자화 레벨들 중 0.4375에 가장 근접하는 양자화 레벨에 대응되는 값인 $0.25(=2^{-2})$ 에 대한 로그 양자화를 수행함으로써 00010을 제 1 양자화 값으로 결정할 수 있다.
- [0065] 다시 도 5로 돌아오면, 520 단계에서, 뉴럴 네트워크 장치는 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 및 파라미터 간의 오차와 임계값을 비교할 수 있다.
- [0066] 뉴럴 네트워크 장치는 제 1 양자화 값을 하기 수학적 2에 의해 역양자화(dequantization, inverse quantization)하여 제 1 역양자화 값을 획득할 수 있다.

수학식 2

$$\tilde{x} = \begin{cases} \text{sign}(q)2^{-|q|} \\ 0 \end{cases}$$

[0067]

[0068] 상기 수학식 2에서 q 는 제 1 양자화 값에 해당할 수 있고, $\tilde{x}$ 는 제 1 양자화 값을 역양자화한 제 1 역양자화 값에 해당할 수 있다.

[0069] 제 1 역양자화 값과 파라미터 간의 오차가 임계값보다 작은 경우에는, 파라미터를 제 1 양자화 값으로 양자화할 수 있다. 그러나, 오차가 임계값보다 큰 경우에는 뉴럴 네트워크의 인식률이 저하되는 문제가 발생할 수 있으므로, 파라미터를 제 1 양자화 값을 포함하는 두 개 이상의 양자화 값으로 표현할 수 있다.

[0070] 도 7은 뉴럴 네트워크 장치가 제 1 역양자화 값과 파라미터의 차이가 임계값보다 큰 지 판단하는 일 예를 나타낸다.

[0071] 도 6에서 상술한 바와 같이, 파라미터의 실제값이 0.5(=2-1)인 경우에 제 1 양자화 값은 00001이 된다. 수학식 1에 기초하여, 제 1 양자화 값인 00001을 역양자화하여 제 1 역양자화 값 0.5를 획득할 수 있다. 파라미터의 실제값인 0.5와 제 1 역양자화 값인 0.5의 차이는 0이다. 따라서, 파라미터의 실제값과 제 1 역양자화 값의 차이가 임계값인 1/128 보다 작으므로, 파라미터는 최종적으로 00001로 양자화될 수 있다.

[0072] 마찬가지로, 파라미터의 실제값이 -0.25(=2-2)인 경우에도 파라미터의 실제값과 11110을 역양자화한 제 1 역양자화 값의 차이는 0이므로, 파라미터는 최종적으로 11110으로 양자화될 수 있다.

[0073] 한편, 파라미터의 실제값이 0.4375인 경우에 제 1 양자화 값은 00001이 된다. 00001을 역양자화한 제 1 역양자화 값은 0.5이므로, 파라미터의 실제값인 0.4375와 제 1 역양자화 값인 0.5의 차이는 -0.0625 이다. 따라서, 파라미터는 제 1 양자화 값인 00001을 포함하는 두 개 이상의 양자화 값으로 표현될 수 있으며, 이는 도 8에서 후술하도록 한다.

[0074] 한편, 도 7에서 임계값은 1/128로 설정되어 있으나 이에 제한되는 것은 아니며 임계값은 다양한 값에 해당할 수 있다. 예를 들어, 임계값은 뉴럴 네트워크의 인식률 및 파라미터의 양자화에 따른 데이터의 크기에 기초하여 결정될 수 있다. 임계값이 작을수록 파라미터의 실제값을 높은 정확도로 양자화할 수 있으므로 뉴럴 네트워크의 인식률이 증가하나, 파라미터를 여러 개의 양자화 값으로 표현함에 따라 데이터의 크기가 증가할 수 있다. 따라서, 뉴럴 네트워크의 인식률 및 양자화 값에 해당하는 데이터의 크기 간의 트레이드-오프(trade off) 관계를 고려하여 임계값을 결정할 수 있다.

[0075] 다시 도 5로 돌아오면, 530 단계에서, 뉴럴 네트워크 장치는 비교 결과 오차가 임계값보다 큰 경우, 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정할 수 있다.

[0076] 도 8은 뉴럴 네트워크 장치가 제 2 양자화 값을 결정하는 일 예를 나타내는 도면이다.

[0077] 뉴럴 네트워크 장치는 제 1 양자화 값을 결정할 때와 마찬가지로, 복수의 양자화 레벨들 중에서 오차에 가장 근접하는 양자화 레벨에 대응되는 값에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정할 수 있다. 이때, 제 2 양자화 값은 제 1 양자화 값을 나타내는 비트 수와 동일한 비트 수로 표현될 수 있다. 예를 들어 제 1 양자화 값을 5 비트로 나타내는 경우, 뉴럴 네트워크 장치는 제 2 양자화 값 또한 5 비트로 표현할 수 있다. 다만, 이에 제한되는 것은 아니며 제 1 양자화 값과 제 2 양자화 값은 서로 다른 비트 수로 표현될 수도 있다.

[0078] 도 7에서 상술한 바와 같이, 파라미터의 실제값인 0.4375과 제 1 양자화 값 00001을 역양자화한 제 1 역양자화 값인 0.5간의 오차는 -0.0625이다. 뉴럴 네트워크 장치는 오차에 해당하는 -0.0625(-2-4)에 대해 로그 양자화를 수행할 수 있다. 이 때, 오차 값이 음수이므로 뉴럴 네트워크 장치는 2의 보수법을 이용하여, 제 2 양자화 값을 11110으로 결정할 수 있다.

[0079] 다시 도 5로 돌아오면, 540 단계에서, 뉴럴 네트워크 장치는 파라미터를 제 1 양자화 값 및 제 2 양자화 값이 그룹핑된 값으로 양자화할 수 있다.

[0080] 예를 들어, 그룹핑된 값은 제 1 양자화 값 다음에 제 2 양자화 값이 연속적으로 위치하여 하나의 그룹을 형성한 값에 해당할 수 있다. 도 8을 참조하면, 파라미터의 실제값인 0.4375인 경우, 뉴럴 네트워크 장치는 파라미터를

제 1 양자화 값에 해당하는 00001 및 제 2 양자화 값에 해당하는 11110이 그룹핑된 값으로 양자화할 수 있다.

[0081] 한편, 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내기 위해 추가적으로 코드화 과정이 이루어질 수 있으며 이는 이하 도 9 및 도 10에서 후술하도록 한다.

[0082] 도 9는 태그 비트(tag bit)를 이용하여 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내는 일 예를 나타내는 도면이다.

[0083] 뉴럴 네트워크 장치는 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내기 위해, 제 1 양자화 값 및 제 2 양자화 값 각각에 태그 비트를 추가할 수 있다.

[0084] 예를 들어, 뉴럴 네트워크 장치는 제 1 양자화 값에 연속된 제 2 양자화 값이 있음을 나타내는 제 1 태그 비트를 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가할 수 있다. 또한, 제 2 양자화 값에 연속된 양자화 값이 없음을 나타내는 제 2 태그 비트를 제 2 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가할 수 있다. 제 1 태그 비트가 1일 때 제 2 태그 비트는 0일 수 있으나 이에 제한되는 것은 아니며, 제 1 태그 비트가 0일 때 제 2 태그 비트는 1일 수 있다.

[0085] 도 9를 참조하면, 제 1 태그 비트 및 제 2 태그 비트 중 어느 하나는 양자화 값들 각각을 나타내는 비트들 중 첫 번째 비트의 앞에 추가될 수 있다. 이 때, 제 1 태그 비트는 1에 해당하고, 제 2 태그 비트는 0에 해당할 수 있다. 양자화 값 $q[n-1]$, $q[n]2$ 및 $q[n+1]$ 각각을 나타내는 비트들 중 첫 번째 비트의 앞에는 제 2 태그 비트에 해당하는 0이 추가되어 있으므로, $q[n-1]$, $q[n]2$ 및 $q[n+1]$ 각각에는 연속된 양자화 값이 없음을 알 수 있다. 한편, $q[n]1$ 을 나타내는 비트들 중 첫 번째 비트의 앞에는 제 1 태그 비트에 해당하는 1이 추가되어 있으므로, $q[n]1$ 에는 연속된 양자화 값이 있음을 알 수 있다. 따라서, $q[n]1$ 및 $q[n]2$ 은 연속적으로 위치하여 그룹핑된 값에 해당하고, 그룹핑된 값은 하나의 파라미터를 양자화한 값을 나타낼 수 있다.

[0086] 일 예로써, 제 1 양자화 값이 00001이고 제 2 양자화 값이 11100인 경우, 00001을 나타내는 비트들 중 첫 번째 비트의 앞에는 제 1 태그 비트에 해당하는 1을 추가하고, 11100을 나타내는 비트들 중 첫 번째 비트의 앞에는 제 2 태그 비트에 해당하는 0을 추가할 수 있다. 이를 통해, 제 1 양자화 값 00001 및 제 2 양자화 값 11100는 그룹핑되어 하나의 파라미터를 양자화한 값을 나타낼 수 있다.

[0087] 도 10은 코드 값을 이용하여 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내는 일 예를 나타내는 도면이다.

[0088] 뉴럴 네트워크 장치는 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값을 나타내기 위해, 특정 코드 값을 추가할 수 있다. 코드 값은 제 1 양자화 값 또는 제 2 양자화 값과 동일한 수의 비트들로 구성될 수 있으나, 이에 제한되는 것은 아니다.

[0089] 예를 들어, 뉴럴 네트워크 장치는 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 및 제 2 양자화 값을 나타내는 비트들 중 마지막 비트의 뒤 중 어느 하나에, 제 1 양자화 값 및 상기 제 2 양자화 값이 연속된 값을 나타내는 코드 값을 추가할 수 있다.

[0090] 도 10을 참조하면, 코드 값은 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞에 추가될 수 있다. 양자화 값 $q[n-1]$, $q[n]2$ 및 $q[n+1]$ 각각을 나타내는 비트들 중 첫 번째 비트의 앞에는 코드 값이 추가되지 않았으므로, $q[n-1]$, $q[n]2$ 및 $q[n+1]$ 각각에는 연속된 양자화 값이 없음을 알 수 있다. 한편, $q[n]1$ 을 나타내는 비트들 중 첫 번째 비트의 앞에는 코드 값이 추가되어 있으므로, $q[n]1$ 에는 연속된 양자화 값이 있음을 알 수 있다. 따라서, $q[n]1$ 및 $q[n]2$ 은 연속적으로 위치하여 그룹핑된 값에 해당하고, 그룹핑된 값은 하나의 파라미터를 양자화한 값을 나타낼 수 있다.

[0091] 일 예로써, 제 1 양자화 값이 00001이고 제 2 양자화 값이 11100인 경우, 00001을 나타내는 비트들 중 첫 번째 비트의 앞에 코드값(-M)을 추가할 수 있다. 이를 통해, 제 1 양자화 값 00001 및 제 2 양자화 값 11100는 그룹핑되어 하나의 파라미터를 양자화한 값을 나타낼 수 있다.

[0092] 도 11은 뉴럴 네트워크 장치가 파라미터를 양자화하는 방법의 일 예를 설명하기 위한 흐름도이다.

[0093] 1110 단계에서, 뉴럴 네트워크 장치는 초기 조건으로써 $k=1$, $\delta=x$, $q1=0$ 으로 설정할 수 있다. x 는 뉴럴 네트워크 장치가 양자화하고자 하는 파라미터의 실제값에 해당할 수 있다.

[0094] 1120 단계에서, 뉴럴 네트워크 장치는 k 에 해당하는 값이 L 보다 큰 지 판단할 수 있다. L 은 파라미터를 복수의 양자화 값들이 그룹핑된 값으로 양자화하는 경우, 최대도 그룹핑될 수 있는 양자화 값들의 개수에 해당할 수 있다.

다. k 에 해당하는 값이 L 보다 크지 않은 경우에는 1130 단계로 진행한다. k 에 해당하는 값이 L 보다 큰 경우에는 1180 단계로 진행한다.

[0095] 1130 단계에서, 뉴럴 네트워크 장치는 δ 에 해당하는 값이 0인지 여부를 판단할 수 있다. δ 에 해당하는 값이 0인 경우에는 1140 단계로 진행한다. δ 에 해당하는 값이 0이 아닌 경우에는 1180 단계로 진행한다.

[0096] 1140 단계에서, 뉴럴 네트워크 장치는 로그 양자화를 수행함으로써 양자화 값 q_k 를 결정할 수 있다. q_k 는 하기 수학식 3에 의해 결정될 수 있다.

수학식 3

[0097]
$$q_k = \text{sign}(\delta) \text{round}(-\log_2(\text{abs}(\delta)))$$

[0098] 상기 수학식 3에서 $\text{sign}(\delta)$ 은 δ 의 부호에 따라 1 또는 -1로 결정될 수 있고, $\text{abs}(\delta)$ 은 δ 의 절대값에 해당하며, $\text{round}(-\log_2(\text{abs}(\delta)))$ 은 $-\log_2(\text{abs}(\delta))$ 에 해당하는 값을 소수점 첫째 자리에서 반올림한 값에 해당할 수 있다.

[0099] 1150 단계에서, 뉴럴 네트워크 장치는 파라미터의 실제값 및 양자화 값 q_1 내지 q_k 가 그룹핑된 값을 역양자화한 값 간의 오차를 산출할 수 있다. 오차는 하기 수학식 4에 의해 결정될 수 있다.

수학식 4

[0100]
$$\delta = \text{abs}(x - \text{dequantize}(q_1 \dots q_k))$$

[0101] 상기 수학식 4에서 $\text{dequantize}(q_1 \dots q_k)$ 는 양자화 값 q_1 내지 q_k 가 그룹핑된 값을 역양자화한 값에 해당할 수 있고, $\text{abs}(x - \text{dequantize}(q_1 \dots q_k))$ 은 $x - \text{dequantize}(q_1 \dots q_k)$ 의 절대값에 해당할 수 있다. 한편, 복수의 양자화 값들이 그룹핑된 값을 역양자화하는 것에 관해서는 도 12에서 후술한다.

[0102] 1160 단계에서, 뉴럴 네트워크 장치는 δ 에 해당하는 값이 임계값 0보다 큰 지 판단할 수 있다. δ 에 해당하는 값이 임계값 0보다 큰 경우에는 1170 단계로 진행한다. δ 에 해당하는 값이 임계값 0보다 크지 않은 경우에는 1180 단계로 진행한다.

[0103] 1170 단계에서, 뉴럴 네트워크 장치는 k 에 해당하는 값을 1만큼 증가시킬 수 있다.

[0104] 1180 단계에서, 뉴럴 네트워크 장치는 최종적으로 파라미터를 양자화 값 q_1 내지 q_k 가 그룹핑된 값으로 양자화할 수 있다.

[0105] 도 12는 뉴럴 네트워크 장치가 파라미터를 역양자화하는 방법의 일 예를 설명하기 위한 흐름도이다.

[0106] 1210 단계에서, 뉴럴 네트워크 장치는 초기 조건으로써 $k=1$, $\tilde{x} = 0$ 으로 설정할 수 있다. 이 때, 역양자화의 대상이 되는 값은 양자화 값 q_1 내지 q_n 이 그룹핑된 값일 수 있다. $\tilde{x}$ 는 그룹핑된 값을 역양자화한 값에 해당할 수 있다.

[0107] 1220 단계에서, 뉴럴 네트워크 장치는 q_k 에 해당하는 값이 0인지 여부를 판단할 수 있다. q_k 에 해당하는 값이 0이 아닌 경우에는 1230 단계로 진행한다. q_k 에 해당하는 값이 0인 경우에는 1240 단계로 진행한다.

[0108] 1230 단계에서, 뉴럴 네트워크 장치는 역양자화를 수행함으로써 역양자화 값 $\tilde{x}$ 를 결정할 수 있다. $\tilde{x}$ 는 하기 수학식 5에 의해 결정될 수 있다.

수학식 5

[0109] $\tilde{x} += \text{sign}(q_k) 2^{-|q_k|}$

[0110] 상기 수학식 5에서 $\text{sign}(\delta)$ 은 δ 의 부호에 따라 1 또는 -1로 결정될 수 있다. 양자화 값 q_1 내지 q_n 이 그룹핑된 값을 역양자화한 값 $\tilde{x}$ 는 양자화 값 q_1 내지 q_n 각각을 역양자화한 값을 누적함으로써 결정될 수 있다.

[0111] 1240 단계에서, 뉴럴 네트워크 장치는 k 에 해당하는 값이 n 인 지 여부를 판단할 수 있다. k 에 해당하는 값이 n 인 경우에는 1260 단계로 진행한다. k 에 해당하는 값이 n 이 아닌 경우에는 1250 단계로 진행한다.

[0112] 1250 단계에서, 뉴럴 네트워크 장치는 k 에 해당하는 값을 1만큼 증가시킬 수 있다.

[0113] 1260 단계에서, 뉴럴 네트워크 장치는 양자화 값 q_1 내지 q_n 이 그룹핑된 값을 역양자화한 값으로써 $\tilde{x}$ 를 출력할 수 있다. $\tilde{x}$ 는 양자화 값 q_1 내지 q_n 각각을 역양자화한 값을 누적한 값에 해당할 수 있다.

[0114] 도 13은 파라미터를 양자화하는 방식에 따른 뉴럴 네트워크의 정확도를 비교한 도면이다.

[0115] 이미지 분류의 정확도를 산출하기 위한 뉴럴 네트워크의 모델은 AlexNet, SqueezeNet 및 VGG-S를 포함할 수 있다. 도 13을 참조하면, 선형 양자화 방식, 로그 양자화 방식 및 본 실시예에 따라 양자화 에러값이 임계값 이상인 경우에는 두 개 이상의 양자화 값으로 표현하는 로그 양자화 방식 각각이 적용되는 경우 뉴럴 네트워크에서의 분류 정확도(%)가 표시되어 있다. 한편, float 32b는 파라미터를 양자화하지 않은 경우를 나타낸다.

[0116] 도 13에 따라, 기존의 로그 양자화 방식 대비 본 실시예에 따른 로그 양자화 방식을 적용하는 경우, 2% 내지 4% 정도 정확도가 향상됨을 알 수 있다.

[0117] 도 14는 MAC(Multiplication and Accumulation) 연산을 수행하는 연산 장치의 일 예를 나타내는 도면이다.

[0118] 뉴럴 네트워크 장치는 예를 들어 뉴럴 네트워크의 레이어에서 처리되는 웨이트 값들을 양자화하여 메모리에 저장할 수 있다. 이후, 뉴럴 네트워크 장치는 양자화된 웨이트 값들을 역양자화하여 입력 액티베이션 값들과 컨볼루션 연산을 수행할 수 있다.

[0119] 이 때, 웨이트 값이 제 1 양자화 값 및 제 2 양자화 값이 그룹핑된 값으로 양자화된 경우 뉴럴 네트워크 장치는 그룹핑된 값을 역양자화하고, 그룹핑된 값을 역양자화한 값과 입력 액티베이션 값 간의 컨볼루션 연산을 수행할 수 있다. 한편, 뉴럴 네트워크 장치는 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각을 산출하고, 제 1 역양자화 값 및 제 2 역양자화 값을 더함으로써 그룹핑된 값을 역양자화할 수 있다.

[0120] 이에 따라, 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각은 동일한 입력 액티베이션 값과 연산이 수행되어야 한다.

[0121] 도 14는 웨이트 값들이 태그 비트를 이용하여 양자화된 경우 MAC 연산을 수행하는 연산 장치이며, q 는 웨이트 값들이 양자화된 값을 나타내고 X 는 입력 액티베이션 값을 나타낸다. MAC 연산을 수행하는 연산 장치는 태그 비트가 1인 경우 이를 MUX(Multiplexer) 선택 신호로 사용하여 전 사이클(cycle)의 입력 액티베이션 값 X 가 계속 연산에 이용되도록 할 수 있다. 이 때, 양자화 값에 추가된 태그 비트가 1인 경우에는 해당 양자화 값에 연속적으로 위치하는 또 다른 양자화 값이 있음을 나타낼 수 있다. 태그 비트가 0인 경우에는, 연속적으로 위치하는 양자화 값이 없으므로 전 사이클의 입력 액티베이션 값 X 와는 다른 새로운 입력 액티베이션 값 X 가 입력되어 연산이 수행될 수 있다.

[0122] 도 15는 MAC(Multiplication and Accumulation) 연산을 수행하는 연산 장치의 다른 예를 나타내는 도면이다.

[0123] 도 15는 웨이트 값들이 코드값을 이용하여 양자화된 경우 MAC 연산을 수행하는 연산 장치이며, 마찬가지로 q 는 웨이트 값들이 양자화된 값을 나타내고 X 는 입력 액티베이션 값을 나타낸다. MAC 연산을 수행하는 연산 장치는 코드값이 입력된 경우 이를 MUX(Multiplexer) 선택 신호로 사용할 수 있다. 이에 따라, 코드값 이후에 연속적으로 입력되는 제 1 양자화 값 및 제 2 양자화 값에 대해서는 동일한 입력 액티베이션 값 X 가 계속 연산에 이용되도록 할 수 있다. 한편, 코드값 및 입력 액티베이션 값 X 간의 연산 결과는 결과값 Y 에 누적되지 않도록 할 수

있다.

- [0124] 도 16은 일 실시예에 따른 전자 시스템을 나타내는 블록도이다.
- [0125] 도 16을 참고하면, 전자 시스템(1600)은 뉴럴 네트워크를 기초로 입력 데이터를 실시간으로 분석하여 유효한 정보를 추출하고, 추출된 정보를 기초로 상황 판단을 하거나 또는 전자 시스템(1600)이 탑재되는 전자 디바이스의 구성들을 제어할 수 있다. 예컨대 전자 시스템(1600)은 드론(drone), 첨단 운전자 보조 시스템(Advanced Drivers Assistance System; ADAS) 등과 같은 로봇 장치, 스마트 TV, 스마트폰, 의료 디바이스, 모바일 디바이스, 영상 표시 디바이스, 계측 디바이스, IoT 디바이스 등에 적용될 수 있으며, 이 외에도 다양한 종류의 전자 디바이스들 중 적어도 하나에 탑재될 수 있다.
- [0126] 전자 시스템(1600)은 CPU(1610), RAM(1620), 뉴럴 네트워크 장치(1630), 메모리(1640), 센서 모듈(1650) 및 통신 모듈(1660)을 포함할 수 있다. 전자 시스템(1600)은 입출력 모듈, 보안 모듈, 전력 제어 장치 등을 더 포함할 수 있다. 전자 시스템(1600)의 하드웨어 구성들 중 일부는 적어도 하나의 반도체 칩에 탑재될 수 있다. 뉴럴 네트워크 장치(1630)는 앞서 설명된 뉴럴 네트워크 전용 하드웨어 가속기 자체 또는 이를 포함하는 장치일 수 있다.
- [0127] CPU(1610)는 전자 시스템(1600)의 전반적인 동작을 제어한다. CPU(1610)는 하나의 프로세서 코어(Single Core)를 포함하거나, 복수의 프로세서 코어들(Multi-Core)을 포함할 수 있다. CPU(1610)는 메모리(1640)에 저장된 프로그램들 및/또는 데이터를 처리 또는 실행할 수 있다. 일 실시예에 있어서, CPU(1610)는 메모리(1640)에 저장된 프로그램들을 실행함으로써, 뉴럴 네트워크 장치(1630)의 기능을 제어할 수 있다. CPU(1610)는 CPU, GPU, AP 등으로 구현될 수 있다.
- [0128] RAM(1620)은 프로그램들, 데이터, 또는 명령들(instructions)을 일시적으로 저장할 수 있다. 예컨대 메모리(1640)에 저장된 프로그램들 및/또는 데이터는 CPU(1610)의 제어 또는 부팅 코드에 따라 RAM(1620)에 일시적으로 저장될 수 있다. RAM(1620)은 DRAM(Dynamic RAM) 또는 SRAM(Static RAM) 등의 메모리로 구현될 수 있다.
- [0129] 뉴럴 네트워크 장치(1630)는 수신되는 입력 데이터를 기초로 뉴럴 네트워크의 연산을 수행하고, 수행 결과를 기초로 정보 신호를 생성할 수 있다. 뉴럴 네트워크는 Convolutional Neural Networks(CNN), Recurrent Neural Networks(RNN), Deep Belief Networks, Restricted Boltzman Machines 등을 포함할 수 있으나 이에 제한되지 않는다. 뉴럴 네트워크 장치(1630)는 앞서 설명된 고정 소수점 타입으로 양자화된 뉴럴 네트워크를 이용하여 처리를 수행하는 하드웨어로서, 앞서 설명된 뉴럴 네트워크 전용 하드웨어 가속기에 해당될 수 있다.
- [0130] 정보 신호는 음성 인식 신호, 사물 인식 신호, 영상 인식 신호, 생체 정보 인식 신호 등과 같은 다양한 종류의 인식 신호 중 하나를 포함할 수 있다. 예를 들어, 뉴럴 네트워크 장치(1630)는 비디오 스트림에 포함되는 프레임 데이터를 입력 데이터로서 수신하고, 프레임 데이터로부터 프레임 데이터가 나타내는 이미지에 포함된 사물에 대한 인식 신호를 생성할 수 있다. 그러나, 이에 제한되는 것은 아니며, 전자 시스템(1600)이 탑재된 전자 장치의 종류 또는 기능에 따라 뉴럴 네트워크 장치(1630)는 다양한 종류의 입력 데이터를 수신할 수 있고, 입력 데이터에 따른 인식 신호를 생성할 수 있다.
- [0131] 메모리(1640)는 데이터를 저장하기 위한 저장 장소로서, OS(Operating System), 각종 프로그램들, 및 각종 데이터를 저장할 수 있다. 실시예에 있어서, 메모리(1640)는 뉴럴 네트워크 장치(1630)의 연산 수행 과정에서 생성되는 중간 결과들, 예컨대 출력 피치 맵을 출력 피치 리스트 또는 출력 피치 매트릭스 형태로 저장할 수 있다. 실시예에 있어서, 메모리(1640)에는 압축된 출력 피치 맵이 저장될 수 있다. 또한, 메모리(1640)는 뉴럴 네트워크 장치(1630)에서 이용되는 양자화된 뉴럴 네트워크 데이터, 예컨대, 파라미터들, 웨이트 맵 또는 웨이트 리스트를 저장할 수 있다.
- [0132] 메모리(1640)는 DRAM일 수 있으나, 이에 한정되는 것은 아니다. 메모리(1640)는 휘발성 메모리 또는 불휘발성 메모리 중 적어도 하나를 포함할 수 있다. 불휘발성 메모리는 ROM, PROM, EPROM, EEPROM, 플래시 메모리, PRAM, MRAM, RRAM, FRAM 등을 포함한다. 휘발성 메모리는 DRAM, SRAM, SDRAM, PRAM, MRAM, RRAM, FeRAM 등을 포함한다. 실시예에 있어서, 메모리(1640)는 HDD, SSD, CF, SD, Micro-SD, Mini-SD, xD 또는 Memory Stick 중 적어도 하나를 포함할 수 있다.
- [0133] 센서 모듈(1650)은 전자 시스템(1600)이 탑재되는 전자 장치 주변의 정보를 수집할 수 있다. 센서 모듈(1650)은 전자 장치의 외부로부터 신호(예컨대 영상 신호, 음성 신호, 자기 신호, 생체 신호, 터치 신호 등)를 센싱 또는 수신하고, 센싱 또는 수신된 신호를 데이터로 변환할 수 있다. 이를 위해, 센서 모듈(1650)은 센싱 장치, 예컨대 마이크, 촬상 장치, 이미지 센서, 라이다(LIDAR; light detection and ranging) 센서, 초음파 센서, 적외선

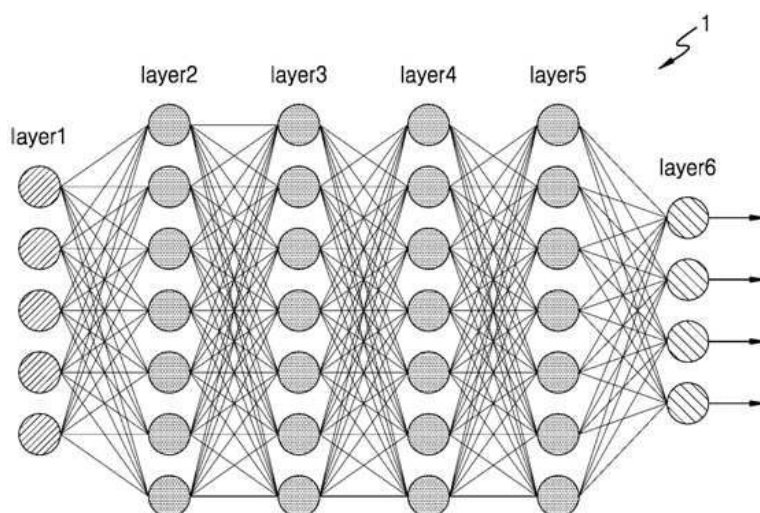
센서, 바이오 센서, 및 터치 센서 등 다양한 종류의 센싱 장치 중 적어도 하나를 포함할 수 있다.

- [0134] 센서 모듈(1650)은 변환된 데이터를 뉴럴 네트워크 장치(1630)에 입력 데이터로서 제공할 수 있다. 예를 들어, 센서 모듈(1650)은 이미지 센서를 포함할 수 있으며, 전자 장치의 외부 환경을 촬영하여 비디오 스트림을 생성하고, 비디오 스트림의 연속하는 데이터 프레임을 뉴럴 네트워크 장치(1630)에 입력 데이터로서 순서대로 제공할 수 있다. 그러나 이에 제한되는 것은 아니며 센서 모듈(1650)은 다양한 종류의 데이터를 뉴럴 네트워크 장치(1630)에 제공할 수 있다.
- [0135] 통신 모듈(1660)은 외부 장치와 통신할 수 있는 다양한 유선 또는 무선 인터페이스를 구비할 수 있다. 예컨대 통신 모듈(1660)은 유선 근거리통신망(Local Area Network; LAN), Wi-fi(Wireless Fidelity)와 같은 무선 근거리 통신망 (Wireless Local Area Network; WLAN), 블루투스(Bluetooth)와 같은 무선 개인 통신망(Wireless Personal Area Network; WPAN), 무선 USB (Wireless Universal Serial Bus), Zigbee, NFC (Near Field Communication), RFID (Radio-frequency identification), PLC(Power Line communication), 또는 3G (3rd Generation), 4G (4th Generation), LTE (Long Term Evolution) 등 이동 통신망(mobile cellular network)에 접속 가능한 통신 인터페이스 등을 포함할 수 있다.
- [0136] 이하, 도 17을 참조하여 뉴럴 네트워크 장치(1630)를 보다 상세히 설명하기로 한다.
- [0137] 도 17은 뉴럴 네트워크 장치의 구성의 일 예를 도시한 블록도이다.
- [0138] 뉴럴 네트워크 장치(1700)는 PC(personal computer), 서버 디바이스, 모바일 디바이스, 임베디드 디바이스 등의 다양한 종류의 디바이스들로 구현될 수 있다. 예를 들어, 뉴럴 네트워크 장치(1700)는 뉴럴 네트워크를 이용한 음성 인식, 영상 인식, 영상 분류 등을 수행하는 스마트폰, 태블릿 디바이스, AR(Augmented Reality) 디바이스, IoT(Internet of Things) 디바이스, 자율주행 자동차, 로봇틱스, 의료기기 등으로 구현될 수 있으나, 이에 제한되지 않는다. 나아가서, 뉴럴 네트워크 장치(1700)는 위와 같은 디바이스에 탑재되는 전용 하드웨어 가속기(HW accelerator)에 해당될 수 있고, 뉴럴 네트워크 장치(1700)는 뉴럴 네트워크 구동을 위한 전용 모듈인 NPU(neural processing unit), TPU(Tensor Processing Unit), Neural Engine 등과 같은 하드웨어 가속기일 수 있으나, 이에 제한되지 않는다.
- [0139] 도 17을 참조하면, 뉴럴 네트워크 장치(1700)는 프로세서(1710) 및 메모리(1730)를 포함한다. 도 17에 도시된 뉴럴 네트워크 장치(1700)에는 본 실시예들과 관련된 구성요소들만이 도시되어 있다. 따라서, 뉴럴 네트워크 장치(1700)에는 도 17에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있음은 당해 기술분야의 통상의 기술자에게 자명하다.
- [0140] 프로세서(1710)는 뉴럴 네트워크의 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 파라미터를 양자화할 수 있다.
- [0141] 먼저, 프로세서(1710)는 파라미터에 대한 로그 양자화(log quantization)를 수행함으로써 제 1 양자화 값을 결정할 수 있다.
- [0142] 또한, 프로세서(1710)는 제 1 양자화 값에 대한 역양자화(dequantization)에 의해 획득된 제 1 역양자화 값 및 파라미터 간의 오차와 임계값을 비교할 수 있다. 이 때, 임계값은, 뉴럴 네트워크의 인식을 및 파라미터의 양자화에 따른 데이터의 크기 간의 트레이드-오프 관계에 기초하여 결정될 수 있다.
- [0143] 프로세서(1710)는 비교 결과 오차가 임계값보다 큰 경우, 오차에 대한 로그 양자화를 수행함으로써 제 2 양자화 값을 결정할 수 있다. 한편, 오차가 임계값보다 작은 경우에는 파라미터를 제 1 양자화 값으로 양자화할 수 있다.
- [0144] 최종적으로, 프로세서(1710)는 파라미터를 제 1 양자화 값 및 제 2 양자화 값이 그룹핑된 값으로 양자화할 수 있다. 또한, 프로세서(1710)는 제 1 양자화 값 및 제 2 양자화 값이 하나의 파라미터를 양자화한 값임을 내기 위해, 제 1 양자화 값 및 제 2 양자화 값 각각에 태그 비트를 추가할 수 있다. 예를 들어, 프로세서(1710)는 제 1 양자화 값에 연속된 제 2 양자화 값이 있음을 나타내는 제 1 태그 비트를 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가하고, 제 2 양자화 값에 연속된 양자화 값이 없음을 나타내는 제 2 태그 비트를 제 2 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 또는 마지막 비트의 뒤에 추가할 수 있다. 또는, 프로세서(1710)는 제 1 양자화 값을 나타내는 비트들 중 첫 번째 비트의 앞 및 제 2 양자화 값을 구성하는 비트들 중 마지막 비트의 뒤 중 어느 하나에, 제 1 양자화 값 및 제 2 양자화 값이 연속된 값임을 나타내는 코드 값을 추가할 수도 있다.

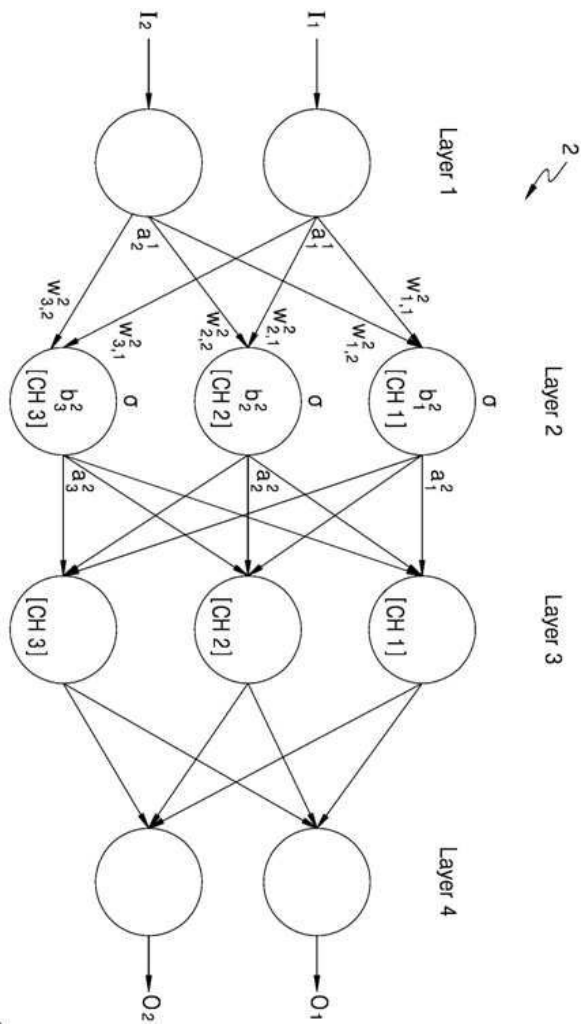
- [0145] 한편, 프로세서(1710)는 뉴럴 네트워크의 레이어에서 처리되는 웨이트 값들을 양자화한 경우에 제 1 양자화 값 및 제 2 양자화 값이 그룹핑된 값을 역양자화하고, 역양자화한 값과 입력 액티베이션 값들간의 컨볼루션(convolution) 연산을 수행할 수 있다. 예를 들어, 프로세서는 제 1 양자화 값을 역양자화한 값인 제 1 역양자화 값 및 제 2 양자화 값을 역양자화한 값인 제 2 역양자화 값 각각을 산출하고, 제 1 역양자화 값 및 제 2 역양자화 값을 더하여 그룹핑된 값을 역양자화 값을 획득할 수 있다.
- [0146] 프로세서(1710)는 뉴럴 네트워크 장치(1700) 내에 구비된 CPU(central processing unit), GPU(graphics processing unit), AP(application processor) 등으로 구현될 수 있으나, 이에 제한되지 않는다.
- [0147] 메모리(1730)는 뉴럴 네트워크 장치(1700)에서 처리된 데이터들 및 처리될 데이터들을 저장할 수 있다. 예를 들어, 메모리(1730)는 레이어에서 처리되는 입력 액티베이션 값들 및 웨이트 값들 중 적어도 하나를 포함하는 파라미터, 제 1 양자화 값 및 제 2 양자화 값 등을 저장할 수 있다. 메모리(1730)는 DRAM(dynamic random access memory), SRAM(static random access memory) 등과 같은 RAM(random access memory), ROM(read-only memory), EEPROM(electrically erasable programmable read-only memory), CD-ROM, 블루레이 또는 다른 광학 디스크 스토리지, HDD(hard disk drive), SSD(solid state drive), 또는 플래시 메모리일 수 있으나, 이에 한정되지 않는다.
- [0148] 한편, 상술한 방법은 컴퓨터에서 실행될 수 있는 프로그램으로 작성 가능하고, 컴퓨터로 읽을 수 있는 기록매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다. 또한, 상술한 방법에서 사용된 데이터의 구조는 컴퓨터로 읽을 수 있는 기록매체에 여러 수단을 통하여 기록될 수 있다. 상기 컴퓨터로 읽을 수 있는 기록매체는 마그네틱 저장매체(예를 들면, 롬, 램, USB, 플로피 디스크, 하드 디스크 등), 광학적 판독 매체(예를 들면, 시디롬, 디브이디 등)와 같은 저장매체를 포함한다.
- [0149] 이제까지 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 상기된 기재의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 하며, 권리 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점을 포함하는 것으로 해석되어야 할 것이다.

도면

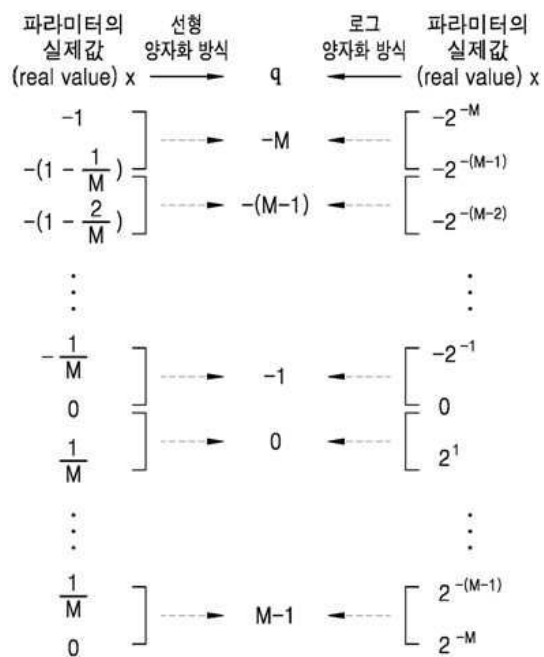
도면1



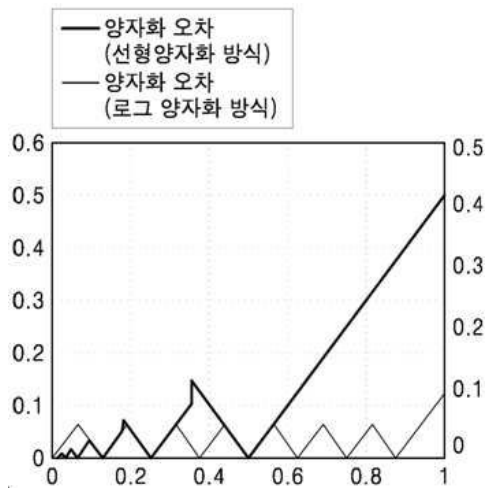
도면2



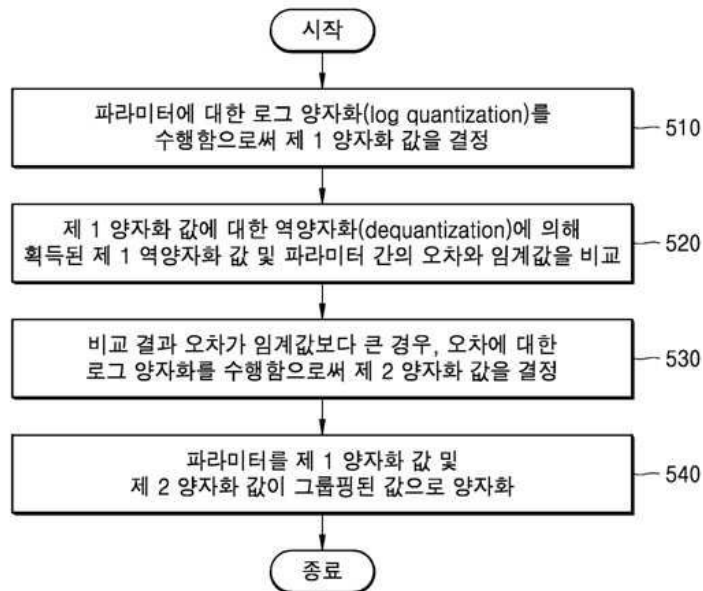
도면3



도면4



도면5



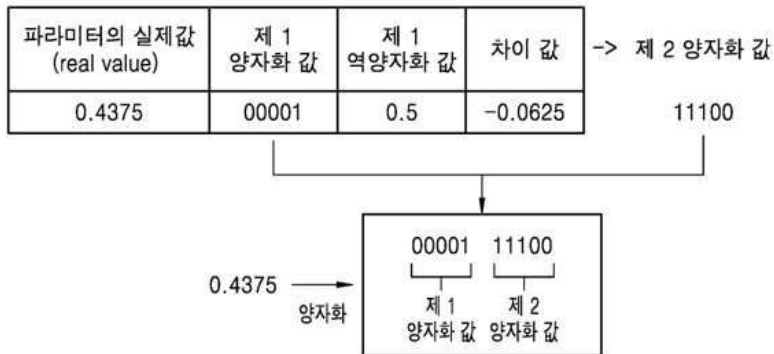
도면6

파라미터의 실제값 (real value)	제 1 양자화 값
0.5	00001
-0.25	11110
0.4375	00001

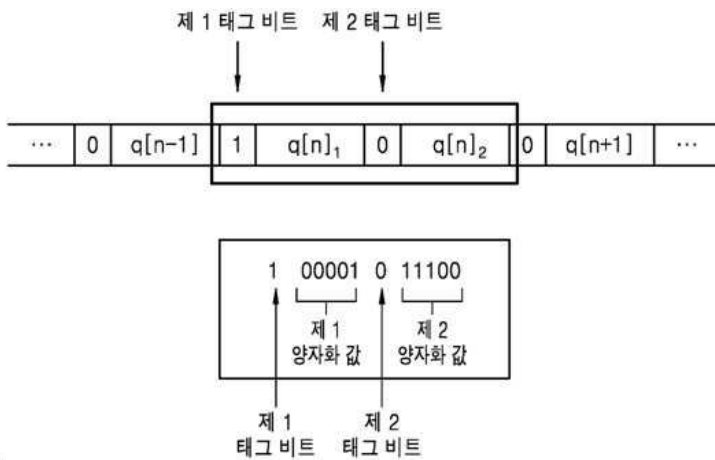
도면7

파라미터의 실제값 (real value)	제 1 양자화 값	제 1 역양자화 값	차이 값	임계값
0.5	00001	0.5	0	< 1/128
-0.25	11110	-0.25	0	< 1/128
0.4375	00001	0.5	-0.0625	> 1/128

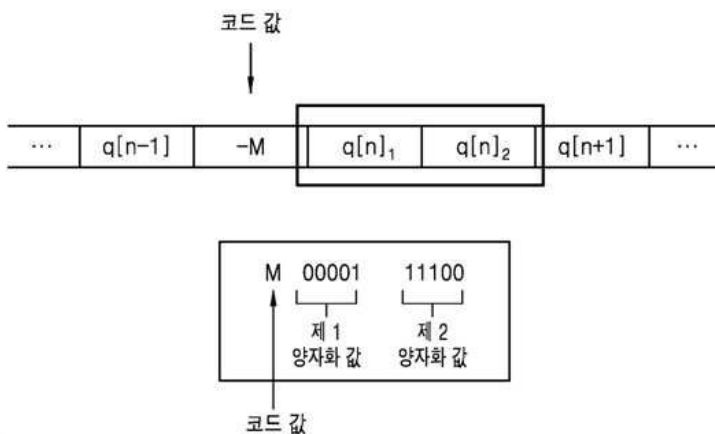
도면8



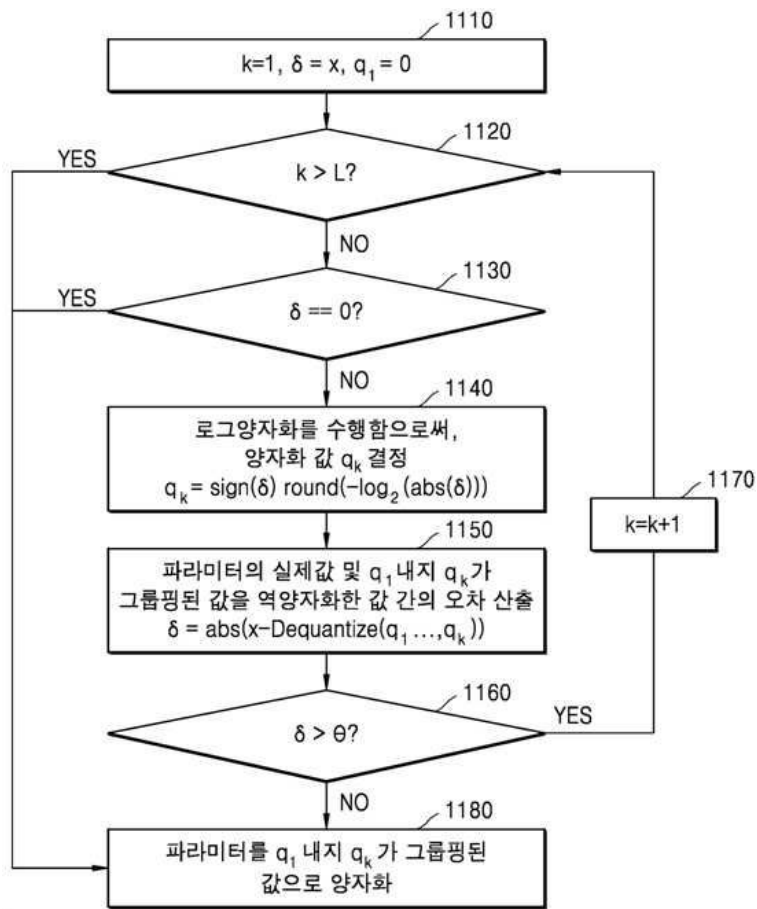
도면9



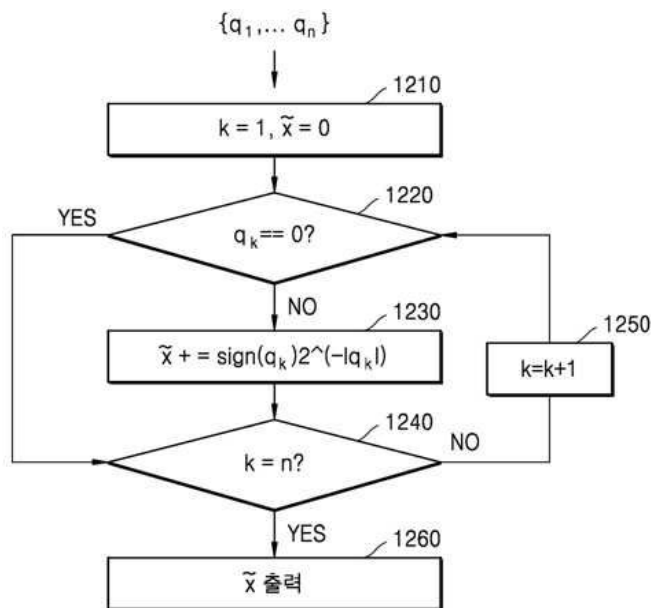
도면10



도면11



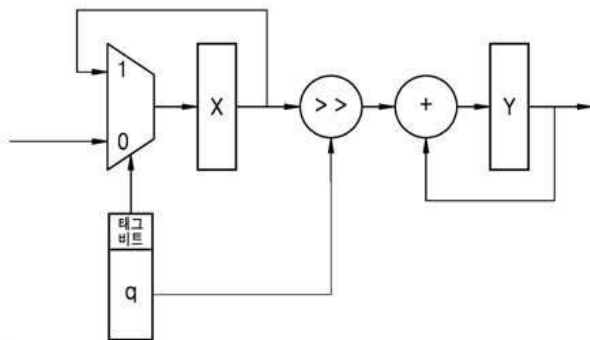
도면12



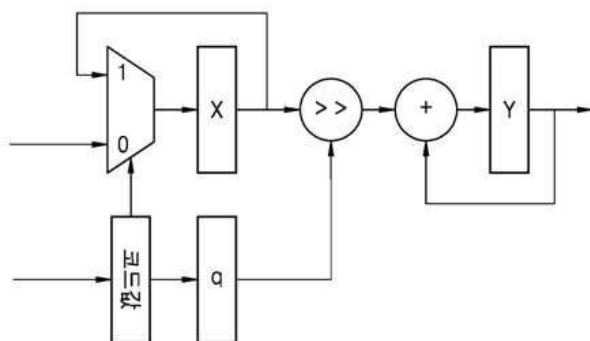
도면13

AlexNet	Precision	LinQ	LogQ	SLQ
	4b	1.23	75.82	78.57 (4.08b) 79.32 (4.46b)
	5b	3.94	76.24	78.9 (5.10b)
	6b	74.15	76.3	79.41 (5.57b)
	float 32b	80.7		
SqueezeNet	Precision	LinQ	LogQ	SLQ
	4b	0.58	77.94	78.98 (4.03b) 80.16 (4.22b)
	5b	58.08	78.14	80.82 (5.02b) 80.14 (5.28b)
	6b	80.44	78.14	80.90 (6.28b)
	float 32b	81.06		
VGG-S	Precision	LinQ	LogQ	SLQ
	5b	0.46	81.45	83.19 (5.03b) 83.70 (5.22b)
	6b	51.53	81.72	83.72 (5.99b)
	7b	82.7	81.52	83.79 (7.19b)
	float 32b	83.41		

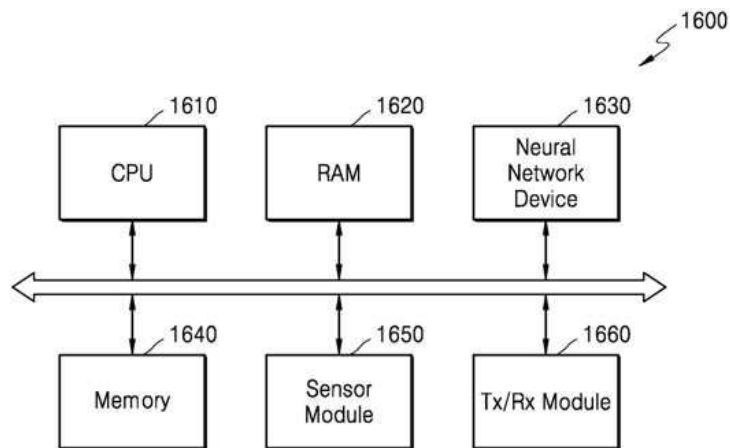
도면14



도면15



도면16



도면17

