



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0069394
(43) 공개일자 2024년05월20일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) G06N 3/04 (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2022-0150855
(22) 출원일자 2022년11월11일
심사청구일자 2022년11월11일

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
심현욱
경기도 수원시 영통구 삼성로 130 (매탄동)
오상윤
울산광역시 울주군 언양읍 유니스트길 50
이중은
울산광역시 울주군 언양읍 유니스트길 50
(74) 대리인
특허법인 무한

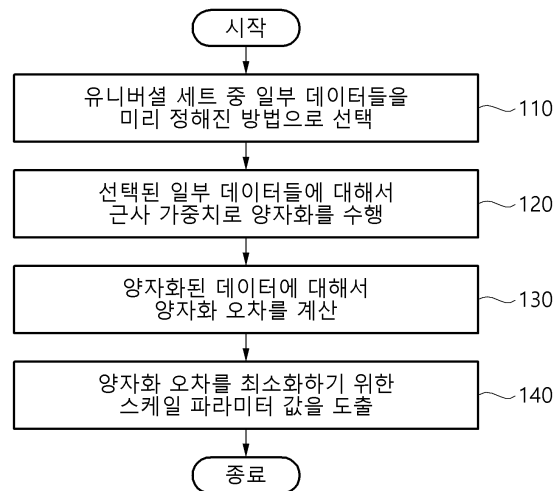
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 양자화 스케일링 장치 및 양자화를 위한 스케일링 파라미터의 계산 방법

(57) 요약

실시예는, 양자화 스케일링을 위한 장치 및 양자화 스케일링을 계산하는 방법에 대한 것이다. 실시예에 따른 양자화 스케일링 장치는, 데이터 세트의 데이터 비트들에 대해 미리 정해진 복수의 양자화 레벨을 지원하는 코드북 기반의 시프트 연산을 수행하는 시프터; 및 데이터 비트들의 양자화 스케일을 미리 정해진 그룹 별로 다르게 설정하여, 시프터를 제어하는 디코더를 포함하고, 시프터는 디코더의 제어에 따라 데이터 비트들을 양자화하여 출력할 수 있다.

대표도 - 도1



명세서

청구범위

청구항 1

양자화 스케일링을 위한 장치에 있어서,

데이터 세트의 데이터 비트들에 대해 미리 정해진 복수의 양자화 레벨을 지원하는 코드북 기반의 시프트 연산을 수행하는 시프터; 및

상기 데이터 비트들의 양자화 스케일을 미리 정해진 그룹 별로 다르게 설정하여, 상기 시프터를 제어하는 디코더

를 포함하고,

상기 시프터는 상기 디코더의 제어에 따라 상기 데이터 비트들을 양자화하여 출력하는,

장치.

청구항 2

제1항에 있어서,

상기 양자화 스케일은,

미리 정해진 방법으로 상기 데이터 세트를 근사 가중치로 양자화할 시, 양자화 오차를 최소화하기 위한 스케일 파라미터에 기초하여 결정되는,

장치.

청구항 3

제2항에 있어서,

상기 데이터 세트는,

임의로 생성된 유니버설 세트 중 Lloyd-Max 양자화 기법에 기초하여 생성된 서브 세트와 유사도가 높은 상기 일부 데이터들로부터 선택된 서브 세트에 해당하는,

장치.

청구항 4

제1항에 있어서,

상기 그룹은, 채널 또는 레이어 중 적어도 하나인,

장치.

청구항 5

제1항에 있어서,

상기 스케일 파라미터는 아래 수학적식에 기초하여 반복 계산되어 도출되는,

장치.

수학식:

$$\forall i, \quad \alpha q_i = \arg \min_{p \in S_Q} |p - w_i|$$

$$\mathcal{L} = \sum_i (w_i - \alpha q_i)^2$$

$$\alpha^* = \frac{\sum_i w_i \cdot q_i}{\sum_i q_i^2}$$

-여기서, α 는 상기 스케일 파라미터, w_i 는 서브 세트의 원소, q_i 는 상기 서브 세트의 양자화 포인트, L 은 상기 양자화 오차이고, $S_Q = \{\alpha q_i\}$ 임-

청구항 6

제1항에 있어서,

상기 데이터 비트들은 2의 로그 레벨로 양자화 되는,

장치.

청구항 7

양자화 스케일링을 위한 스케일 파라미터 계산 방법에 있어서,

양자화된 유니버설 세트 중 일부 데이터들을 미리 정해진 방법으로 선택하는 단계;

상기 선택된 일부 데이터들에 대해서 근사 가중치로 양자화를 수행하는 단계;

상기 양자화된 데이터에 대해서 양자화 오차를 계산하는 단계; 및

상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계

를 포함하는,

스케일 파라미터 계산 방법.

청구항 8

제7항에 있어서,

일부 데이터들을 미리 정해진 방법으로 선택하는 단계는,

상기 유니버설 세트들 중 Lloyd-Max 양자화 기법에 기초하여 생성된 서브 세트와 유사도가 높은 상기 일부 데이터를 선택하는 단계

를 포함하는,

스케일 파라미터 계산 방법.

청구항 9

제7항에 있어서,

상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계는,
 상기 스케일 파라미터의 초기 값을 결정하는 단계;
 상기 양자화 오차의 변화에 기초하여 상기 스케일 파라미터를 업데이트하는 단계; 및
 상기 양자화 오차의 변화가 미리 정해진 기준 이하에 해당하면, 상기 스케일 파라미터를 출력하는 단계
 를 포함하는,
 스케일 파라미터 계산 방법.

청구항 10

제7항에 있어서,
 상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계는,
 아래 수학적식에 기초하여 반복적으로 계산함으로써 상기 양자화 오차를 최소화하는 상기 스케일 파라미터 값을
 도출하는 단계
 를 포함하는,
 스케일 파라미터 계산 방법.
 수학적식:

$$\forall i, \quad \alpha q_i = \arg \min_{p \in S_Q} |p - w_i|$$

$$\mathcal{L} = \sum_i (w_i - \alpha q_i)^2$$

$$\alpha^* = \frac{\sum_i w_i \cdot q_i}{\sum_i q_i^2}$$

-여기서, α 는 상기 스케일 파라미터, w_i 는 서브 세트의 원소, q_j 는 상기 서브 세트의 양자화 포인트, L 은 상기 양자화 오차이고, $S_Q = \{\alpha q_j\}$ 임-

청구항 11

제7항에 있어서,
 상기 스케일 파라미터는,
 채널 별로 또는 레이어 별로 다르게 설정되는,
 스케일 파라미터 계산 방법.

청구항 12

하드웨어와 결합되어 제7항 내지 제11항 중 어느 하나의 항의 방법을 실행시키기 위하여 기록 매체에 저장된 컴퓨터 프로그램.

청구항 13

양자화 스케일링을 위한 스케일 파라미터를 계산하는 장치에 있어서,
 하나 이상의 프로세서;
 메모리; 및
 상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그램을 포함하고,
 상기 프로그램은,
 양자화된 유니버설 세트 중 일부 데이터들을 미리 정해진 방법으로 선택하는 단계;
 상기 선택된 일부 데이터들에 대해서 근사 가중치로 양자화를 수행하는 단계;
 상기 양자화된 데이터에 대해서 양자화 오차를 계산하는 단계; 및
 상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계
 를 포함하는,
 장치.

청구항 14

제13항에 있어서,
 일부 데이터들을 미리 정해진 방법으로 선택하는 단계는,
 상기 유니버설 세트들 중 Lloyd-Max 양자화 기법에 기초하여 생성된 서브 세트와 유사도가 높은 상기 일부 데이터들을 선택하는 단계
 를 포함하는,
 장치.

청구항 15

제13항에 있어서,
 상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계는,
 상기 스케일 파라미터의 초기 값을 결정하는 단계;
 상기 양자화 오차의 변화에 기초하여 상기 스케일 파라미터를 업데이트하는 단계; 및
 상기 양자화 오차의 변화가 미리 정해진 기준 이하에 해당하면, 상기 스케일 파라미터를 출력하는 단계
 를 포함하는,
 장치.

청구항 16

제13항에 있어서,
 상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계는,
 아래 수학적식에 기초하여 반복적으로 계산함으로써 상기 양자화 오차를 최소화하는 상기 스케일 파라미터 값을
 도출하는 단계
 를 포함하는,

장치.

수학식:

$$\forall i, \quad \alpha q_i = \arg \min_{p \in S_Q} |p - w_i|$$

$$\mathcal{L} = \sum_i (w_i - \alpha q_i)^2$$

$$\alpha^* = \frac{\sum_i w_i \cdot q_i}{\sum_i q_i^2}$$

-여기서, α 는 상기 스케일 파라미터, w_i 는 서브 세트의 원소, q_i 는 상기 서브 세트의 양자화 포인트, L 은 상기 양자화 오차이고, $S_Q = \{\alpha q_i\}$ 임-

청구항 17

제13항에 있어서,

상기 스케일 파라미터는,

채널 별로 또는 레이어 별로 다르게 설정되는,

장치.

발명의 설명

기술 분야

[0001] 실시예는, 양자화 스케일링 장치 및 양자화를 위한 스케일링 파라미터의 계산 방법에 관한 것이다.

배경 기술

[0003] 딥 러닝이 많은 기계 학습 작업에서 가장 높은 성능을 발휘하는 방법이 되면서 하드웨어 DNN(Deep Neural Network) 가속기에 대한 관심이 증가하고 있다. 이미지 향상 및 초해상도 애플리케이션 등의 컴퓨터 비전 작업을 위해서는 DNN에 대해 종종 매우 높은 컴퓨팅 및 메모리가 요구된다. 이러한 DNN 가속기의 하드웨어 비용과 메모리 공간을 줄이기 위해 양자화는 매우 효과적인 접근 방식이 될 수 있다.

[0004] 인공 신경망의 양자화는 양자화 함수(quantizer)와 이를 활용하는 양자화 알고리즘(quantization algorithm)으로 구성된다. 양자화 함수는 특정한 스케일 및 가용 비트수를 고려한 양자화 표현 값의 집합인 코드북(codebook)을 정의하여 대상 데이터를 정수 값으로 변환하여 데이터 복잡도를 줄이는 역할을 한다. 한편, 양자화 알고리즘은 양자화 함수를 최적화하여 신경망의 데이터 복잡도를 줄이면서도 이로 인한 성능 저하를 최소화하는 알고리즘을 의미한다.

[0005] 일반적으로 양자화 알고리즘의 양자화 함수 최적화는 코드북, 입력 데이터와 같이 양자화 과정에서 필요한 대상을 스케일링하는 양자화 파라미터를 활용한다.

발명의 내용

해결하려는 과제

과제의 해결 수단

- [0007] 실시예에 따른 양자화 스케일링 장치는, 데이터 세트의 데이터 비트들에 대해 미리 정해진 복수의 양자화 레벨을 지원하는 코드북 기반의 시프트 연산을 수행하는 시프터; 및 데이터 비트들의 양자화 스케일을 미리 정해진 그룹 별로 다르게 설정하여, 시프터를 제어하는 디코더를 포함하고, 시프터는 디코더의 제어에 따라 데이터 비트들을 양자화하여 출력할 수 있다.
- [0008] 상기 양자화 스케일은, 미리 정해진 방법으로 상기 데이터 세트를 근사 가중치로 양자화할 시, 양자화 오차를 최소화하기 위한 스케일 파라미터에 기초하여 결정될 수 있다.
- [0009] 상기 데이터 세트는, 임의로 생성된 유니버설 세트 중 Lloyd-Max 양자화 기법에 기초하여 생성된 서브 세트와 유사도가 높은 상기 일부 데이터들로부터 선택된 서브 세트에 해당할 수 있다.
- [0010] 상기 그룹은, 채널 또는 레이어 중 적어도 하나일 수 있다.
- [0011] 상기 스케일 파라미터는 아래 수학식에 기초하여 반복 계산되어 도출될 수 있다.
- [0012] 수학적식:

$$\forall i, \quad \alpha q_i = \arg \min_{p \in S_Q} |p - w_i|$$

$$\mathcal{L} = \sum_i (w_i - \alpha q_i)^2$$

$$\alpha^* = \frac{\sum_i w_i \cdot q_i}{\sum_i q_i^2}$$

- [0013]
- [0014] -여기서, α 는 상기 스케일 파라미터, w_i 는 서브 세트의 원소, q_i 는 상기 서브 세트의 양자화 포인트, L 은 상기 양자화 오차이고, $S_Q = \{\alpha q_i\}$ 임-
- [0015] 상기 데이터 비트들은 2의 로그 레벨로 양자화 될 수 있다.
- [0016] 양자화 스케일링을 위한 스케일 파라미터 계산 방법에 있어서, 양자화된 유니버설 세트 중 일부 데이터들을 미리 정해진 방법으로 선택하는 단계; 상기 선택된 일부 데이터들에 대해서 근사 가중치로 양자화를 수행하는 단계; 상기 양자화된 데이터에 대해서 양자화 오차를 계산하는 단계; 및 상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계를 포함하는, 스케일 파라미터 계산 방법이 제공될 수 있다.
- [0017] 일부 데이터들을 미리 정해진 방법으로 선택하는 단계는, 상기 유니버설 세트들 중 Lloyd-Max 양자화 기법에 기초하여 생성된 서브 세트와 유사도가 높은 상기 일부 데이터들을 선택하는 단계를 포함할 수 있다.
- [0018] 상기 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계는, 상기 스케일 파라미터의 초기 값을 결정하는 단계; 상기 양자화 오차의 변화에 기초하여 상기 스케일 파라미터를 업데이트하는 단계; 및 상기 양자화 오차의 변화가 미리 정해진 기준 이하에 해당하면, 상기 스케일 파라미터를 출력하는 단계를 포함할 수 있다.

도면의 간단한 설명

- [0020] 도 1은 실시예에서, 양자화 스케일링을 위한 스케일 파라미터를 도출하는 방법의 흐름도이다.
- 도 2는 실시예에서, 유니버설 세트 중 데이터들을 선택하는 방법을 설명하기 위한 도면이다.
- 도 3은 실시예에서, 양자화 스케일링 방법의 전체 스킴을 시각적으로 나타낸 도면이다.
- 도 4는 실시예에서, 양자화 스케일을 수행하는 실시예에 대한 것이다.
- 도 5는 실시예에서, 양자화 스케일링 장치의 구조를 설명하기 위한 도면이다.
- 도 6은 실시예에서, 스케일 파라미터를 도출하는 방법을 수행하기 위한 장치의 구성을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0021] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다. 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.
- [0022] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0023] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0025] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.
- [0026] 또한, 실시 예의 구성 요소를 설명하는 데 있어서, 제1, 제2, A, B, (a), (b) 등의 용어를 사용할 수 있다. 이러한 용어는 그 구성 요소를 다른 구성 요소와 구별하기 위한 것일 뿐, 그 용어에 의해 해당 구성 요소의 본질이나 차례 또는 순서 등이 한정되지 않는다. 어떤 구성 요소가 다른 구성요소에 "연결", "결합" 또는 "접속"된다고 기재된 경우, 그 구성 요소는 그 다른 구성요소에 직접적으로 연결되거나 접속될 수 있지만, 각 구성 요소 사이에 또 다른 구성 요소가 "연결", "결합" 또는 "접속"될 수도 있다고 이해되어야 할 것이다.
- [0027] 어느 하나의 실시 예에 포함된 구성요소와, 공통적인 기능을 포함하는 구성요소는, 다른 실시 예에서 동일한 명칭을 사용하여 설명하기로 한다. 반대되는 기재가 없는 이상, 어느 하나의 실시 예에 기재한 설명은 다른 실시 예에도 적용될 수 있으며, 중복되는 범위에서 구체적인 설명은 생략하기로 한다.
- [0029] 실시예에 따른 양자화 방법은 서브 세트 양자화에 기초한다. 이를 위해 시뮬레이션된 양자화 기능의 범위에 대한 양자화 포인트 세트를 정의할 수 있다. 실시예에서는 4비트로 양자화 하는 경우를 중심으로 설명하도록 한다.
- [0030] 코드북 기반의 유니버설 세트에 대해서 서브 세트들을 설계할 수 있다. 이때 양자화 정밀도에 의해 서브 세트에 포함되는 데이터 수가 결정될 수 있다. 앞서 설명한 바와 같이 실시예는 4비트의 서브 세트를 설계할 수 있다.
- [0031] 실시예에서 n 비트의 양자화의 경우(부호 비트를 포함함), 0 또는 양수에서 $N=2n-1$ 값을 선택할 수 있다. 이때, 0을 포함하거나 포함하지 않을 수 있다. 실시예에 따른 $n=4$ 이면 $N=8$ 이고, $n=3$ 이면, $N=4$ 이다. 서브 세트를 선택하는 방법에는 여러가지 방법이 적용될 수 있다.
- [0032] 양자화를 위한 서브 세트가 선택되면, 0에 대칭이 되는 음수의 서브 세트는 부호를 통해 간단히 정의될 수 있다. 양자화 포인트가 대칭되도록 정의하면, 양자화 장치를 설계하기에 하드웨어 복잡성을 줄일 수 있다.
- [0033] 실시예에서 생성된 서브 세트에 대해 스케일을 결정하기 위한 스케일 파라미터가 정의될 수 있다. 스케일 파라미터는 유니버설 세트에 대한 양자화 결과에 기초하여 결정될 수 있고, 하드웨어적으로 후속 계층의 일부로 구현될 수 있다.
- [0035] 도 1은 실시예에서, 양자화 스케일링을 위한 스케일 파라미터를 도출하는 방법의 흐름도이다.
- [0036] 실시예는 코드북의 유니버설 세트(universal set)에 대해 복수의 서브 세트들을 생성하고, 생성된 서브 세트 중

코드북 유사도 및 로이드 맥스(Lloyd-Max) 양자화를 통해 최적의 서브 세트를 도출할 수 있다. 또한, 도출된 서브 세트에 기초하여 양자화 오차를 최소화하는 스케일 파라미터를 획득할 수 있다. 이를 위해, 실시예에서는 신경망을 양자화하는 데에 학습을 위한 데이터 세트를 필요로 하지 않는다.

[0037] 단계(110)에서 장치는, 양자화된 유니버설 세트 중 일부 데이터들을 미리 정해진 방법으로 선택한다.

[0038] 도 2는 실시예에서, 유니버설 세트 중 데이터들을 선택하는 방법을 설명하기 위한 도면이다.

[0039] 실시예에서, 유니버설 세트는 임시 가중치에 의한 코드북을 탐색함으로써 생성될 수 있다. 장치는, 유니버설 세트 중, 로이드 맥스 양자화 기법을 통해 부동 소수점으로 구성된 이상적인 세트, 코드북 L(Lloyd-Max Codebook L)을 생성할 수 있다.

[0040] 한편, 유니버설 세트에 포함되는 원소들에 대해서 코드북 L에 포함되는 원소와 유사도가 계산되며, 그 결과에 기초하여 일부 데이터들이 선택될 수 있다.

[0041] 예컨대, 계산된 유사도 값은 비어있는 큐(Queue, D)에 보관될 수 있다. 유니버설 세트들에 대한 유사도의 계산이 종료되면, 장치는 D에 저장된 값들을 값이 낮은 순서대로 정렬할 수 있다. 장치는 코드북 L과 유사도가 높은 원소들을 선택되, 예컨대 3%에 해당하는 데이터들이 선택될 수 있다.

[0042] 실시예를 통해 선택된 데이터들은 새로운 코드북의 후보인 서브 세트들을 생성하는 데에 활용될 수 있다.

[0044] 다시 돌아가, 단계(120)에서 장치는, 선택된 일부 데이터들에 대해서 근사 가중치로 양자화를 수행한다.

[0045] 실시예에서는, 선택된 일부 데이터들에 대해서 복수의 서브 세트들이 생성될 수 있다. 서브 세트들을 생성하는 과정에서 선택된 모든 데이터들에 대해서 근사 가중치 w 를 대상으로 양자화를 수행하고, 양자화된 근사 가중치 값을 획득할 수 있다.

[0046] 단계(130)에서 장치는, 양자화된 데이터에 대해서 양자화 오차를 계산한다.

[0047] 실시예에서, 서브 세트에 포함되는 근사 가중치 w 와 양자화된 근사 가중치 w_q 간의 차이를 계산할 수 있다.

[0048] 단계(140)에서 장치는, 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출한다.

[0049] 실시예에서, 상기 계산된 결과에 기초하여 양자화 오차를 줄이기 위한 스케일 파라미터 값을 계산할 수 있다. 실시예에 따른 스케일 파라미터는 반복적으로 누적되어 계산될 수 있다. 예컨대, 1차 계산된 스케일 파라미터에 기초하여 서브 세트에 대한 양자화를 다시 수행하고 양자화 오차를 다시 계산할 수 있으며, 계산된 양자화 오차에 기초하여 스케일 파라미터가 다시 계산될 수 있다.

[0050] 이러한 사이클은, 미리 정해진 기준까지 양자화 오차가 감소하도록 또는 더 이상 양자화 오차가 감소하지 않을 때까지 이루어질 수 있다. 이에, 계산된 스케일 파라미터에 기초하여 근사 가중치를 양자화 오차가 적은 양자화 포인트로 양자화되도록 할 수 있다.

[0052] 도 3은 실시예에서, 양자화 스케일링 방법의 전체 스킴을 시각적으로 나타낸 도면이다.

[0053] 실시예는 앞서 설명한 바와 같이, $N=8$ 인 부호 비트를 포함하는 4 비트의 양자화에 관한 것이다.

[0054] 실시예에서, 코드북 기반의 유니버설 세트가 생성될 수 있고, 유니버설 세트에서 로이드 맥스에 기초한 방법으로 추려진 데이터들 중 N 개의 데이터를 이용하여 양자화 포인트를 나타내는 서브 세트가 정의될 수 있다.

[0055] 실시예에서, 시뮬레이션된 양자화 범위에 해당하는 서브 세트를 정의할 수 있다. 시뮬레이션에 적용된 양자화 장치는 양자화 장치 및 역양자화 장치를 연속적으로 적용하며, 입력 및 출력은 동일한 스케일을 갖는다. 이에, 서브 세트는 양자화 오차 없이 양자화가 가능한 입력 도메인의 점 집합에 해당할 수 있다. 시뮬레이션에 따라 입력이 서브 세트에 가까운 원소들로 근사화될 수 있다. 다른 양자화 체계와 달리, 실시예는 입력 이외에 양자화 로스를 최소화하기 위한 스케일 파라미터를 이용하는 방식으로 일반적인 양자화 장치를 이용하여 설계될 수 있다.

[0056] [수학식 1]

$$Q(x, S_Q) = \arg \min_{p \in S_Q} |x - p|.$$

[0057]

[0058] 상기 수학식 1은 시뮬레이션된 양자화 장치에 의해 서브 세트의 양자화 결과에 해당한다. x 는 입력, S_Q 는 양자화 로스가 최소화된 서브 세트를 의미하고, p 는 서브 세트에 포함되는 양자화 포인트를 의미한다.

[0059] 예컨대, 선형 양자화를 위한 서브 세트는 아래의 수학식 2로 정의될 수 있다.

[0060] [수학식 2]

$$S_Q^{lin}(\alpha) = \{\alpha \cdot i \mid i = -N, -N + 1, \dots, N - 1\}$$

[0061]

[0062] 여기서, α 는 스케일 파라미터에 해당하고, $2N$ 은 양자화 포인트의 수를 나타낸다. 즉, k 비트 양자화의 경우, $N=2^k-1$ 로 나타낼 수 있고, 0을 중심으로 대칭인 형태를 적용할 수 있다.

[0063] 실시예에 따른 로그 스케일 양자화를 위한 서브 세트는 아래의 수학식 3과 같이 정의될 수 있다.

[0064] [수학식 3]

$$S_Q^{log}(\alpha) = \{-\alpha 2^{-i} \mid i = 0, 1, \dots, N - 1\} \cup \{0\} \\ \cup \{\alpha 2^{-i} \mid i = 0, 1, \dots, N - 2\}$$

[0065]

[0066] 상기과 같은 로그 스케일 양자화의 정확도를 향상시키기 위한 방법으로, 아래의 수학식 4와 같이 정의되는 2의 양자화 포인트에 대한 서브 세트가 이용될 수 있다.

[0067] [수학식 4]

$$S_Q^{2log}(\alpha) = \{q_1 + q_2 \mid q_1 \in S_Q^{log}(\alpha), q_2 \in S_Q^{log}(\alpha)\}$$

[0068]

[0069] 여기서, q_1 , q_2 는 각각 스케일 파라미터에 대응하는 양자화 포인트를 의미한다.

[0070] 실시예에 따른 서브 세트의 양자화에 대한 개념은 더 큰 집합인 전체 코드북 기반의 유니버설 세트에서 미리 정의된 크기를 가지는 임의의 코드북에 해당하는 서브 세트를 정의하는 것이다.

[0071] [수학식 5]

$$S_Q^{sq}(\alpha) = \{\text{any } q_i \in S_U(\alpha) \mid i = 1, \dots, 2N\}$$

[0072]

[0073] 수학식 5는 $2N$ 개 이하의 원소를 포함하는 전체 코드북 $S_U(\alpha)$ 의 하위 집합이다. 양자화 정밀도는 양자화 포인트 세트의 크기를 제한하나 유니버설 세트에 대해서는 제한되지 않는다.

[0074] 실시예에서, 수학식 1에 따르면 시뮬레이션된 양자화 장치에 대해 정의되어 있고, 스케일 파라미터에 의해 유동적으로 정의될 수 있다. 즉, 스케일 파라미터를 결정하는 단계를 통해 양자화 장치에서 양자화되는 결과가 최적화될 수 있다.

[0075] 실시예에 따른 양자화 방식은 각 계층, 또는 각 채널에 대해서 서로 다른 서브 세트가 선택되어 이루어질 수 있다.

[0076] 모든 실수 값이 유니버설 세트로 사용될 수 있으나, 하드웨어 복잡성이 낮고 효율이 높은 2-로그 스케일 양자화 방식에 대응하는 양자화 포인트 세트와 유사한 세트를 유니버설 세트(즉, S_Q^{2log})로 사용하도록 한다.

[0077] 이하에서는 서브 세트에 대한 양자화가 최적화되는 과정에 대해서 설명하도록 한다.

[0078] 서브 세트에 대한 양자화 오차를 최소화하기 위해 서브 세트의 양자화를 최적화하며, 이를 위해 서브 세트를 확

정하고, 스케일 파라미터 값을 결정할 수 있다.

[0079] 실시예에 따른 최적의 스케일 파라미터를 찾기 위해, 주어진 양자화 포인트 세트에서 최적의 스케일 파라미터를 효율적으로 찾을 수 있다.

[0080] [수학식 6]

$$\forall i, \quad \alpha q_i = \arg \min_{p \in S_Q} |p - w_i|.$$

$$\mathcal{L} = \sum_i (w_i - \alpha q_i)^2.$$

[0081]

[0082] 수학식 6은 양자화 오차를 계산하는 방법에 대한 것이다. q_i 를 서브 세트에 해당하는 양자화 포인트 세트라고 가정하면(스케일 파라미터의 적용 이전), 스케일 파라미터를 적용한 최종 양자화 포인트 세트는 $S_Q = \{\alpha q_i\}$ 라고 할 수 있다. 서브 세트 값인 w_i 와 양자화된 값 사이의 손실을 계산할 수 있고, αq_i 를 w_i 의 가장 가까운 양자화 포인트로 가정한다.

[0083] [수학식 7]

$$\alpha^* = \frac{\sum_i w_i \cdot q_i}{\sum_i q_i^2}$$

[0084]

[0085] 앞서 설명한 바와 같이, 양자화 오차(L)을 최소화하기 위한 스케일 파라미터를 찾기 위해 상기와 같은 α 에 대한 L의 도함수를 0으로 설정할 수 있다.

[0086] 앞서, 수학식 6에서 αq_i 를 찾는 것은 α 의 현재 값에 의존하므로, α 가 일정 값으로 수렴할 때까지 수학식 6 및 수학식 7의 과정이 반복될 수 있다. 이때, α 는 1을 초기 값으로 가지도록 설정할 수 있다.

[0087] 실험적 결과로, 실시예의 최적화 방법이 α 에 대한 허용 오차가 $1e-5$ 일 때, 평균 17회의 반복 이내에서 수렴 값을 획득할 수 있다.

[0089] 도 4는 실시예에서, 양자화 스케일을 수행하는 실시예에 대한 것이다.

[0090] 이하에서는, L2 양자화 오차를 최소화하는 주어진 양자화 포인트 세트 $S(\alpha) = \{\alpha q_i\}$ 에 대한 최적의 스케일 파라미터를 찾는 실시예에 대해 설명한다.

[0091] 이를 위해 먼저 α 를 1로 초기화하고, α 가 더 이상 변경되지 않을 때까지 수학식 7을 적용하여 α 를 업데이트할 수 있다. 수학식 7의 우항의 q_i 는 아래의 수학식 8과 같이 $q(w_i, \alpha)$ 를 약식으로 표기한 것이다.

[0092] [수학식 8]

$$q_i := q(w_i, \alpha) = \frac{1}{\alpha} \cdot \arg \min_{p \in S(\alpha)} |p - w_i|$$

[0093]

[0094] 수학식 7에 따라 α 가 업데이트되면 $q_i = q(w_i, \alpha)$ 가 업데이트될 수도 있고 업데이트되지 않을 수도 있다. 만약, q_i 가 업데이트되지 않으면(즉, $\forall_i, q(w_i, \alpha) = q(w_i, \alpha^*)$)이면), 양자화 오차는 α 에 대한 간단한 2차 함수가 되고, α 의 최적 값은 수학식 7에 의해서 결정될 수 있고, 이후 반복을 종료할 수 있다.

[0095] 도 4에 의하면, α 가 $\Delta \alpha = \alpha' - \alpha$ 만큼 증가하면 가장 가까운 양자화 포인트가 변경될 수 있다. 실시예에서, $q_1=1$, $q_2=2$ 이다.

[0096] 일부 q_i 가 업데이트되면(즉, $\exists_i, q(w_i, \alpha) \neq q(w_i, \alpha^*)$), 이는 L 값이 일시적으로 감소하거나 동일하거나, 또는 일시적으로 증가하나 이후 감소하게 된다. 이를 확인하기 위해 도 4의 실시예를 들어 설명하도록 한다.

- [0097] 실시예에서 $S(a) = \{a_{q_1}, a_{q_2}\}$ 를 사용하는 첫 번째 반복($t=1$)에서는 양자화 포인트 세트에 양자화된 가중치 값 w 하나만 고려한다. 첫 번째 반복에서는 $a' > a$ 라고 가정하고 a 를 a' 로 업데이트할 수 있다.
- [0098] 이후, 두 번째 반복($t=2$)에서 w 는 $S(a') = \{a'_{q_1}, a'_{q_2}\}$ 를 사용하여 양자화될 수 있다. 더 단순화하기 위해 양자화 포인트 q_1, q_2 가 $q_1=1, q_2=2$ 라고 가정한다. 도 4에서 $A=1.5a, C=1.5a', B=A+0.5\Delta a$ 이고, 여기서 $\Delta a = a' - a$ 이다. A, B, C 의 포인트를 기준으로 세 가지 예시를 들 수 있다.
- [0099] 예시 1: a' 와 $1.5a (=A)$ 사이의 모든 점은 $a_{q_1}(t=1)$ 에서 또는 $a'_{q_1}(t=2)$ 에서 양자화될 수 있다. 이 경우 $q(w, a)$ 는 업데이트되지 않는다.
- [0100] 예시 2: 유사하게, $1.5a' (=C)$ 와 $2a$ 사이의 모든 점은 $a_{q_2}(t=1)$ 에서 또는 $a'_{q_2}(t=2)$ 에서 양자화될 수 있다. 마찬가지로, $q(w, a)$ 는 업데이트되지 않는다.
- [0101] 예시 3: w 가 점 A 와 C 사이에 있으면 $q(w, a)$ 가 q_2 에서 q_1 으로 업데이트될 수 있다. 이에 대한 두 가지 일례를 들 수 있다.
- [0102] 먼저, w 가 B 보다 크면 $t=2$ 에서의 양자화 오차가 $t=1$ 의 양자화 오차에 비해 감소하게 된다. 반대로 w 가 B 보다 작으면 $t=1$ 에 비해 양자화 오차가 증가할 수 있다. 후자의 경우 수학적 7에 의한 양자화 오차가 단조롭게 감소된다고 할 수 없다. 다만, 진동하고 수렴하지 않기 위해서는 수학적 7에 의해 a 를 감소시켜야 하고 이에 따라 다음 사이클에 $t=1$ 의 상황으로 갈 수 있어야 하나 이는 불가능하다. w 가 B 보다 작을 때 양자화 오차는 $w - a'$ 이고, 이는 a 가 증가할 때 최소화되기 때문이다. 두 a 값 사이에 무한 진동이 있을 수 없으며 양자화 오차가 일시적으로 증가하지만 결국에는 최소값으로 수렴할 수 있다.
- [0103] 실시예는 $a' > a$ 의 경우에 대한 것이나, 반대로 $a' < a$ 의 수렴도 유사한 방식으로 나타낼 수 있다.
- [0104] 실시예에 따른 알고리즘은 아래와 같이 표현될 수 있다.
- [0105] [표 1]

Algorithm 1: FINDBESTQPS

Input: w : pretrained weight, S_U : universal set, $N: 2^{k-1}$, $\text{loss}(w, S)$: loss function

Result: S_Q : the QPS that minimizes loss

```

1  $l_{\min} \leftarrow \infty$ 
2 for  $S$  in each  $N$ -element subset of  $S_U$  do
3    $\alpha \leftarrow \text{FINDSCALEFACTOR}(S, w)$ 
4    $l_{\text{curr}} \leftarrow \text{loss}(w, \alpha S)$ 
5   if  $l_{\text{curr}} < l_{\min}$  then
6      $S_Q \leftarrow \alpha S$ 
7      $l_{\min} \leftarrow l_{\text{curr}}$ 
8   end
9 end
10 return  $S_Q$ 

```

- [0106]
- [0107] 최적의 서브 세트를 찾기 위해 고려할 서브 세트의 수는 $\frac{|S_U|}{|S_Q|}$ 개이다. 따라서, 최소 양자화 오차를 제공하는 모든 경우에 대해서 고려될 수 있다. 각 레이어 또는 각 채널에 대해서 실시예에 따른 스케일 파라미터가 도출될 수 있다.
- [0109] 이하에서는 유니버설 세트를 생성하는 방법에 대해 설명하도록 한다.
- [0110] 유니버설 세트 S_U 의 초기 후보는 $S_Q^{2\log}$ 로 표현될 수 있으며, 이는 2개의 시프터 및 1개의 덧셈기만을 하드웨어로 구현될 수 있다. 실시예는 이러한 하드웨어에 최적화를 위한 설계 옵션을 고려할 수 있다.
- [0111] 실시예는 tb 의 형태를 고려한다. 여기서 a 와 b 는 아래의 표 2에 나열된 것처럼 0 또는 $2k$ 로 표현될 수 있다.

스케일 파라미터에 의해 서브 세트는 스케일된 버전과 동일하므로 다음과 같이 스케일되었다고 가정하며, 값은 1이하에 해당한다.

[표 2]

Option	A	B
1	$\{2^{-1}, 2^{-2}, 2^{-3}, 0\}$	$\{2^{-1}, 2^{-2}, 2^{-4}, 0\}$
2	$\{1, 2^{-2}, 2^{-4}, 0\}$	$\{2^{-1}, 2^{-3}, 2^{-5}, 0\}$
3	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 0\}$	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 0\}$
4	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 0\}$	$\{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}, 0\}$
5 (chosen)	$\{1, 2^{-1}, 2^{-3}, 0\}$	$\{1, 2^{-2}, 2^{-4}, 0\}$

실시예에 따른 유니버설 세트의 설계는 하드웨어의 복잡성과 표현력 면에서 균형 있게 조절될 수 있다. 예컨대, 옵션 4는 23개의 원소가 있는 유니버설 세트를 생성하는 반면, 옵션 3과 옵션 5는 각각 17개와 16개 원소를 갖지만 옵션 4가 가장 복잡하다. 실시예는 옵션 5가 선택되어 간단한 하드웨어로 구현될 수 있다.

양자화 스케일링을 위한 하드웨어는 도 5와 같이 표현될 수 있다.

도 5는 양자화 스케일링을 위한 양자화 장치에 대한 것으로, MAC(Multiply-and-accumulate)을 이용한 하드웨어 설계에 대해 설명하기 위한 도면이다.

도 5(a)는 MAC 기반의 양자화 장치의 구조를 나타내며, 도 5(b)는 양자화 장치의 디코더의 구조를 나타낸다.

실시예는 선형으로 양자화되고, 가중치는 서브 세트로 양자화되는 것으로 가정할 수 있다. 유니버설 세트를 구현하는 간단한 방법으로, 두 개의 시프터와 하나의 덧셈기를 이용할 수 있다. 여기서, 시프터는 배럴 시프터라 불리는 가변량 시프터에 해당할 수 있다.

모든 유니버설 세트의 정의를 활용하여 도식된 바와 같이 단순한 디자인의 양자화 장치로 최적화될 수 있다. 도 5(a)에 따르면, 각 시프터(shifter 1, 2)는 MUX(Multiplexer) 및 일정량이 시프트되는 2개의 시프터로 대체될 수 있다. 시프터는 양자화 대상에 해당하는 데이터 비트들에 대해 미리 정해진 복수의 양자화 레벨을 지원하는 코드북 기반의 시프트 연산을 수행할 수 있다.

설계에 사용되는 MUX는 배럴 시프터에 대비하여 경제적이다. 예컨대, 일정량의 상수를 시프트하는 데에 논리 게이트가 사용되지 않으며, 단순한 와이어를 통해 구현될 수 있다. 실시예에 따른 하드웨어 구성은 유니버설 세트가 생성되는 방식에 기초하여 설계될 수 있다. 여기서, 각 항에는 0을 포함하여 4개의 결과가 출력될 수 있다.

각 MUX는 독립적으로 입력을 선택할 수 있으므로 16개의 서로 다른 조합이 가능하며, 그 중 양자화된 가중치 값의 나머지가 제한되어 실제 사용되는 경우는 4개이다. 이에, 양자화 포인트 세트는 실제로 사용될 4가지의 경우를 선택하여 디코더로 구현하는 데에 이용될 수 있다.

도 5(b)의 디코더는 16비트 레지스터(4×4 비트) 및 4비트의 4-to-1 MUX로 구성될 수 있다. 16비트 레지스터는 표 1의 알고리즘에 의해 선택된 양자화 포인트 세트(즉, 2비트 로그 값)를 저장하고 양자화 입자값에 따라 동일한 계층 또는 채널의 모든 MAC 간에 공유할 수 있다. 실시예에서, 디코더는 양자화되는 데이터 비트들의 양자화 스케일을 동일한 채널 또는 레이어 등의 그룹에 적용할 수 있도록 MAC을 구성하는 시프터의 출력을 제어할 수 있다. 따라서, 데이터 비트들에 대한 양자화 스케일은 그룹별로 다르게 설정될 수 있다.

이후 디코더는 단순한 MUX로 축소되지만 이 MUX도 MAC 어레이의 데이터 흐름에 따라 MAC 어레이 내의 여러 MAC 간에 공유될 수 있다. 요약하면, 최적화된 MAC의 하드웨어 비용은 매우 적다. 2개의 4-to-1 MUX, 1개의 덧셈기, 1개의 누산기, 1개의 4-to-1 MUX에 의해 구현될 수 있으며, 하드웨어 데이터 흐름에 따라 여러 MAC 간에 공유되어 양자화될 수 있다.

도 6은 실시예에서, 스케일 파라미터를 도출하는 방법을 수행하기 위한 장치의 구성을 도시한 도면이다.

일 실시예에 따른 장치(600)는 프로세서(610), 메모리(630) 및 통신 인터페이스(650)를 포함할 수 있다. 프로

세서(610), 메모리(630) 및 통신 인터페이스(650)는 통신 버스(605)를 통해 서로 통신할 수 있다.

[0128] 일 실시예에 따른 프로세서(610)는 양자화 스케일링을 위한 스케일 파라미터 계산 방법을 수행한다. 이에, 장치(600)는 프로그램에 의해, 양자화된 유니버설 세트 중 일부 데이터들을 미리 정해진 방법으로 선택하는 단계; 선택된 일부 데이터들에 대해서 근사 가중치로 양자화를 수행하는 단계; 양자화된 데이터에 대해서 양자화 오차를 계산하는 단계; 및 계산된 양자화 오차를 최소화하기 위한 스케일 파라미터 값을 도출하는 단계를 실행할 수 있다.

[0129] 메모리(630)는 휘발성 메모리 또는 비 휘발성 메모리일 수 있고, 프로세서(610)는 프로그램을 실행하고, 장치(600)를 제어할 수 있다. 프로세서(610)에 의하여 실행되는 프로그램 코드는 메모리(630)에 저장될 수 있다. 장치(600)는 입출력 장치(미도시)를 통하여 외부 장치(예를 들어, 퍼스널 컴퓨터 또는 네트워크)에 연결되고, 데이터를 교환할 수 있다. 장치(600)는 스마트 폰, 태블릿 컴퓨터, 랩톱 컴퓨터, 데스크톱 컴퓨터, 텔레비전, 웨어러블 장치, 보안 시스템, 스마트 홈 시스템 등 다양한 컴퓨팅 장치 및/또는 시스템에 탑재될 수 있다.

[0131] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 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. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

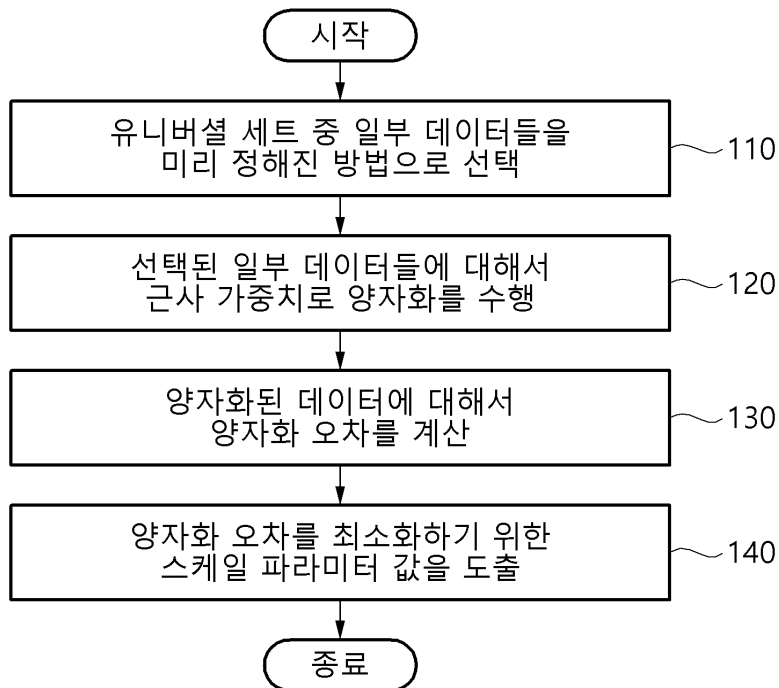
[0133] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 매체에 저장될 수 있다.

[0135] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 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.

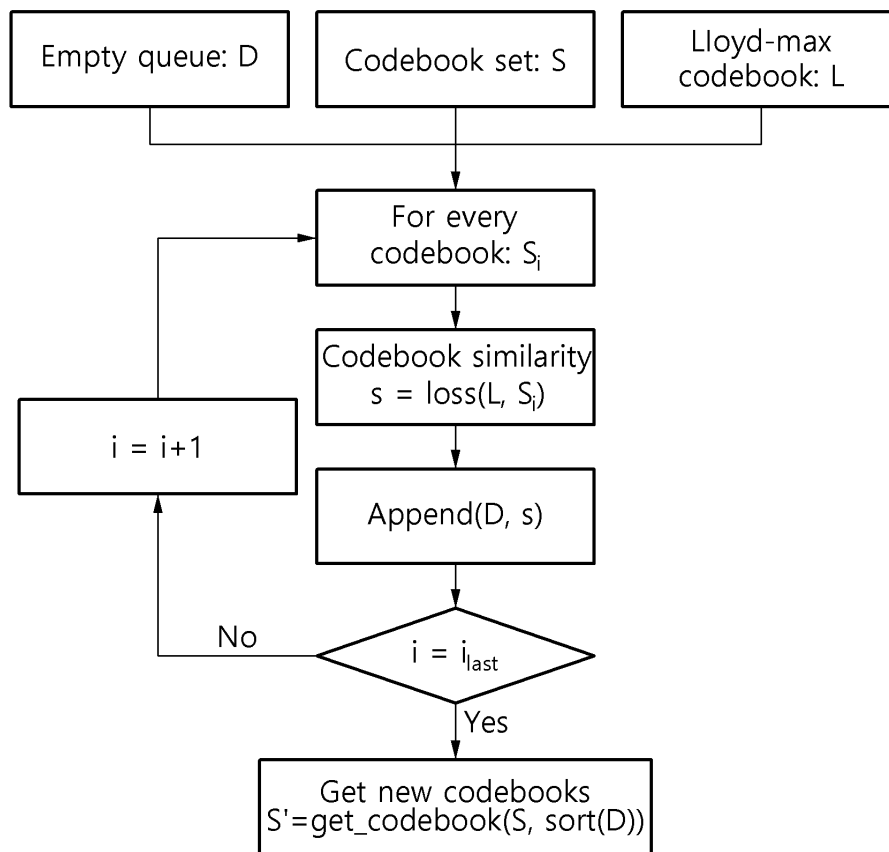
[0136] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

도면

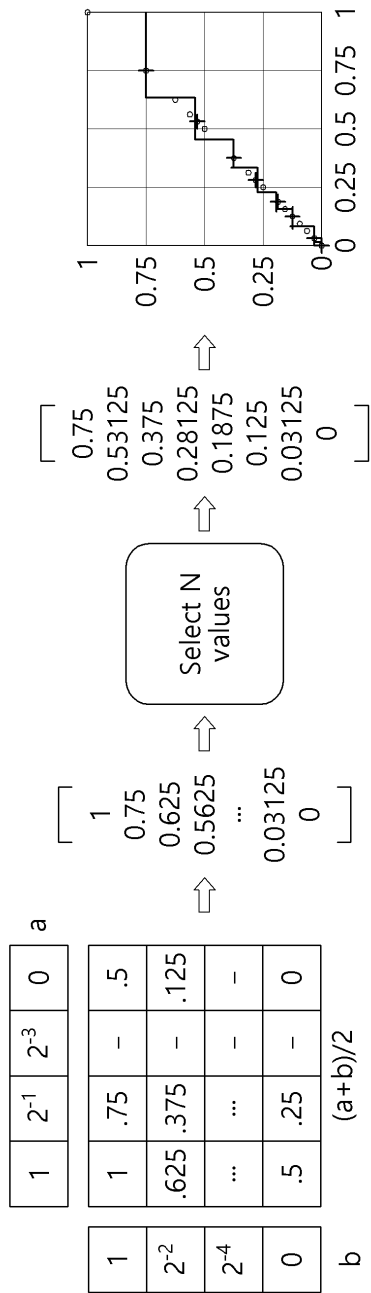
도면1



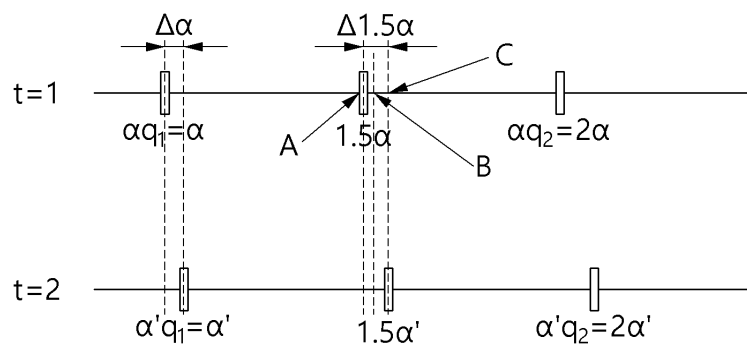
도면2



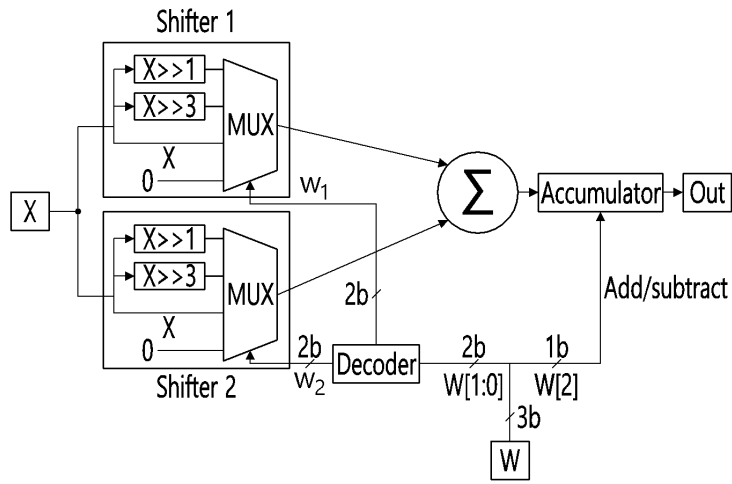
도면3



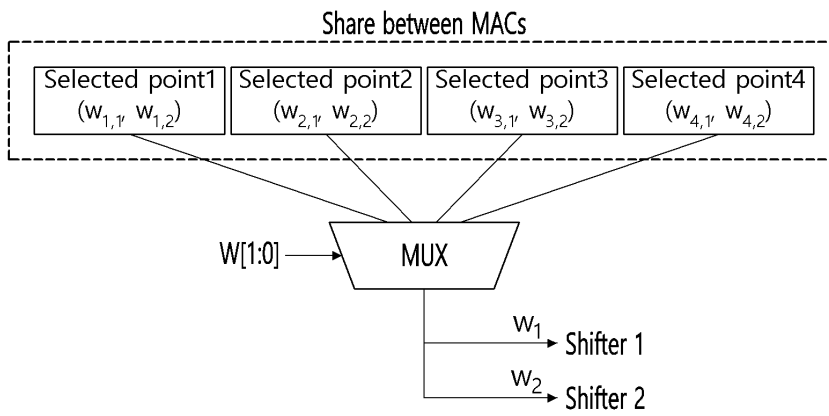
도면4



도면5



(a)



(b)

도면6

