(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0069693
(43) 공개일자 2023년05월19일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) *G06N 3/04* (2023.01)
(52) CPC특허분류
G06N 3/08 (2023.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2021-0155942
(22) 출원일자 2021년11월12일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
장준우
부산광역시 해운대구 해운대로 265, 101동 1204호 (재송동, 센텀천일스카이원)
박재우
울산광역시 울주군 언양읍 유니스트길 50
(뒷면에 계속)
(74) 대리인
특허법인 무한

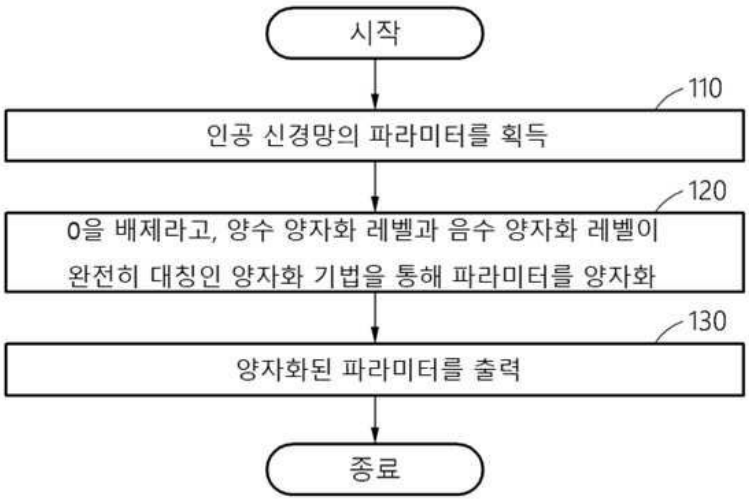
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **인공 신경망의 양자화 방법 및 이를 수행하는 장치**

(57) 요 약

실시예는, 인공 신경망의 양자화 방법 및 이를 수행하는 장치에 대한 것이다. 실시예에 따른 양자화 방법은, 상기 인공 신경망의 파라미터를 획득하는 단계; 0을 양자화 레벨에서 배제하고, 적어도 하나의 양수 양자화 레벨과 적어도 하나의 음수 양자화 레벨이 서로 완전히 대칭인 양자화 기법을 이용하여, 상기 파라미터를 양자화하는 단계; 상기 양자화된 파라미터를 출력하는 단계를 포함할 수 있다.

대 표 도 - 도1



(72) 발명자
아심 파이즈
울산광역시 울주군 언양읍 유니스트길 50
이종은
울산광역시 울주군 언양읍 유니스트길 50

명 세 서

청구범위

청구항 1

인공 신경망의 양자화 방법에 있어서,

상기 인공 신경망의 파라미터를 획득하는 단계;

0을 양자화 레벨에서 배제하고, 적어도 하나의 양수 양자화 레벨과 적어도 하나의 음수 양자화 레벨이 서로 완전히 대칭인 양자화 기법을 이용하여, 상기 파라미터를 양자화하는 단계;

상기 양자화된 파라미터를 출력하는 단계

를 포함하는,

양자화 방법.

청구항 2

제1항에 있어서,

상기 파라미터를 양자화하는 단계는,

하기 수학식에 기초하여 상기 파라미터를 양자화하는 단계

를 포함하는,

양자화 방법.

수학식:

$$v_{bar} = \text{clamp}(\text{round}(v/s + 0.5) - 0.5, -2^{b-1} + 0.5, 2^{b-1} - 0.5)$$

여기서, v는 상기 파라미터이고, s는 상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈이고, b는 미리 정해진 양자화 비트 수임

청구항 3

제1항에 있어서,

상기 파라미터는,

양자화 인식 훈련(quantization-aware training)을 통해 학습되어 결정되는,

양자화 방법.

청구항 4

제1항에 있어서,

상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈는,

상기 파라미터와 함께 조인트 학습에 의해 결정되는,

양자화 방법.

청구항 5

제1항에 있어서,

상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈는 하기 수학식에 기초하여 결정되는,

양자화 방법.

수학식:

$$\frac{\partial v}{\partial s} = \begin{cases} -\frac{v}{s} + \left(\lceil \frac{v}{s} \rceil - 0.5\right) & if\ -Q_n < \left(\lceil \frac{v}{s} \rceil - 0.5\right) \leq Q_p \\ Q_n & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \leq -Q_n \\ Q_p & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \geq Q_p \end{cases}$$

여기서, v는 상기 파라미터이고, s는 상기 스텝 사이즈이고, $-Q_n$은 최저 양자화 레벨이고, Q_p는 최대 양자화 레벨임

청구항 6

제1항에 있어서,

상기 양자화된 파라미터 기반의 MAC 연산은,

XNOR-Popcount 구조를 가지는 BNN(Binary Neural Network) 하드웨어에 의해 수행되는,

양자화 방법.

청구항 7

제1항에 있어서,

상기 양자화된 파라미터는,

0을 기준으로 대칭인 형태이며, 양수와 음수에 균등하게 할당되는,

양자화 방법.

청구항 8

하드웨어와 결합되어 제1항 내지 제7항 중 어느 하나의 항의 방법을 실행시키기 위하여 매체에 저장된 컴퓨터 프로그램.

청구항 9

인공 신경망의 양자화 방법을 위한 장치에 있어서,

무선 통신을 위한 통신부;

하나 이상의 프로세서;

메모리; 및

상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그

램을 포함하고,

상기 프로그램은,

상기 인공 신경망의 파라미터를 획득하는 단계;

0을 양자화 레벨에서 배제하고, 적어도 하나의 양수 양자화 레벨과 적어도 하나의 음수 양자화 레벨이 서로 완전히 대칭인 양자화 기법을 이용하여, 상기 파라미터를 양자화하는 단계; 및

상기 양자화된 파라미터를 출력하는 단계

를 포함하는,

장치.

청구항 10

제9항에 있어서,

상기 파라미터를 양자화하는 단계는,

하기 수학식에 기초하여 상기 파라미터를 양자화하는 단계

를 포함하는,

장치.

수학식:

$$v_{bar} = clamp(round\ (v/s + 0.5) - 0.5, -2^{b-1} + 0.5, 2^{b-1} - 0.5)$$

여기서, v는 상기 파라미터이고, 상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈이고, b는 미리 정해진 양자화 비트 수임

청구항 11

제9항에 있어서,

상기 파라미터는,

양자화 인식 훈련(quantization-aware training)을 통해 학습되어 결정되는,

장치.

청구항 12

제9항에 있어서,

상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈는,

상기 파라미터와 함께 조인트 학습에 의해 결정되는,

장치.

청구항 13

제9항에 있어서,

상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈는 하기 수학식에 기초하여 결정되는,

장치.

수학식:

$$\frac{\partial v}{\partial s} = \begin{cases} -\frac{v}{s} + \left(\lceil \frac{v}{s} \rceil - 0.5\right) & if\ -Q_n < \left(\lceil \frac{v}{s} \rceil - 0.5\right) \leq Q_p \\ Q_n & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \leq -Q_n \\ Q_p & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \geq Q_p \end{cases}$$

여기서, v는 상기 파라미터이고, s는 상기 스텝 사이즈이고, $-Q_n$은 최저 양자화 레벨이고, Q_p는 최대 양자화 레벨임

청구항 14

제9항에 있어서,

상기 양자화된 파라미터 기반의 MAC 연산은,

XNOR-popcount 구조를 가지는 BNN(Binary Neural Network) 하드웨어에 적용되는,

장치.

청구항 15

제9항에 있어서,

상기 양자화된 파라미터는,

0을 기준으로 대칭인 형태이며, 양수와 음수에 균등하게 할당되는,

장치.

발명의 설명

기 술 분 야

[0001] 실시예들은, 인공 신경망의 양자화 방법 및 이를 수행하는 장치에 관한 것이다.

배 경 기 술

[0003] 인공지능 분야에서 연산량을 줄이면서 전력 효율성을 향상시키는 방법 중 하나가 양자화(quantization) 기술이다. 양자화는, 정확하고 세밀한 단위로 표현한 입력값을 보다 단순화한 단위의 값으로 변환하는 다양한 기술을 포괄적으로 의미하는 용어이다. 양자화 기술은 정보를 표현하는 데 필요한 비트의 수를 줄이기 위한 것이다.

[0004] 일반적으로 인공 신경망은 활성 노드, 노드 간의 연결, 각 연결과 관련한 가중치 매개변수(weight parameter)로 구성되는데, 여기서 양자화되는 대상은 가중치 매개변수와 활성 노드 연산. 신경망을 하드웨어에서 진행하면 곱셈 및 덧셈 연산을 수백만 회 실행한다.

[0005] 만약 양자화된 매개변수로 저비트(lower-bit)의 수학 연산을 수행하고, 신경망의 중간 계산값도 함께 양자화한다면, 연산 속도는 빨라지고 성능도 향상된다. 더불어, 인공 신경망을 양자화하게 되면, 메모리 액세스를 줄이고 연산 효율성도 높일 수 있으므로 전력 효율성도 향상될 수 있다.

[0006] 그러나, 양자화로 인해 인공 신경망의 정확도가 떨어질 수 있다. 이에, 정확도에 영향을 주지 않으면서 연산 효율 및 전력 효율이 높아 지도록 양자화 기술이 발전하고 있다.

[0007] 이와 관련하여 국제특허 WO2020/248424(Method for determining quantization parameters in neural network and related products)에서는 인공 신경망에서 양자화 파라미터를 결정하는 방법을 제시한다.

발명의 내용

해결하려는 과제

[0009] 실시예에 따른 발명은, 양수 및 음수 각각에 균등한 양자화 파라미터를 제공하며, 0을 중심으로 대칭 구조를 가지는 양자화 방법을 제공하고자 한다.

과제의 해결 수단

[0011] 인공 신경망의 양자화 방법에 있어서, 상기 인공 신경망의 파라미터를 획득하는 단계; 0을 양자화 레벨에서 배제하고, 적어도 하나의 양수 양자화 레벨과 적어도 하나의 음수 양자화 레벨이 서로 완전히 대칭인 양자화 기법을 이용하여, 상기 파라미터를 양자화하는 단계; 상기 양자화된 파라미터를 출력하는 단계를 포함하는, 양자화 방법이 제공될 수 있다.

[0012] 상기 파라미터를 양자화하는 단계는, 하기 수학식에 기초하여 상기 파라미터를 양자화하는 단계를 포함할 수 있다.

[0013] 수학식:

[0014] vbar = clamp(round (v/s + 0.5) - 0.5, - 2b-1 + 0.5, 2b-1 - 0.5)

[0015] 여기서, v는 상기 파라미터이고, s는 상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈이고, b는 미리 정해진 양자화 비트 수임

[0016] 상기 파라미터는, 양자화 인식 훈련(quantization-aware training)을 통해 학습되어 결정될 수 있다.

[0017] 상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈는, 상기 파라미터와 함께 조인트 학습에 의해 결정될 수 있다.

[0018] 상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈는 하기 수학식에 기초하여 결정될 수 있다.

[0019] 수학식:

[0020]
$$\frac{\partial v}{\partial s} = \begin{cases} -\frac{v}{s} + \left(\lceil \frac{v}{s} \rceil - 0.5\right) & if\ -Q_n < \left(\lceil \frac{v}{s} \rceil - 0.5\right) \leq Q_p \\ Q_n & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \leq -Q_n \\ Q_p & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \geq Q_p \end{cases}$$

[0021] 여기서, v는 상기 파라미터이고, s는 상기 스텝 사이즈이고, -Qn은 최저 양자화 레벨이고, Qp는 최대 양자화 레벨임

[0022] 상기 양자화된 파라미터 기반의 MAC 연산은, XNOR-Popcount 구조를 가지는 BNN(Binary Neural Network) 하드웨어에 의하여 수행될 수 있다.

[0023] 상기 양자화된 파라미터는, 0을 기준으로 대칭인 형태이며, 양수와 음수에 균등하게 할당될 수 있다.

[0024] 인공 신경망의 양자화 방법을 위한 장치에 있어서, 무선 통신을 위한 통신부; 하나 이상의 프로세서; 메모리; 및 상기 메모리에 저장되어 있으며 상기 하나 이상의 프로세서에 의하여 실행되도록 구성되는 하나 이상의 프로그램을 포함하고, 상기 프로그램은, 상기 인공 신경망의 파라미터를 획득하는 단계; 학습을 통해 상기 인공 신경망의 양자화 구간을 결정하기 위한 스텝 사이즈를 획득하는 단계; 및 미리 정해진 양자화 비트 수 및 상기 스텝 사이즈에 기초하여 상기 파라미터를 양자화하는 단계를 포함하는, 장치가 제공될 수 있다.

발명의 효과

[0026] 실시예에 따른 발명은, 양수 및 음수 각각에 균등한 양자화 파라미터를 제공하며, 0을 중심으로 대칭 구조를 가지는 양자화 방법을 제공할 수 있다.

도면의 간단한 설명

[0028] 도 1은 실시예에서, 인공 신경망의 양자화 방법을 설명하기 위한 흐름도이다.

도 2는 실시예에서, 양자화 파라미터에 대한 그래프이다.

도 3은 실시예에서, 양자화를 위한 장치를 설명하기 위한 블록도이다.

도 4a는 실시예에서, 양자화 레벨에 따른 구간의 정규 분포를 나타낸 그래프이다.

도 4b, c는 실시예에서, CLQ와 실시예의 양자화 방법에 실제 데이터가 매핑되는 확률을 도시한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0029] 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.

[0030] 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.

[0031] 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.

[0033] 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.

[0034] 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.

[0035] 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.

[0037] 인공 신경망의 가중치 파라미터를 양자화하기 위해서 일반적으로 $[-2^{(b-1)}, 2^{(b-1)-1}]$에 매핑되는 대칭 양자화기를 사용할 수 있다. 여기서, b는 양자화 비트 수이다. 양자화된 인공 신경망(Quantized neural network, QNN)은 3비트 이내의 정밀도가 낮은 양자화 시에 성능 저하가 발생한다. 일반적인 양자화 방식은 양수 및 음수 양자화 수준을 동일하지 않게 할당하는데(예컨대, -1, 0, 1, 2 등), 양수 및 음수의 비대칭으로 인해 정밀도가 낮은 양자화 수준에서 오류 및 성능 저하가 발생할 수 있다.

[0038] 인공 신경망의 구현 시 해당 노드와 그 연결망의 모델은 추론과 학습에서 가중치의 곱셈 값을 합산하여 하나의 뉴런에 전달하는 수많은 MAC(multiplier-accumulator) 연산들과 활성화 함수에서의 곱셈을 통해 이루어진다. MAC 연산들은 인공 신경망의 크기에 비례하여 그 크기가 결정되며, 또한, MAC에 필요한 피연산자의 데이터와 출

력 데이터는 인공 신경망이 구현되는 메모리에 저장된다.

[0039] 인공신경망 구현에서는 MAC 연산기와 메모리가 하드웨어 형태로 존재한다. 좁은 의미로는 이들 MAC 연산과 메모리가 하드웨어로 매핑되어 병렬 형태로 구현되는 것을 인공 신경망의 하드웨어 형태 구현이라고 볼 수 있으나, MAC 연산에 사용되는 곱셈기와 덧셈기의 효율을 높이거나 메모리의 사용량을 줄일 수 있다.

[0040] 한편, BNN(Binary Neural Network, 이진 신경망)은 심층 신경망(Quantized neural networks, QNN)의 메모리 및 계산 비용을 높이기 위한 방법으로 제시되는 개념이다. 이진 신경망은 가중치 및 활성화 텐서의 각 값을 +1 및 -1로 양자화하여 1bit만으로 표현할 수 있으나 예측 정확도는 상대적으로 낮다.

[0041] 이진 신경망의 하드웨어는, 논리 연산인 XNOR 연산을 통해 곱셈을 구현할 수 있으며 레지스터 내의 1로 설정된 비트의 수를 알 수 있는 Pop-Count 명령을 통해 누적 덧셈을 구현할 수 있다. 이진 신경망은 실수나 정수 곱셈, 덧셈이 필요하지 않게 됨으로 연산 속도를 개선할 수 있다. 또한 기존 32Bit에서 1Bit로 줄어들기 때문에 이론적으로 메모리 대역폭이 32배 증가하게 된다.

[0042] 이진 신경망은 입력과 가중치를 모두 1Bit로 변환 후 XNOR 연산을 한다. XNOR 연산 결과에 근사 값을 곱하여 32Bit에서 1Bit로의 변환으로 인한 손실을 보정할 수 있다.

[0043] 실시예에 따라, BNN 하드웨어에서 비트 연산을 사용하여 심층 신경망에 대해 효율적인 하드웨어를 구현이 가능한 양자화 방법을 제공하고자 한다.

[0045] 도 1은 실시예에서, 인공 신경망의 양자화 방법을 설명하기 위한 흐름도이다.

[0046] 단계(110)에서 장치는, 인공 신경망의 파라미터를 획득한다.

[0047] 실시예에 따른 양자화 방법은 파라미터 간에 균일한 구간을 가지면서 양수면과 음수면 사이의 대칭 구조를 가지고 0을 양자화 레벨로 포함하지 않도록 할 수 있다. 즉, 0을 양자화 레벨에서 배제하고, 양수의 양자화 레벨들과 음수 양자화 레벨들이 서로 완전히 대칭 구조를 가질 수 있다. 예컨대, {-1.5, -0.5, 0.5, 1.5}와 같이 분수 레벨로 양자화될 수 있고, {-3, -1, 1, 3}과 같이 정수로 양자화 하기 위해 2로 양자화 구간을 위한 스텝 사이즈가 결정될 수 있다.

[0048] 일반적인 선형 양자화(Conventional Linear Quantization, CLQ)의 파라미터 레벨은 비트 수에 따라 [-2^(b-1), 2^(b-1)-1]으로 표현될 수 있다. 예컨대, 2비트 인 경우, {-2, -1, 0, 1}으로 표현될 수 있다. 또는, 양극과 음극 사이의 비대칭을 반대로 결정할 수도 있다.

[0049] Reduced Symmetric Quantization(RSQ)의 경우, 양자화 파라미터를 실시예의 수준에 대비하여 하나 덜 사용하여 L=-2b-1+1 및 U=2b-1-1 수준으로 예컨대 {-1, 0, 1}로 양자화를 수행하며, 0을 기준으로 완벽한 대칭을 이룰 수 있다. 이러한 양자화 방법은 양자화 수준이 적어져 성능이 저하될 수 있다.

[0050] Extended Symmetric Quantization(ESQ)는 하나 이상의 양자화 수준을 사용하여 0을 대칭으로 이루는 형태를 보이며, 2bit 이상을 요구될 수 있다. L=-2b-1 및 U=2b-1 수준으로 양자화되며, 예를 들어, {-2, -1, 0, 1, 2}로 양자화될 수 있다.

[0051] Non-Uniform Symmetric Quantization(NSQ)는 2b 양자화 레벨이 0을 포함하지 않는 대칭 형태를 포함하며, 예컨대 {-2, -1, 1, 2}로 양자화하는 방법을 제시하나 양자화 레벨 간의 간격이 동일하지 않다.

[0052] 단계(120)에서 장치는, 0을 양자화 포인트에서 배제하고, 적어도 하나의 양수 양자화 포인트와 적어도 하나의 음수 양자화 포인트가 서로 완전히 대칭인 양자화 기법을 이용하여 파라미터를 양자화한다.

[0053] 실시예에서, 인공 신경망의 학습 시, 파라미터 및 파라미터의 양자화 구간에 대해 함께 훈련이 이루어질 수 있다. 실시예에 따른 학습 방법은 선형 양자화를 위해 개발된 다양한 훈련 방법이 적용될 수 있다. 양자화된 파라미터에 대한 학습을 위해 양자화 인식 훈련이 적용될 수 있다. 예를 들어, LSQ와 동일한 방식으로 양자화 구간이 훈련될 수 있다.

[0054] 실시예에서, 이러한 대칭 형태의 양자화 파라미터를 학습하기 위해 아래의 수학식 1과 같은 미분 공식을 이용할 수 있다.

[0055] [수학식 1]

$$\frac{\partial v}{\partial s} = \begin{cases} -\frac{v}{s} + \left(\lceil \frac{v}{s} \rceil - 0.5\right) & if\ -Q_n < \left(\lceil \frac{v}{s} \rceil - 0.5\right) \le Q_p \\ Q_n & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \le -Q_n \\ Q_p & if\ \left(\lceil \frac{v}{s} \rceil - 0.5\right) \ge Q_p \end{cases}$$

[0056]

[0057] 경사 하강법을 사용하여 양자화 구간의 스텝 사이즈 s를 최적하기 위해 상기의 수학식 1과 같은 미분 공식을 사용할 수 있다. 여기서, v는 입력 값이고, Qn은 양자화 구간의 최소 값의 절대값이고, Qp는 양자화 구간의 최대 값을 의미한다.

[0058] 경사 하강법은 실함수의 기울기 변화를 통해 손실 함수를 줄이는 방법으로, 초기 시점에 대한 기울기를 구하여 기울기의 반대 방향으로 이동하는 과정을 통해 기울기를 수렴시키는 것으로 오차를 줄이는 과정을 포함할 수 있다. 실시예에서, 수렴시킨 손실 기울기를 계산할 수 있다. 스텝 사이즈의 기울기는 기울기의 스케일링과 유사하게 $g = 1/\sqrt{N_W 2^p}$ 로 스케일링될 수 있다. 여기서, g는 스텝 사이즈의 스케일링이고, N_w는 양자화 파라미터의 수이고, p는 비트 폭(bit-width)를 의미한다.

[0059] 실시예에서, 가중치는 $2\langle|v|\rangle/\sqrt{Q}$ 로 초기화될 수 있다. 여기서, <.>은 분포의 평균에 대해 표기하는 방법으로 사용될 수 있다.

[0060] 실시예에서, 훈련을 통해 획득한 양자화 방법은 아래의 수학식 3과 같이 나타낼 수 있다.

[0061] [수학식 3]

$$\dot{v} = \left\lfloor \frac{v}{s} + 0.5 \right\rfloor - 0.5$$
$$\bar{v} = \mathrm{clip}\left(\dot{v}, -Q, Q\right)$$
$$\hat{v} = \bar{v} \times s$$

[0062]

[0063] 여기서, clip() 함수는, clip(리스트, 최소값, 최대값)으로 나타내고, 리스트 안에 있는 값들을 최소값과 최대값 사이의 값들로 변환시킨 array를 리턴할 수 있고, clip(x; a; b) = min(max(x; a); b)로 나타낼 수 있다.

[0064] v는 임의의 입력 값이고, s는 스텝 사이즈를 의미한다. 앞서 훈련을 통해 Q= 2^{b-1}-0.5로 결정될 수 있고, b는 양자화 밀도, 즉 미리 정해진 비트 수를 의미한다. v는 정수가 아니며, 그럼에도 불구하고 실시예에 따른 b비트의 양자화 방법을 통해 보다 정확히 표현될 수 있다. $\bar{v}$ 는 b비트 하드웨어에서 계산된 값을 나타내고, $\hat{v}$ 는 훈련 목적으로 정의되고 사용되는 v의 축소 버전에 해당한다. 실시예에 따른 양자화 장치는 입력 분포의 양수 및 음수에 대해 동일하게 표현될 수 있다.

[0065] 단계(130)에서 장치는, 양자화된 파라미터를 출력한다.

[0066] 실시예에서, 양자화된 파라미터는 앞서 설명된 바와 같이 0을 양자화 레벨에서 배제하고, 양수의 양자화 레벨들과 음수 양자화 레벨들이 서로 완전히 대칭 구조를 가질 있다.

[0067] 일실시예에서, 양자화 방법은 효율적인 하드웨어 및 소프트웨어로 구현될 수 있다. 해당 사항에 대해서는 이후 자세히 설명하도록 한다.

[0069] 도 2는 실시예에서, 양자화 파라미터에 대한 그래프이다.

[0070] 도 2(a)는 일반적인 선형 양자화 방법과 실시예에 따른 양자화 방법에 따른 결과를 도시하며, 도 2(b)는 실시예에 따른 양자화 파라미터의 스텝 사이즈에 대한 기울기를 도시한 그래프이다.

[0071] 도 2(a)의 그래프는2bit의 부호 데이터를 양자화한 실시예에 대한 것이다. 도 2(a)에 도시된 바와 같이, 동일한 스텝 사이즈를 가지는 선형 양자화 방법에 대해서 0 주변의 값을 양자화한 결과에 있어서 차이가 있으며, 실시예에 따르면 0을 기준으로 상방 및 하방의 간격이 동일한 형태로 양자화가 가능하다. 입력이 정수인 곳을 제

외하고, 모든 입력 값에 대해서 반올림 연산자를 적용할 수 있다.

[0072] 실시예에 따른 그래프는 앞서 도 1을 통해 설명된 경사 하강법을 통해 최적화된 스텝 사이즈에 의해 결정된 양자화 구간에 기초하여 도시되어 있다. 도 2(b)에 도시된 바에 따르면, 실시예에 따른 양자화 방법에 의해 양자화 구간 내에 포함되는 입력 값에 대해서 일정한 기울기 내에서 양자화 결과를 획득할 수 있음을 알 수 있다.

[0074] 실시예에 따른 양자화 방법은, 낮은 비트의 양자화된 가중치, 예컨대 3비트 이하에서 최대 엔트로피에 가까운 효율의 하드웨어 및 소프트웨어로 구현될 수 있다.

[0075] 대표적으로 이진 신경망이 있다. 이진 신경망은 앞서 기재한 바와 같이, 기존 인공신경망의 속도를 대폭적으로 상승시키고 인공신경망 모델의 메모리 용량을 대폭 줄일 수 있다는 점에서 획기적인 방식이나, 기존의 부동소수점인 웨이트와 활성화 함수를 -1과 1로 표현하기 때문에 정보의 손실이 발생한다는 단점이 있다. 이러한 정보 손실은 결과적으로 정확도 저하로 이어지며, 사물을 인식하거나 물건을 탐지하는 데 있어 성능 저하를 가져올 수 있다.

[0076] 예를 들어, 1.4와 0.2 둘 다 양수이기 때문에 1로 매핑하게되는 경우, 크기가 7배나 차이나는 두 값들이 같은 값으로 매핑된다면 양자화 에러(quantization error)가 매우 커질 수 있다. 이에, 종래 이진 인공 신경망에서는 스케일 팩터(scale factor)를 이용하여 데이터들의 크기(magnitude)를 고려한 이진 양자화를 수행하였다. 그러나, 스케일 팩터 역시 학습을 통해 결정해야한다는 제한이 있다.

[0077] 실시예에의 양자화 방법은 이진 신경망 하드웨어에 효율적으로 매핑될 수 있다. 이진 신경망을 통해 이진 가중치, 예컨대 +1 및 -1의 가중치 파라미터가 적용될 수 있다. 이러한 가중치 파라미터를 적용함으로써 하드웨어로 구현 시 곱셈기를 제거할 수 있고, 신경망 구조를 간소화함으로써 빠른 연산 속도를 제공할 수 있다.

[0078] 실시예에서, 이진 신경망에서는 이진 인코딩 시, 일반적인 2' complement 방법 대신 0을 -1로 해석할 수 있다. 예를 들어, 010=-1, 1, -1로 인코딩하고, 해당 입력은 -(2^2)+(2^1)-(2^0)=-3으로 표현될 수 있다.

[0079] 실시예에 따른 이진 신경망은 XNOR-Popcount를 사용하여 MAC 연산을 구현할 수 있다. 이러한 하드웨어 구현은 부호 확장을 위한 추가 비트를 제거하는 데에 용이하다.

[0080] 이하에서는, 2bit의 부호 데이터를 XNOR-Popcount하는 실시예에 대해서 설명하도록 한다.

[0081] 2bit 이진수 x = x1 x0은 정수를 나타내며, X = 2*(-1)^x1 + (-1)^x0로 표현될 수 있다. 2bit 이진수 y = y1 y0은 정수를 나타내며, Y = 2*(-1)^y1 + (-1)^y0로 표현될 수 있다.

[0082] 이러한 X와 Y의 곱은 XY = 4*(-1)^(x1+y1) + 2*(-1)^(x0+y1) + 2*(-1)^(x1+y0) + (-1)^(x0+y0)로 나타낼 수 있다.

[0083] 한편, 1bit의 이진수 x, y, z = xnor(x, y)의 경우 Z = (-1)^z, X = (-1)^x, Y = (-1)^y이므로 XY = - Z로 나타낼 수 있다. 해당 식을 풀어서 계산하면 XY = (-1)^(x+y) = (-1)^xor(x, y) = (-1)^[1+xnor(x, y)] = -1*(-1)^xnor(x, y) = - Z로 나타나고,

[0084] 또한 실시예에 따른 양자화 인코딩에서 Z = 2*z-1이며, 결국 X Y = 1 - 2z = 1 - 2 xnor(x,y)로 표현될 수 있다.

[0085] 정리하면, X Y는 다음과 같이 XNOR-Popcount를 이용하여 다시 나타낼 수 있다.

[0086] XY = 4*(1-2xnor(x1, y1)) + 2*(1-2xnor(x0, y1)) + 2*(1-2xnor(x1, y0)) + (1-2xnor(x0, y0)) = 9 - 8xnor(x1, y1) - 4(xnor(x0, y1) + xnor(x1, y0)) - 2 xnor(x0, y0)

[0087] 따라서, XY 곱은 4개의 XNOR 연산, 3개의 시프트 연산(2bit) 및 4개의 더하기 연산을 사용하여 계산될 수 있다. 상수 항을 바이어스(Bias) 항으로 결합하고 모든 항을 2로 나눔으로써 추가적으로 단순화될 수 있다. 이 경우 4개의 XNOR, 2개의 시프트 및 3개의 추가만 필요하게 된다.

[0088] 참고로, 대안으로 2's complement 인코딩이 사용될 수 있다. 이러한 경우, X Y를 계산하는 좋은 방법으로 보다 복잡한 부호가 있는 승수(multiplier)를 사용하는 것이다.

[0089] 혹은, 다음과 같이 2-초과를 포함하는 오프셋 바이너리를 사용할 수 있다. X' = X + 2 ≥ 0, Y' = Y + 2 ≥ 0 이며, 여기서 X와 Y는 x와 y에 대한 일반적인 2's complement로 해석할 수 있다. 따라서 X' 및 Y'는 부호 없는

버전(2-초과 코드)이다.

[0090] 이에, X Y 곱은 X Y = (X'-2)(Y'-2) = X'Y'-2(X'+Y') + 4로 계산될 수 있다. 해당 식은 2비트 곱셈, 하나의 시프트(3비트) 및 세 개의 덧셈이 필요하며, 부호 없는 2bit 곱셈에 해당하는 경우, 4개의 AND 연산과 3개의 시프트 연산이 추가로 필요하게 된다.

[0091] 따라서, 양자화 인코딩은 2bit 곱셈에 더 효율적이며, 양자화 시 2bit x 2bit 곱셈과 1bit x 2bit 곱셈을 XNOR-popcount BNN 하드웨어에서 추가 하드웨어(예: 부호 있는 또는 부호 없는 곱셈기)를 추가하지 않고도 수행할 수 있다.

[0093] 도 3은 일 실시예에서, 양자화를 위한 장치를 설명하기 위한 블록도이다.

[0094] 도 3을 참조하면, 일 실시예에 따른 장치(300)는 프로세서(310), 메모리(330) 및 통신 인터페이스(350) 포함할 수 있다. 프로세서(310), 메모리(330) 및 통신 인터페이스(350)는 통신 버스(305)를 통해 서로 통신할 수 있다.

[0095] 일 실시예에 따른 프로세서(310)는 인공 신경망의 양자화 방법을 수행한다. 양자화 방법은 인공 신경망의 파라미터를 획득하는 단계; 0을 양자화 레벨에서 배제하고, 적어도 하나의 양수 양자화 레벨과 적어도 하나의 음수 양자화 레벨이 서로 완전히 대칭인 양자화 기법을 이용하여, 파라미터를 양자화하는 단계; 양자화된 파라미터를 출력하는 단계를 포함할 수 있다.

[0096] 실시예의 양자화 방법은 파라미터 간에 균일한 구간을 가지면서 양수면과 음수면 사이의 대칭 구조를 가지고 0을 양자화 레벨로 포함하지 않도록 할 수 있다. 즉, 0을 양자화 레벨에서 배제하고, 양수의 양자화 레벨들과 음수 양자화 레벨들이 0을 기준으로 서로 완전히 대칭 구조를 가지며, 양수 및 음수 각각에 양자화 레벨들이 균등하게 분포되도록 학습될 수 있다.

[0097] 실시예에서, 인공 신경망의 학습 시, 파라미터 및 파라미터의 양자화 구간에 대해 함께 훈련이 이루어질 수 있다. 실시예에 따른 학습 방법은 선형 양자화를 위해 개발된 다양한 훈련 방법이 적용될 수 있다. 양자화된 파라미터에 대한 학습을 위해 양자화 인식 훈련이 적용될 수 있다.

[0098] 일 실시예에 따른 장치(300)는 낮은 비트의 양자화된 가중치, 예컨대 3비트 이하에서 최대 엔트로피에 가까운 효율의 하드웨어 및 소프트웨어로 예컨대, XNOR-Popcount 구조를 가지는 이진 신경망 하드웨어를 통해 구현될 수 있다.

[0099] 메모리(330)는 휘발성 메모리 또는 비 휘발성 메모리일 수 있고, 프로세서(310)는 프로그램을 실행하고, 장치(300)를 제어할 수 있다. 프로세서(310)에 의하여 실행되는 프로그램 코드는 메모리(330)에 저장될 수 있다. 장치(300)는 입출력 장치(미도시)를 통하여 외부 장치(예를 들어, 퍼스널 컴퓨터 또는 네트워크)에 연결되고, 데이터를 교환할 수 있다. 장치(300)는 스마트 폰, 테블릿 컴퓨터, 랩톱 컴퓨터, 데스크톱 컴퓨터, 텔레비전, 웨어러블 장치, 보안 시스템, 스마트 홈 시스템 등 다양한 컴퓨팅 장치 및/또는 시스템에 탑재될 수 있다.

[0101] 도 4는 실시예에서, 2비트로 양자화된 양자화 구간의 확률 분포를 설명하기 위한 도면이다.

[0102] 도 4(a)는 실시예에서, 양자화 레벨에 따른 구간의 정규 분포를 나타낸 그래프이다.

[0103] 도 4(a)의 x축은 양자화 레벨을 표시하며, y축은 실제 데이터 별로 확률 분포를 나타낸다. 실시예에서, 가우시간 분포와 유사하게 나타날 수 있다.

[0104] 실시예에 따른 양자화 방법은 양자화를 통해 양자화 레벨에 따른 효율을 최대화하는 데에 유용하게 적용될 수 있다.

[0105] 만약, 데이터가 양자화될 때, 양자화 레벨마다 매핑되는 데이터가 가능한 한 균일하게 분포되어야 하는 경우, 높은 양자화 효율을 제공할 수 있으며, 또는 양자화 레벨의 분포가 데이터 분포, 예컨대 가우시안 분포와 유사하게 이루어지는 경우 높은 양자화 효율을 제공할 수 있다.

[0106] 실시예에 따른 양자화 방법은 상기의 두 가지 조건에 모두 부합하는 것이다. 예컨대, 실시예에 따라 2 bit로 양자화 하는 경우, 일반적으로 임계값 {-1; 0; 1}를 기준으로 상기의 두 가지 조건을 충족할 수 있다.

[0107] 실시예에서, 도 4(a)와 같이 데이터는 양자화 수준에 걸쳐 균일하게 분포되며, 동시에 양자화 레벨 또한 가우시안 분포에 따른다. 여기서, 도 4(a)의 가우시안 분포는 X~N(0; 1)에서 $P(0 \leq X \leq s) = 0{:}25$로 나타나는 표준 정규 분포의 CDF를 따르는 것으로 가정할 수 있다.

[0108] 도 4(b), (c)는 실시예에서, CLQ와 실시예 각각에 실제 데이터가 매핑되는 결과를 도시한 그래프이다.

[0109] 도 4(b)는 CLQ로 학습되어 결정된 양자화 레벨에 실제 데이터가 매핑되는 확률에 관한 것이며, 도 4(c)는 실시예에 따른 방법으로 학습되어 결정된 양자화 레벨에 실제 데이터가 매핑되는 확률을 도시한 것이다.

[0110] 도 4(b)에 의하면, 양자화 레벨이 (-2, -1, 0. 1)에 해당하여 각 양자화 레벨에 따른 매핑 확률이 적게는 10%, 많게는 40% 가까이 나타나 양자화 효율이 좋다 평가할 수 없으나, 도 4(c)에 의하면, 양자화 레벨 -1.5, -0.5, 0.5, 1.5 각각에 대해서 25% 전후로 매핑 확률이 비교적 균등하게 나타나는 것을 확인할 수 있다.

[0112] 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.

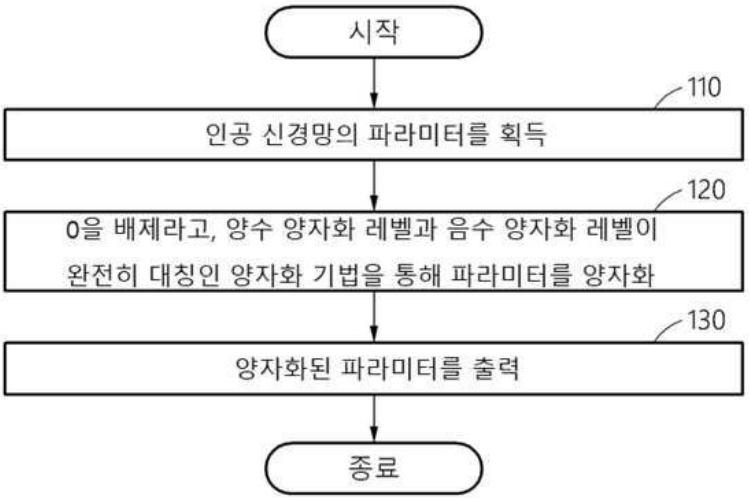
[0114] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0116] 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.

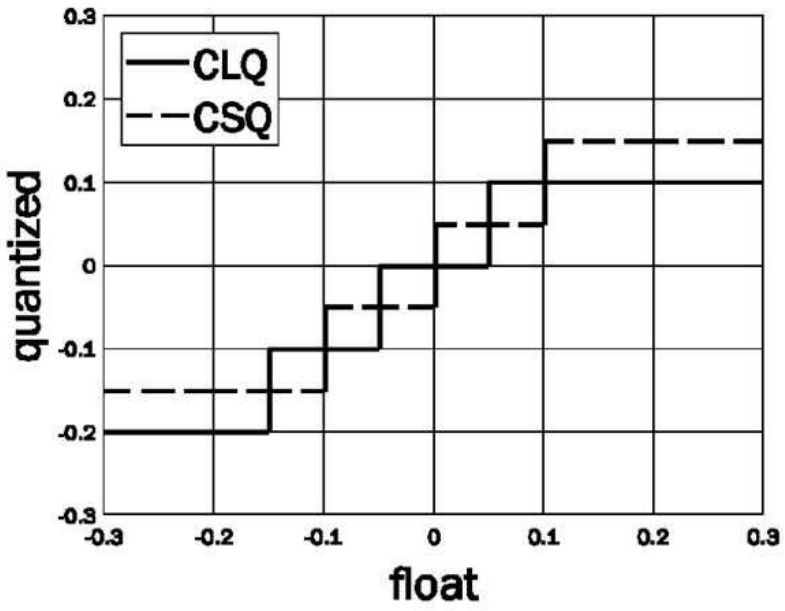
[0117] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 청구범위의 범위에 속한다.

도면

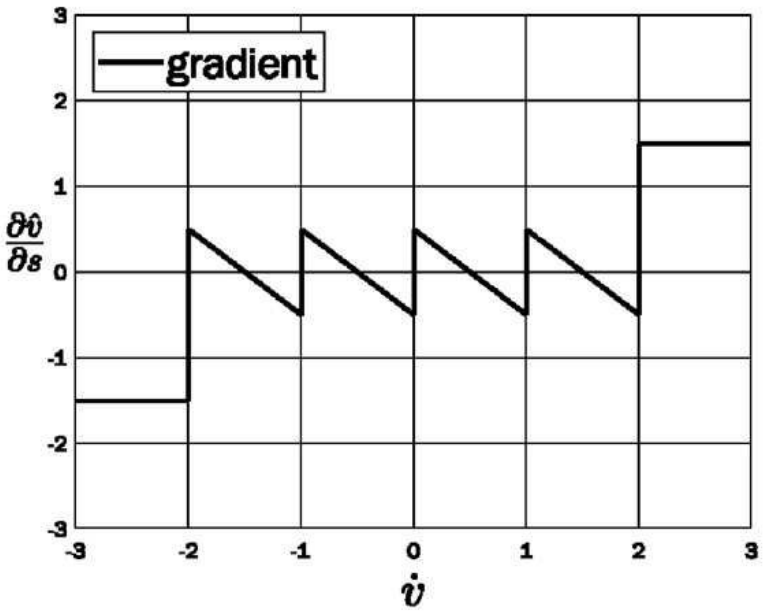
도면1


시작
110
인공 신경망의 파라미터를 획득
120
0을 배제라고, 양수 양자화 레벨과 음수 양자화 레벨이
완전히 대칭인 양자화 기법을 통해 파라미터를 양자화
130
양자화된 파라미터를 출력
종료

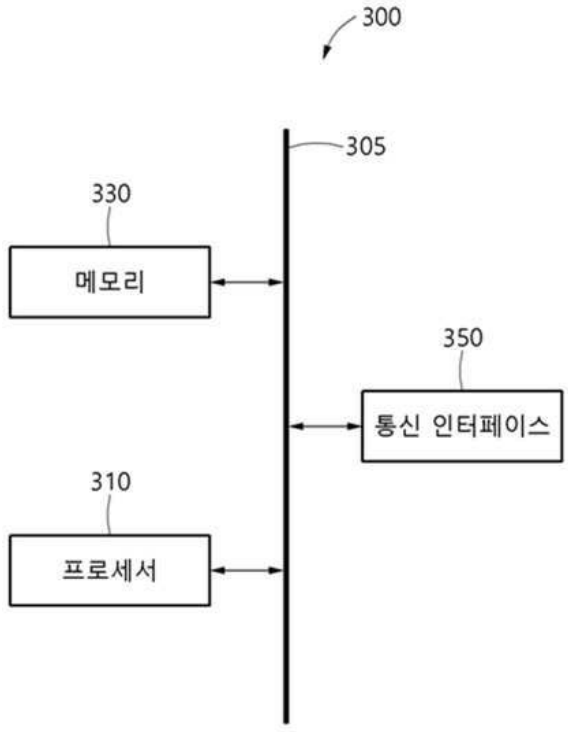
도면2a


CLQ
CSQ
quantized
float
0.3
0.2
0.1
0
-0.1
-0.2
-0.3

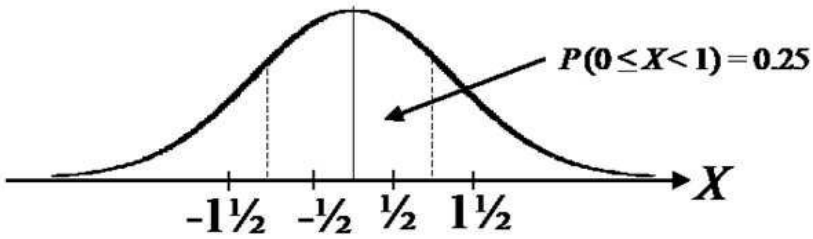
도면2b


gradient
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
$\frac{\partial \hat{v}}{\partial s}$
$\dot{v}$

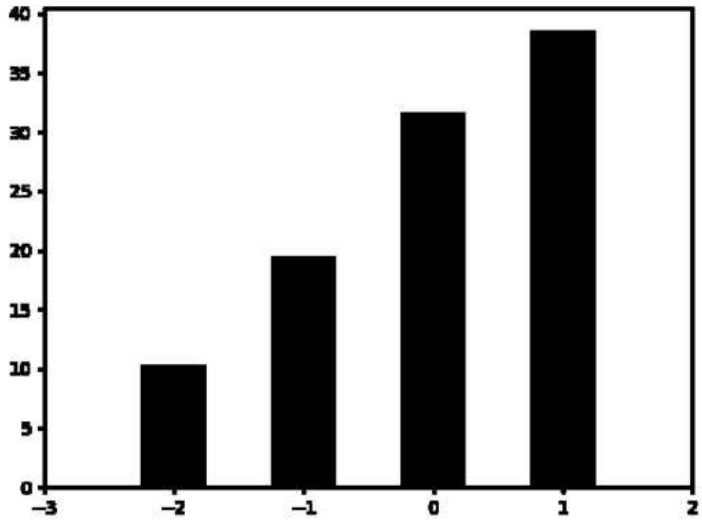
도면3


300
305
330
메모리
350
통신 인터페이스
310
프로세서

도면4a



도면4b



도면4c