(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년07월10일
(11) 등록번호 10-2684061
(24) 등록일자 2024년07월08일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
G06N 7/00 (2023.01)
(52) CPC특허분류
G06N 3/04 (2023.01)
G06N 3/08 (2023.01)
(21) 출원번호 10-2020-0135673
(22) 출원일자 2020년10월20일
심사청구일자 2020년10월20일
(65) 공개번호 10-2022-0051947
(43) 공개일자 2022년04월27일
(56) 선행기술조사문헌
KR1020190130443 A
KR1020190134965 A
KR1020200101521 A
US20200302224 A1

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
함범섭
서울특별시 강남구 압구정로61길 37, 72동 506호 (압구정동, 한양아파트)
김도형
경기도 고양시 덕양구 백양로 8, 1704동 203호(화정동, 옥빛마을17단지아파트)
이중협
서울특별시 서대문구 연희로8길 26(연희동)
(74) 대리인
민영준

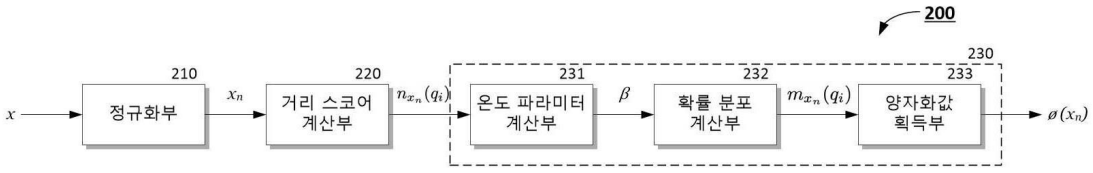
전체 청구항 수 : 총 17 항 심사관 : 유주영

(54) 발명의 명칭 **인공 신경망을 위한 양자화기 및 양자화 방법**

(57) 요 약

본 발명은 양자화 대상인 대상값을 인가받아 기지정된 범위로 정규화하여 정규화값을 출력하는 정규화부, 기지정된 다수의 양자값 중 정규화값에 가장 인접한 2개의 양자값을 선택하고, 선택된 2개의 양자값 각각과 정규화값 사이의 차에 따라 기지정된 방식으로 2개의 거리 기반 스코어를 계산하는 거리 스코어 계산부, 2개의 거리 기반 스코어에 2개의 양자값 각각을 커널 함수에 대입하여 획득되는 커널값을 가중한 가중 거리 기반 스코어를 소프트아그맥스 함수에 대입하여 정규화값이 2개의 양자값 각각 분포할 확률을 나타내는 거리 기반 확률을 계산하고, 계산된 거리 기반 확률을 2개의 양자값에 가중합하여 양자화값을 획득하는 확률 분포 기반 양자화부를 포함하여, 미분 가능한 양자화를 수행할 수 있어 양자화에 따른 인공 신경망의 학습 오차를 최소화할 수 있고, 고속으로 학습을 수행할 수 있는 인공 신경망을 위한 양자화기 및 양자화 방법을 제공할 수 있다.

대 표 도



(52) CPC특허분류
G06N 7/01 (2023.01)

명 세 서

청구범위

청구항 1

인공 신경망에 구비되는 양자화기에 있어서,

양자화 대상인 대상값을 인가받아 기지정된 범위로 정규화하여 정규화값을 출력하는 정규화부;

기지정된 다수의 양자값 중 상기 정규화값에 가장 인접한 2개의 양자값을 선택하고, 선택된 2개의 양자값 각각과 상기 정규화값 사이의 차에 따라 기지정된 방식으로 2개의 거리 기반 스코어를 계산하는 거리 스코어 계산부;

상기 2개의 거리 기반 스코어에 상기 2개의 양자값 각각을 커널 함수(Kernel function)에 대입하여 획득되는 커널값을 가중한 가중 거리 기반 스코어를 소프트 아그맥스 함수(soft argmax function)에 대입하여 상기 정규화값이 상기 2개의 양자값 각각 분포할 확률을 나타내는 거리 기반 확률을 계산하고, 계산된 거리 기반 확률을 상기 2개의 양자값에 가중합하여 양자화값을 획득하는 확률 분포 기반 양자화부를 포함하는 인공 신경망을 위한 양자화기.

청구항 2

제1항에 있어서, 상기 확률 분포 기반 양자화부는

거리 기반 스코어에 가중되는 커널값의 가중 비중을 조절하기 위한 온도 파라미터를 상기 2개의 양자값 각각에 대한 거리 기반 스코어에 대응하는 커널값이 가중된 가중 거리 기반 스코어 사이의 차에 따라 계산하는 온도 파라미터 계산부;

상기 온도 파라미터에 따른 비중으로 커널값이 가중된 가중 거리 기반 스코어를 소프트 아그맥스 함수에 대입하여 상기 2개의 양자값 각각에 대한 거리 기반 확률을 계산하는 확률 분포 계산부; 및

상기 온도 파라미터에 의해 발생되는 양자화 오차가 제거되도록 상기 2개의 양자값을 상기 온도 파라미터에 대응하여 변형하고, 변형된 2개의 양자값에 상기 거리 기반 확률을 가중합하여 상기 양자화값을 획득하는 양자화값 획득부를 포함하는 인공 신경망을 위한 양자화기.

청구항 3

제2항에 있어서, 상기 거리 스코어 계산부는

상기 2개의 거리 기반 스코어($n(x_n,q)$)를 수학식

$$n(x_n, q) = e^{-|x_n - q|}$$

(여기서 x_n은 정규화값이고, q는 양자값을 나타낸다.)

에 따라 계산하는 인공 신경망을 위한 양자화기.

청구항 4

제3항에 있어서, 상기 온도 파라미터 계산부는

상기 온도 파라미터를 수학식

$$\beta = \frac{\gamma}{|s_{x_n}(q_f) - s_{x_n}(q_c)|}$$

(여기서 γ는 기지정된 상수값이며, q_f, q_c는 2개의 양자값이며, s_{x_n}은 선택된 2개의 양자값(q_f, q_c)에 대한 거

리 기반 스코어와 커널값을 가중한 가중 거리 기반 스코어를 나타낸다.)

에 따라 계산하는 인공 신경망을 위한 양자화기.

청구항 5

제4항에 있어서, 상기 확률 분포 계산부는

상기 2개의 양자값 각각에 대한 거리 기반 확률(m_{x_n})을 수학식

$$m_{x_n}(q_i) = \frac{\exp(\beta k_{x_n}(q_i) n_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta k_{x_n}(q_j) n_{x_n}(q_j))} = \frac{\exp(\beta s_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta s_{x_n}(q_j))}$$

(여기서 q_i 는 2개의 양자값($q_i \in (q_f, q_c)$)을 나타낸다)

에 따라 계산하는 인공 신경망을 위한 양자화기.

청구항 6

제5항에 있어서, 상기 양자화값 획득부는

2개의 양자값(q_f, q_c)을 수학식

$$q_f^* = q_f - \frac{1}{e^{\gamma}-1}$$
$$q_c^* = q_c + \frac{1}{e^{\gamma}-1}$$

따라 변형된 양자값(q_f^*, q_c^*)으로 변형하는 인공 신경망을 위한 양자화기.

청구항 7

제6항에 있어서, 상기 양자화값 획득부는

상기 양자화값을 수학식

$$\phi(x_n) = \sum_{i\in\{f,c\}} m_{x_n}(q_i) q_i^*$$

에 따라 획득하는 인공 신경망을 위한 양자화기.

청구항 8

제1항에 있어서, 상기 정규화부는

기지정된 상한값(u)과 하한값(l)에 따라 상기 대상값을 클리핑하여 클리핑값(x_c)을 획득하고, 획득된 클리핑값(x_c)을 수학식

$$x_n = (2^k - 1)\frac{x_c - l}{u - l}$$

에 따라 정규화하여 기지정된 양자화 비트수(k)에 대응하는 양자화 범위로 정규화하는 인공 신경망을 위한 양자화기.

청구항 9

인공 신경망을 위한 양자화기에서 수행되는 양자화 방법으로서,

양자화 대상인 대상값을 인가받아 기지정된 범위로 정규화하여 정규화값을 출력하는 단계;

기지정된 다수의 양자값 중 상기 정규화값에 가장 인접한 2개의 양자값을 선택하고, 선택된 2개의 양자값 각각과 상기 정규화값 사이의 차에 따라 기지정된 방식으로 2개의 거리 기반 스코어를 계산하는 단계;

상기 2개의 거리 기반 스코어에 상기 2개의 양자값 각각을 커널 함수에 대입하여 획득되는 커널값을 가중한 가중 거리 기반 스코어를 소프트 아그맥스 함수에 대입하여 상기 정규화값이 상기 2개의 양자값 각각 분포할 확률을 나타내는 거리 기반 확률을 계산하고, 계산된 거리 기반 확률을 상기 2개의 양자값에 가중합하여 양자화값을 획득하는 단계를 포함하는 인공 신경망을 위한 양자화 방법.

청구항 10

제9항에 있어서, 상기 양자화값을 획득하는 단계는

상기 2개의 양자값 각각에 대한 거리 기반 스코어에 대응하는 커널값이 가중된 가중 거리 기반 스코어 사이의 차에 따라 거리 기반 스코어에 가중되는 커널값의 가중 비중을 조절하기 위한 온도 파라미터를 계산하는 단계;

상기 온도 파라미터에 따른 비중으로 커널값이 가중된 가중 거리 기반 스코어를 소프트 아그맥스 함수에 대입하여 상기 2개의 양자값 각각에 대한 거리 기반 확률을 계산하는 단계;

상기 온도 파라미터에 의해 발생되는 양자화 오차가 제거되도록 상기 온도 파라미터에 대응하여 상기 2개의 양자값을 변형하는 단계; 및

변형된 2개의 양자값에 상기 거리 기반 확률을 가중합하여 상기 양자화값을 계산하는 단계를 포함하는 인공 신경망을 위한 양자화 방법.

청구항 11

제10항에 있어서, 상기 거리 기반 스코어를 계산하는 단계는

상기 2개의 거리 기반 스코어($n(x_n,q)$)를 수학식

$$n(x_n, q) = e^{-|x_n - q|}$$

(여기서 x_n은 정규화값이고, q는 양자값을 나타낸다.)

에 따라 계산하는 인공 신경망을 위한 양자화 방법.

청구항 12

제11항에 있어서, 상기 온도 파라미터를 계산하는 단계는

상기 온도 파라미터를 수학식

$$\beta = \frac{\gamma}{|s_{x_n}(q_f) - s_{x_n}(q_c)|}$$

(여기서 γ는 기지정된 상수값이며, q_f, q_c는 2개의 양자값이며, s_{x_n}은 선택된 2개의 양자값(q_f, q_c)에 대한 거리 기반 스코어와 커널값을 가중한 가중 거리 기반 스코어를 나타낸다.)

에 따라 계산하는 인공 신경망을 위한 양자화 방법.

청구항 13

제12항에 있어서, 상기 거리 기반 확률을 계산하는 단계는

상기 2개의 양자값 각각에 대한 거리 기반 확률(m_{x_n})을 수학식

$$m_{x_n}(q_i) = \frac{\exp(\beta k_{x_n}(q_i) n_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta k_{x_n}(q_j) n_{x_n}(q_j))} = \frac{\exp(\beta s_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta s_{x_n}(q_j))}$$

(여기서 q_i 는 2개의 양자값($q_i \in (q_f, q_c)$)을 나타낸다)

에 따라 계산하는 인공 신경망을 위한 양자화 방법.

청구항 14

제13항에 있어서, 상기 양자값을 변형하는 단계는

2개의 양자값(q_f, q_c)을 수학식

$$q_f^* = q_f - \frac{1}{e^{\gamma}-1}$$
$$q_c^* = q_c + \frac{1}{e^{\gamma}-1}$$

따라 변형된 양자값(q_f^*, q_c^*)으로 변형하는 인공 신경망을 위한 양자화 방법.

청구항 15

제14항에 있어서, 상기 양자화값을 계산하는 단계는

상기 양자화값을 수학식

$$\phi(x_n) = \sum_{i\in\{f,c\}} m_{x_n}(q_i) q_i^*$$

에 따라 획득하는 인공 신경망을 위한 양자화 방법.

청구항 16

제9항에 있어서, 상기 정규화값을 출력하는 단계는

기지정된 상한값(u)과 하한값(l)에 따라 상기 대상값을 클리핑하여 클리핑값(x_c)을 획득하는 단계; 및

획득된 클리핑값(x_c)을 수학식

$$x_n = (2^k - 1)\frac{x_c - l}{u - l}$$

에 따라 정규화하여 기지정된 양자화 비트수(k)에 대응하는 양자화 범위로 정규화하는 단계를 포함하는 인공 신경망을 위한 양자화 방법.

청구항 17

제9항 내지 제16항 중 어느 한 항에 따른 양자화 방법을 실행하기 위한 프로그램 명령어가 기록된, 컴퓨터가 판독 가능한 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 양자화기 및 양자화 방법에 관한 것으로, 인공 신경망을 위한 양자화기 및 양자화 방법에 관한 것이

다.

배 경 기 술

[0002] 최근 인간의 두뇌가 패턴을 인식하는 방법을 모사하여 두뇌와 비슷한 방식으로 여러 정보를 처리하도록 구성된 인공 신경망(artificial neural network)에 대한 연구가 활발하게 진행되고 있다. 현재 인공 신경망은 객체 분류, 객체 검출, 음성 인식, 자연어 처리와 같은 매우 다양한 분야에서 적용되고 있을 뿐만 아니라 그 적용 분야가 계속 확장되어 가고 있다.

[0003] 이러한 인공 신경망은 입력맵(또는 특징맵)과 미리 학습에 의해 획득된 가중치맵 사이에 기지정된 연산을 수행하여 요구되는 동작을 수행한다. 이때 입력맵(또는 특징맵)과 가중치맵은 대부분 행렬(또는 벡터) 형식을 가지므로, 인공 신경망은 대량의 행렬 연산(예를 들면 컨볼루션 연산)을 요구하게 된다. 그리고 인공 신경망은 실제로 이용되기 이전에 대량의 반복 학습이 수행되어 가중치맵의 가중치가 업데이트되어야만 요구되는 성능을 나타낼 수 있다. 즉 인공 신경망은 학습 및 실제 운용 과정에서 대량의 행렬 연산이 수행되어야 한다. 그러므로 연산 효율성이 인공 신경망의 성능을 크게 좌우하게 된다. 그러나 일반적으로 가중치맵과 입력맵(또는 특징맵)의 각 원소인 가중치와 입력값은 부동 소수점(floating point) 형식의 값을 가지므로 연산 효율성이 낮다.

[0004] 이에 최근에는 인공 신경망에서 연산되어야 하는 가중치 또는 가중치와 입력값을 양자화하여 연산을 수행하도록 함으로써 연산 효율성을 향상시키고자 하는 연구가 활발하게 수행되고 있다.

[0005] 도 1은 기존의 인공 신경망을 위한 양자화기의 개략적 구조를 나타낸다.

[0006] 도 1을 참조하면, 양자화기(100)는 정규화부(110), 양자화부(120)를 포함할 수 있다.

[0007] 정규화부(110)는 양자화되어야 하는 대상인 대상값(x)을 인가받아 기지정된 범위로 정규화하여 정규화값(x_n)을 출력한다.

[0008] 양자화부(120)는 정규화부(110)에서 인가되는 정규화값(x_n)을 기지정된 간격 단위로 이산 분포되어 있는 양자값($q = [0, 1, \cdots 2^k-1]$)과 비교하여 가장 인접한 양자값으로 변환하여 대상값(x)에 대응하는 양자화값($Q(x_n)$)을 획득하여 출력한다. 일반적으로 양자화부(120)는 인가된 정규화값(x_n)에 대해 부호 함수(signum function)나 반올림 함수(round function) 등을 이용하여 양자화값($Q(x_n)$)을 획득할 수 있다.

[0009] 이와 같이, 양자화기(100)는 부동 소수점 형식의 대상값(x)을 양자화하여 양자화값($Q(x_n)$)을 출력하여, 인공 신경망이 부동 소수점 형식이 아니라 기지정된 양자값(q) 중 하나로 변환된 양자화값($Q(x_n)$)으로 연산을 수행할 수 있도록 함으로써, 인공 신경망의 연산 효율성을 향상시킬 수 있다.

[0010] 다만 인공 신경망의 경우, 학습 시에 기지정된 방식으로 계산되는 손실(L)을 역전파(Backward propagation)하여 가중치를 반복 업데이트하는 방식으로 수행되어야 한다. 이러한 역전파 방식의 학습 과정을 위해서는 양자화기가 손실 역전파(Loss Backpropagation)를 수행하여 가중치를 업데이트 수 있어야 하며, 손실 역전파를 위해서는 양자화기가 미분 가능한 함수의 형태로 양자화 연산을 수행해야 한다.

[0011] 그러나 양자화부(120)가 정규화값(x_n)을 양자화값($Q(x_n)$)으로 변환하는 연산은 부호 함수나 반올림 함수와 같은 이산 함수(discrete function)에 의한 연산으로 미분이 불가능하므로, 인공 신경망을 학습시킬 수 없다는 문제가 있다.

[0012] 도 2는 기존의 이산화부의 연산 함수와 이의 미분 함수 그래프의 일 예를 나타낸다.

[0013] 도 2의 (a)는 부호 함수를 이용하여 정규화값(x_n)을 양자화값($Q(x_n)$)으로 변환하는 그래프를 나타내고, (b)는 (a)의 부호 함수에 대한 미분 함수의 그래프를 나타낸다.

[0014] 도 2의 (a)와 (b)에 도시된 바와 같이, 양자화부(120)가 이산 함수에 기반하여 정규화값(x_n)을 부호 함수와 같은 양자화값($Q(x_n)$)으로 변환하는 경우, 그에 대한 미분 함수는 (b)에 도시된 바와 같이, 정규화값(x_n)이 0인 경우($x_n = 0$)를 제외($x_n \neq 0$)하면 모두 0으로 출력되므로, 그래디언트 소실(gradient vanishing) 문제가 발생한다. 이로 인해 손실 역전파가 이루어 지지 않아 학습이 불가능한 문제가 있다.

[0015] 이러한 문제를 해결하기 위해 기존에는 (d)에 도시된 바와 같이, 손실 역전파 시, (b)의 미분 함수를 STE(Straight through estimator)로 대체하여 손실 역전파가 가능하도록 함으로써 인공 신경망이 학습될 수 있도록 하는 기법이 제안된 바 있다.

[0016] 그러나 STE는 (c)에 도시된 바와 같이, 하드 하이퍼볼릭 탄젠트(hard hyperbolic tangent: htanth)에 대한 미분 함수로서, 그래디언트를 기지정된 상수(여기서는 일 예로 1)로 균일하게 설정하여 역전파 가능하도록 대체한다. 그러나 STE로 미분 함수를 대체하는 것은 양자화를 위한 순방향 전파 시와 역전파 시에 서로 다른 함수를 사용하는 것을 의미한다. 이로 인해 그래디언트 불일치 문제가 발생하며, 그래디언트 불일치 문제는 손실 역전파 시에 추가적인 오차를 유발하여 인공 신경망의 학습 시에 특정값으로 수렴되어야 하는 가중치 데이터가 수렴되기 어렵도록 하여 학습 불안정성을 초래하는 문제가 있다. 이러한 학습 불안정성은 인공 신경망의 학습 정확도뿐만 아니라, 학습 속도를 저해하는 요인이 된다.

[0017] 양자화기를 이산 함수를 대체하여 유사한 특성을 갖는 시그모이드(sigmoid) 함수 또는 하이퍼볼릭 탄젠트(hyperbolic tangent: tanth) 함수 등을 이용하여 양자화기의 동작을 대체하는 방법도 제안된 바 있으나, 이산 함수를 이용하는 이상적인 양자화기의 동작과는 차이가 있어 여전히 오차가 발생한다는 한계가 있다.

선행기술문헌

특허문헌

[0018] (특허문헌 0001) 한국 공개 특허 제10-2019-0134965호 (2019.12.05 공개)

발명의 내용

해결하려는 과제

[0019] 본 발명의 목적은 이산 함수 기반의 양자화 기법과 동일한 수준으로 양자화를 수행하면서도 미분 가능하여 손실 역전파가 가능한 양자화기 및 양자화 방법을 제공하는데 있다.

[0020] 본 발명의 다른 목적은 동일한 함수를 기반으로 양자화에 대한 순방향 및 역방향 전파를 수행할 수 있어 양자화에 따른 인공 신경망의 학습 오차를 최소화할 수 있고, 고속으로 학습을 수행할 수 있도록 하는 인공 신경망을 위한 양자화기 및 양자화 방법을 제공하는데 있다.

과제의 해결 수단

[0021] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 인공 신경망을 위한 양자화기는 양자화 대상인 대상값을 인가받아 기지정된 범위로 정규화하여 정규화값을 출력하는 정규화부; 기지정된 다수의 양자값 중 상기 정규화값에 가장 인접한 2개의 양자값을 선택하고, 선택된 2개의 양자값 각각과 상기 정규화값 사이의 차에 따라 기지정된 방식으로 2개의 거리 기반 스코어를 계산하는 거리 스코어 계산부; 상기 2개의 거리 기반 스코어에 상기 2개의 양자값 각각을 커널 함수(Kernel function)에 대입하여 획득되는 커널값을 가중한 가중 거리 기반 스코어를 소프트 아그맥스 함수(soft argmax function)에 대입하여 상기 정규화값이 상기 2개의 양자값 각각 분포할 확률을 나타내는 거리 기반 확률을 계산하고, 계산된 거리 기반 확률을 상기 2개의 양자값에 가중합하여 양자화값을 획득하는 확률 분포 기반 양자화부를 포함한다.

[0022] 상기 확률 분포 기반 양자화부는 거리 기반 스코어에 가중되는 커널값의 가중 비중을 조절하기 위한 온도 파라미터를 상기 2개의 양자값 각각에 대한 거리 기반 스코어에 대응하는 커널값이 가중된 가중 거리 기반 스코어 사이의 차에 따라 계산하는 온도 파라미터 계산부; 상기 온도 파라미터에 따른 비중으로 커널값이 가중된 가중 거리 기반 스코어를 소프트 아그맥스 함수에 대입하여 상기 2개의 양자값 각각에 대한 거리 기반 확률을 계산하는 확률 분포 계산부; 및 상기 온도 파라미터에 의해 발생되는 양자화 오차가 제거되도록 상기 2개의 양자값을 상기 온도 파라미터에 대응하여 변형하고, 변형된 2개의 양자값에 상기 거리 기반 확률을 가중합하여 상기 양자화값을 획득하는 양자화값 획득부를 포함할 수 있다.

[0023] 상기 거리 스코어 계산부는 상기 2개의 거리 기반 스코어($n(x_n,q)$)를 수학식

[0024] $$n(x_n, q) = e^{-|x_n - q|}$$

[0025] (x_n은 정규화값이고, q는 양자값을 나타낸다.)에 따라 계산할 수 있다.

[0026] 상기 온도 파라미터 계산부는 상기 온도 파라미터를 수학식

[0027] $$\beta = \frac{\gamma}{|s_{x_n}(q_f) - s_{x_n}(q_c)|}$$

[0028] (γ는 기지정된 상수값이며, q_f, q_c는 2개의 양자값이며, s_{x_n}은 선택된 2개의 양자값(q_f, q_c)에 대한 거리 기반 스코어와 커널값을 가중한 가중 거리 기반 스코어를 나타낸다.)에 따라 계산할 수 있다.

[0029] 상기 확률 분포 계산부는 상기 2개의 양자값 각각에 대한 거리 기반 확률(m_{x_n})을 수학식

[0030] $$m_{x_n}(q_i) = \frac{\exp(\beta k_{x_n}(q_i) n_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta k_{x_n}(q_j) n_{x_n}(q_j))} = \frac{\exp(\beta s_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta s_{x_n}(q_j))}$$

[0031] (q_i 는 2개의 양자값($q_i \in (q_f, q_c)$)을 나타낸다)에 따라 계산할 수 있다.

[0032] 상기 양자화값 획득부는 2개의 양자값(q_f, q_c)을 수학식

[0033] $$q_f^* = q_f - \frac{1}{e^{\gamma}-1}$$
$$q_c^* = q_c + \frac{1}{e^{\gamma}-1}$$

[0034] 따라 변형된 양자값(q_f^*, q_c^*)으로 변형할 수 있다.

[0035] 상기 양자화값 획득부는 상기 양자화값을 수학식

[0036] $$\phi(x_n) = \sum_{i\in\{f,c\}} m_{x_n}(q_i) q_i^*$$

[0037] 에 따라 획득할 수 있다.

[0038] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 인공 신경망을 위한 양자화 방법은 양자화 대상인 대상값을 인가받아 기지정된 범위로 정규화하여 정규화값을 출력하는 단계; 기지정된 다수의 양자값 중 상기 정규화값에 가장 인접한 2개의 양자값을 선택하고, 선택된 2개의 양자값 각각과 상기 정규화값 사이의 차에 따라 기지정된 방식으로 2개의 거리 기반 스코어를 계산하는 단계; 상기 2개의 거리 기반 스코어에 상기 2개의 양자값 각각을 커널 함수에 대입하여 획득되는 커널값을 가중한 가중 거리 기반 스코어를 소프트 아그맥스 함수에 대입하여 상기 정규화값이 상기 2개의 양자값 각각 분포할 확률을 나타내는 거리 기반 확률을 계산하고, 계산된 거리 기반 확률을 상기 2개의 양자값에 가중합하여 양자화값을 획득하는 단계를 포함한다.

발명의 효과

[0039] 따라서, 본 발명의 실시예에 따른 인공 신경망을 위한 양자화기 및 양자화 방법은 양자화 대상인 대상값과 기지정된 양자값 사이의 거리에 따라 거리 기반 스코어를 부여하고, 부여된 거리 기반 스코어의 확률 분포에 기반하는 커널 소프트 아그맥스 함수를 양자화 함수로 이용하여 양자화값을 획득함으로써, 이산 함수 기반의 양자화 기법과 동일한 수준으로 양자화를 수행하면서도 미분 가능한 양자화를 수행할 수 있다. 그러므로 동일한 함수를 기반으로 순방향 및 역방향 전파를 수행할 수 있어 양자화에 따른 인공 신경망의 학습 오차를 최소화할 수 있고, 고속으로 학습을 수행할 수 있다.

도면의 간단한 설명

[0040] 도 1은 기존의 인공 신경망을 위한 양자화기의 개략적 구조를 나타낸다.

도 2는 기존의 이산화부에서 이용하는 이산 함수와 이의 미분 함수 그래프의 일 예를 나타낸다.

도 3은 본 발명의 일 실시예에 따른 인공 신경망을 위한 양자화기의 개략적 구조를 나타낸다.

도 4는 양자값에 대한 대상값의 거리 기반 스코어의 일 예를 나타낸다.

도 5는 온도 파라미터의 변화에 따른 양자화값의 변화를 나타낸다.

도 6은 온도 파라미터의 변화에 따른 양자화값의 변화와 이의 미분 그래프의 변화를 나타낸다.

도 7은 정규화값에 따른 선택된 2개의 양자값에 대한 커널 함수로 가중된 거리 기반 스코어의 차이 변화를 나타낸다.

도 8은 본 발명의 일 실시예에 따른 인공 신경망을 위한 양자화 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0041] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0042] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0043] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0044] 도 3은 본 발명의 일 실시예에 따른 인공 신경망을 위한 양자화기의 개략적 구조를 나타낸다.

[0045] 도 3을 참조하면, 본 실시예에 따른 양자화기(200)는 정규화부(210)와 거리 스코어 계산부(220) 및 확률 분포 기반 양자화부(230)를 포함한다.

[0046] 정규화부(210)는 도 1의 정규화부(110)와 마찬가지로 양자화되어야 하는 대상인 대상값(x)을 인가받아 기지정된 범위로 정규화하여 정규화값(x_n)을 출력한다. 여기서 대상값(x)은 인공 신경망에 인가되는 입력맵(또는 특징맵)의 원소이거나 학습에 의해 업데이트되는 가중치맵의 원소일 수 있으며, 부동 소수점 형식의 데이터일 수 있다. 그리고 입력맵(또는 특징맵)과 가중치맵은 행렬 또는 벡터 형식의 데이터일 수 있다.

[0047] 정규화부(210)는 대상값(x)인 인가되면, 수학식 1과 같이, 클리핑 함수(clipping function)를 이용하여 인가된 대상값(x)이 기지정된 상한값(u)과 하한값(l)을 초과하지 않도록 클리핑하여 클리핑값(x_c)을 획득한다.

수학식 1

[0048] $$x_c = \mathrm{clip}(x, \min = l, \max = u)$$

[0049] 그리고 획득된 클립핑값(x_c)을 양자화시키고자 하는 양자화 범위[0, 2^k-1]에 따라 수학식 2와 같이 정규화하여 정규화값(x_n)을 출력한다.

수학식 2

[0050] $$x_n = (2^k - 1)\frac{x_c - l}{u - l}$$

[0051] 여기서 k는 기지정된 양자화 비트수이다.

[0052] 여기서 양자화기의 양자화 비트수가 k이므로, 양자화 범위는 [0, 2^k-1]가 되고, 양자화 범위 내에서 기지정된 간격 단위로 이산 분포되는 양자값(q)은 [0, 1, … 2^k-1])로 설정될 수 있다. 이는 양자화부(230)에서 정규화값(x_n)을 양자화하여 획득되는 양자화값(Q(x_n))이 양자값(q) 중 하나로 획득될 수 있음을 의미한다.

[0053] 한편, 본 실시예에서 양자화기(200)는 기존의 양자화기(100)와 달리 이산 함수를 이용하여 양자화를 수행하지 않고, 기지정된 다수의 양자값(q)과 정규화값(x_n) 사이의 차에 따른 거리 기반 스코어를 계산하고, 계산된 거리 기반 스코어를 바탕으로 정규화값(x_n)의 확률 분포를 계산하여 정규화값(x_n)을 양자화한다.

[0054] 이에 거리 스코어 계산부(220)는 정규화부(210)로부터 정규화값(x_n)을 인가받고, 다수의 양자값(q_i = [0, 1, … 2^k-1]) 중 인가된 정규화값(x_n)에 가장 인접한 2개의 양자값(q_f, q_c)를 선택하고, 정규화값(x_n)과 선택된 2개의 양자값(q_f, q_c) 사이의 거리 기반 스코어(n(x_n,q))를 계산한다.

[0055] 거리 스코어 계산부(220)는 다수의 양자값(q = [0, 1, … 2^k-1]) 중 인가된 정규화값(x_n)에 가장 인접한 2개의 양자값(q_f, q_c)을 각각 내림 함수(floor function)과 올림 함수(ceil function)를 이용하여 획득할 수 있다. 여기서 내림 함수를 이용하여 선택되는 양자값을 제1 양자값(q_f = floor(x_n))이라 하고, 올림 함수를 이용하여 선택되는 양자값을 제2 양자값(q_c = ceil(x_n))이라 할 수 있다. 따라서 제1 양자값(q_f)은 정규화값(x_n) 이하의 값을 갖는 최대 양자값이며, 제2 양자값(q_c)은 정규화값(x_n) 이상의 값을 갖는 최소 양자값이다.

[0056] 제1 및 제2 양자값(q_f, q_c)이 선택되면, 거리 스코어 계산부(220)는 선책된 2개의 양자값(q_f, q_c) 각각과 정규화값(x_n) 사이의 차에 기반하여, 정규화값(x_n)에 대응하는 거리 기반 스코어(n(x_n,q))를 수학식 3에 따라 계산한다.

수학식 3

[0057] $$n(x_n, q) = e^{-|x_n - q|}$$

[0058] 거리 기반 스코어(n(x_n,q))는 수학식 3에 따라 다수의 양자값(q = [0, 1, … 2^k-1]) 각에 대해 계산될 수 있으나, 여기서는 거리 스코어 계산부(220)가 선택된 제1 및 제2 양자값(q_f, q_c)에 대해서만 계산하여 불필요한 연산이 수행되는 것을 방지한다.

[0059] 도 4는 양자값에 대한 대상값의 거리 기반 스코어의 일 예를 나타낸다.

[0060] 도 4에서는 일 예로 선택된 양자값(q_f, q_c) 중 하나가 1인 경우(q = 1)를 가정하여, 정규화값(x_n)에 따라 획득되는 거리 기반 스코어(n(x_n,1))를 도시하였다. 수학식 3에 나타난 바와 같이, 거리 스코어 계산부(220)는 선택된 양자값(q_f, q_c) 각각을 중심으로, 선택된 양자값(q_f, q_c)과 정규화값(x_n) 사이의 차의 절대값에 따른 지수 함수 그래프 형태로 거리 기반 스코어(n(x_n,q))가 계산될 수 있다.

[0061] 여기서 거리 기반 스코어(n(x_n,q))는 선택된 2개의 양자값(q_f, q_c) 각각에 대해 획득되므로, 거리 스코어 계산부

(220)는 정규화값(x_n) 각각에 대해 2개의 거리 기반 스코어($n(x_n,q_f)$, $n(x_n,q_c)$)를 획득할 수 있다.

[0062] 확률 분포 기반 양자화부(230)는 거리 스코어 계산부(220)에서 정규화값(x_n) 각각에 대해 획득된 2개의 거리 기반 스코어($n(x_n,q_f)$, $n(x_n,q_c)$)를 이용하여 정규화값(x_n)의 거리 기반 확률 분포에 따라 계산하여 양자화를 수행한다.

[0063] 본 실시예에서 확률 분포 기반 양자화부(230)는 획득된 2개의 거리 기반 스코어($n(x_n,q_f)$, $n(x_n,q_c)$) 각각에 커널 함수(Kernel function)(k_{x_n})가 가중된 가중 거리 기반 스코어(s_{x_n})를 소프트 아그맥스 함수(soft argmax function)에 대입하여 수학식 4에 따라 정규화값(x_n)의 거리 기반 확률(m_{x_n})을 계산한다.

수학식 4

[0064]
$$m_{x_n}(q_i) = \frac{\exp(\beta k_{x_n}(q_i) n_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta k_{x_n}(q_j) n_{x_n}(q_j))} = \frac{\exp(\beta s_{x_n}(q_i))}{\sum_{j\in\{f,c\}} \exp(\beta s_{x_n}(q_j))}$$

[0065] 여기서 q_i ($q_i \in (q_f, q_c)$)는 양자값, β는 온도 파라미터(temperature)이고, n_{x_n}은 거리 기반 스코어($n_{x_n} = n(x_n,q)$)이다.

[0066] 커널 함수와 소프트 아그맥스 함수가 결합된 수학식 4를 여기서는 커널 소프트 아그맥스 함수(Kernel soft argmax function)라 한다. 즉 수학식 4는 정규화값(x_n)의 확률 분포가 선택된 2개의 양자값(q_f, q_c)에 밀집되도록, 양자값(q_i)을 커널 함수(k_{x_n})에 대입한 커널값($k_{x_n}(q_i)$)을 2개의 거리 기반 스코어($n(x_n,q_i)$)에 가중하여 정규화값(x_n)의 거리 기반 확률(m_{x_n})을 획득할 수 있도록 한다. 여기서 온도 파라미터(β)는 거리 기반 스코어에 가중되는 커널값($k_{x_n}(q_i)$)의 가중 비중을 조절하기 위한 조절 파라미터이다.

[0067] 선택된 2개의 양자값(q_f, q_c) 각각에 대한 거리 기반 확률(m_{x_n})이 계산되면, 선택된 2개의 양자값(q_f, q_c)에 계산된 2개의 거리 기반 확률(m_{x_n})을 가중 평균하여 수학식 5에 따라 정규화값(x_n)에 대한 양자화값(ø(x_n))을 획득할 수 있다.

수학식 5

[0068]
$$\phi(x_n) = \sum_{i\in\{f,c\}} m_{x_n}(q_i) q_i$$

[0069] 즉 선택된 2개의 양자값(q_f, q_c)에 대해 2개의 거리 기반 확률(m_{x_n})를 가중 평균하여 양자화값(ø(x_n))을 획득할 수 있다.

[0070] 도 5는 온도 파라미터의 변화에 따른 양자화값의 변화를 나타내고, 도 6은 온도 파라미터의 변화에 따른 양자화값의 변화와 이의 미분 그래프의 변화를 나타낸다.

[0071] 온도 파라미터(β)는 기지정된 상수값으로 설정될 수 있으며, 도 5에서는 온도 파라미터(β)가 각각 10, 20, 30 및 40인 경우에 정규화값(x_n)에 따라 획득되는 양자화값(ø(x_n))의 변화를 도시하였다.

[0072] 도 5에 도시된 바와 같이, 온도 파라미터(β)가 10인 경우, 정규화값(x_n)에 따라 획득되는 양자화값(ø(x_n))은 시그모이드 함수를 이용하여 양자화한 경우와 유사하게 나타난다. 즉 정규화값(x_n)이 선택된 2개의 양자값(q_f,

q_c)의 중앙((q_f + q_c)/2)에 가까울수록 양자화값($\phi(x_n)$)이 이산 함수를 이용하는 경우와 상이하게 나타나게 되어 양자화 오차가 증가하게 된다. 그러나 온도 파라미터(β)가 20, 30 및 40으로 증가될수록 점차적으로 이상적인 양자화기와 유사한 형태로 양자화값($\phi(x_n)$)의 분포가 변화됨을 알 수 있다.

[0073] 즉 온도 파라미터(β)의 값이 증가될수록 수학식 4 및 수학식 5에 따라 계산되는 양자화값($\phi(x_n)$)의 양자화 오차가 감소하게 된다.

[0074] 한편 도 6에서 상단의 그래프는 도 5에서와 마찬가지로 온도 파라미터(β)의 변화에 따른 양자화값($\phi(x_n)$)의 변화를 나타내고, 하단 그래프는 온도 파라미터(β)의 변화에 따른 양자화값($\phi(x_n)$)의 미분 그래프를 나타낸다.

[0075] 도 6에서는 온도 파라미터(β)가 각각 10, 20, 40 및 80으로 증가되는 경우를 도시하였다. 도 6의 상단에 도시된 바와 같이, 온도 파라미터(β)가 10, 20, 40 및 80으로 증가됨에 따라 양자화값($\phi(x_n)$)이 이산 함수를 이용하는 이상적인 양자화기와 유사한 그래프를 나타냄을 알 수 있다. 그러나 하단에 도시된 바와 같이, 온도 파라미터(β)가 증가함에 따라 양자화값($\phi(x_n)$)의 미분 그래프 또한 이산 함수를 이용하는 경우와 유사하게 점차로 0에 수렴하는 형태의 그래프로 나타나게 된다.

[0076] 이는 이산 함수를 이용하여 양자화기를 구성하는 경우와 마찬가지로 그래디언트 소실(gradient vanishing) 문제를 야기하게 된다.

[0077] 이에 본 실시예에서 확률 분포 기반 양자화부(230)는 온도 파라미터(β)를 상수로 설정하지 않고, 거리 기반 스코어(n_{x_n})와 커널값($k_{x_n}(q_i)$)에 따라 가변되는 변수로 설정하여, 정규화값(x_n)이 선택된 2개의 양자값(q_f, q_c)의 중앙((q_f + q_c)/2)에 가까울수록 큰 값을 갖고, 2개의 양자값(q_f, q_c)에 가까울수록 작은 값을 갖도록 함으로써, 이산 함수를 적용하는 이상적인 양자화기와 유사하게 양자화를 수행하면서도 미분 가능하여 역전파가 가능하도록 한다.

[0078] 다시 도 3을 참조하면, 확률 분포 기반 양자화부(230)는 온도 파라미터 계산부(231), 확률 분포 계산부(2232) 및 양자화값 획득부(233)를 포함할 수 있다.

[0079] 온도 파라미터 계산부(231)는 수학식 6에 따라 거리 기반 스코어(n_{x_n})와 커널값($k_{x_n}(q_i)$)에 따라 가변되는 온도 파라미터(β)를 획득한다.

수학식 6

[0080]
$$\beta = \frac{\gamma}{|s_{x_n}(q_f) - s_{x_n}(q_c)|}$$

[0081] 여기서 γ는 기지정된 상수값이며, s_{x_n}은 선택된 2개의 양자값(q_f, q_c)에 대한 거리 기반 스코어(n_{x_n})와 커널값($k_{x_n}(q_i)$)의 곱($s_{x_n}(q_i) = k_{x_n}(q_i)n_{x_n}(q_i)$)으로 가중 거리 기반 스코어를 나타낸다.

[0082] 수학식 6에서 분자항의 상수값(γ)은 기존의 상수 온도 파라미터에 대응하는 값으로 상수값(γ)이 증가됨에 따라(예를 들면 γ = 80) 양자화값($\phi(x_n)$)은 이상적인 양자화기의 양자화값($Q(x_n)$)과 동일한 값으로 출력되게 된다.

[0083] 그러나 수학식 6의 분모항($|s_{x_n}(q_f) - s_{x_n}(q_c)|$)은 정규화값($x_n$)이 선택된 2개의 양자값($q_f$, q_c)에 가까울수록 증가된다.

[0084] 도 7은 정규화값에 따른 선택된 2개의 양자값에 대한 커널 함수로 가중된 거리 기반 스코어의 차이 변화를 나타낸다.

[0085] 도 7에 나타난 바와 같이, 수학식 6에서 분모항($|s_{x_n}(q_f) - s_{x_n}(q_c)|$)으로 나타나는 선택된 2개의 양자값에

대한 커널 함수로 가중된 거리 기반 스코어의 차는 정규화값(x_n)이 선택된 2개의 양자값(q_f, q_c)의 중앙((q_f + q_c)/2)에 가까울수록 감소하게 되고, 2개의 양자값(q_f, q_c) 중 하나에 가까워질수록 점차로 증가하게 된다.

[0086] 따라서 수학식 6에 따라 계산되는 온도 파라미터(β)는 정규화값(x_n)이 2개의 양자값(q_f, q_c) 중 하나에 가까울수록 감소되는 반면, 2개의 양자값(q_f, q_c)의 중앙((q_f + q_c)/2)에 가까울수록 증가된다.

[0087] 확률 분포 계산부(232)는 온도 파라미터 계산부(231)에서 획득된 온도 파라미터(β)와 거리 스코어 계산부(220)에서 계산된 거리 기반 스코어(n_{x_n})를 수학식 4의 커널 소프트 아그맥스 함수에 대입하여 정규화값(x_n)의 거리 기반 확률(m_{x_n})을 획득한다.

[0088] 그리고 양자화값 획득부(233)는 획득된 거리 기반 확률(m_{x_n})을 기반으로 양자화값(ø(x_n))을 획득한다. 이때 양자화값 획득부(233)가 수학식 5에 따라 선택된 2개의 양자값(q_f, q_c)에 계산된 2개의 거리 기반 확률(m_{x_n})을 가중 평균하여 양자화값(ø(x_n))을 획득하는 경우, 가변되는 온도 파라미터(β)에 의해 여전히 양자화 오차가 발생될 수 있다.

[0089] 수학식 6에 따른 온도 파라미터(β)가 반영하여 수학식 5에 따라 계산되는 양자화값(ø(x_n))은 수학식 7과 같이 나타난다.

수학식 7

[0090]
$$\phi(x_n) = \begin{cases} q_f + \frac{1}{e^{\gamma}+1} & \text{if } |x_n - q_f| < 0.5 \\ q_c - \frac{1}{e^{\gamma}+1} & \text{if } |x_n - q_c| < 0.5 \end{cases}$$

[0091] 즉 양자화값(ø(x_n))이 선택된 2개의 양자값(q_f, q_c) 중 하나로 획득되지 않아 양자화 오차가 발생된다.

[0092] 이에 본 실시예의 양자화값 획득부(233)는 양자화 오차가 제거될 수 있도록 선택된 2개의 양자값(q_f, q_c)을 수학식 8과 같이 변형한다.

수학식 8

[0093]
$$q_f^* = q_f - \frac{1}{e^{\gamma}-1}$$
$$q_c^* = q_c + \frac{1}{e^{\gamma}-1}$$

[0094] 그리고 양자화값 획득부(233)는 변형된 2개의 양자값(q_f^*, q_c^*)에 따라 수학식 5를 수정한 수학식 9에 따라 양자화값(ø(x_n))을 획득한다.

수학식 9

[0095]
$$\phi(x_n) = \sum_{i \in \{f,c\}} m_{x_n}(q_i) q_i^*$$

[0096] 도 8은 본 실시예에 따른 양자화기(200)의 정규화값에 따른 양자화값의 변화와 이의 미분 그래프의 변화를 나타낸다.

[0097] 도 8에 도시된 바와 같이, 본 실시예에 따른 양자화기(200)는 대상값(x)을 정규화한 정규화값(x_n)에 대해 이산 함수를 이용하는 이상적인 이산화기와 동일하게 양자화값($ø(x_n)$)을 획득할 수 있어 양자화 오차가 발생되지 않는다. 뿐만 아니라 정규화값(x_n)과 양자값(q_f, q_c) 사이의 거리에 따라 가변되는 온도 파라미터(β)에 의해 미분값이 0으로 수렴되지 않으므로, 인공 신경망의 학습시에 손실 역전파가 가능하다. 특히 동일한 함수에 기반하여 순방향 전파와 역방향 전파가 수행될 수 있어, 그래디언트 불일치 문제를 야기하지 않는다.

[0098] 도 9는 본 발명의 일 실시예에 따른 인공 신경망을 위한 양자화 방법을 나타낸다.

[0099] 도 3을 참조하여, 도 9의 인공 신경망을 위한 양자화 방법을 설명하면, 양자화 대상이 되는 대상값(x)이 입력된다(S10). 여기서 대상값(x)은 인공 신경망에 인가되는 입력맵(또는 특징맵)의 원소이거나 학습에 의해 업데이트되는 가중치맵의 원소일 수 있다. 대상값(x)이 입력되면, 입력된 대상값(x)을 클리핑하여 클리핑값(x_c)을 획득하고, 클리핑값(x_c)을 기지정된 양자화 범위에 대응하도록 정규화하여 정규화값(x_n)을 획득한다(S20). 여기서 클리핑값(x_c)은 일예로 수학식 1에 따라 획득될 수 있으며, 정규화값(x_n)은 수학식 2에 따라 획득될 수 있다.

[0100] 정규화값(x_n)이 획득되면, 획득된 정규화값(x_n)에 가장 인접한 2개의 양자값(q_f, q_c)을 선택한다(S30). 여기서 2개의 양자값(q_f, q_c)은 각각 내림 함수와 올림 함수에 정규화값(x_n)을 대입하여 이용하여 획득할 수 있다.

[0101] 2개의 양자값(q_f, q_c)이 선택되면, 선택된 2개의 양자값(q_f, q_c)과 정규화값(x_n) 각각 사이의 차에 기반하여, 2개의 양자값(q_f, q_c) 각각과 정규화값(x_n)의 거리를 나타내는 2개의 거리 기반 스코어($n(x_n,q_f)$, $n(x_n,q_c)$)를 수학식 3에 따라 계산한다(S40).

[0102] 그리고 선택된 2개의 양자값(q_f, q_c)과 정규화값(x_n) 각각 사이의 거리에 따라 가변되는 온도 파라미터(β)를 수학식 6에 따라 계산한다(S50). 수학식 6에 따르면, 온도 파라미터(β)는 선택된 2개의 양자값(q_f, q_c)을 커널 함수(k_{x_n})에 대입한 커널값($k_{x_n}(q_i)$)을 2개의 거리 기반 스코어($n(x_n,q_i)$) 중 대응하는 거리 기반 스코어에 가중한 커널 가중 거리 기반 스코어 사이의 차에 따라 가변되는 파라미터로서, 정규화값(x_n)이 2개의 양자값(q_f, q_c) 중 하나에 가까울수록 감소되는 반면, 2개의 양자값(q_f, q_c)의 중앙($(q_f + q_c)/2$)에 가까울수록 증가된다.

[0103] 온도 파라미터(β)가 계산되면, 계산된 온도 파라미터(β)와 커널 함수(k_{x_n})로 가중된 거리 기반 스코어($n(x_n,q_f)$, $n(x_n,q_c)$)를 이용하여 수학식 4에 따라 거리 기반 확률(m_{x_n})을 계산한다(S70). 거리 기반 확률(m_{x_n}) 또한 2개의 양자값(q_f, q_c) 각각에 대응하여 개별적으로 계산될 수 있다.

[0104] 한편 양자값(q_f, q_c)과 정규화값(x_n) 각각 사이의 거리에 따라 가변되는 온도 파라미터(β)에 의해 발생하는 양자화 오차를 제거하기 위해, 선택된 2개의 양자값(q_f, q_c)을 온도 파라미터(β)에 대응하여 수학식 8과 같이 변형하여 변형된 양자값(q_f^*, q_c^*)을 획득한다(S70)

[0105] 그리고 계산된 거리 기반 확률(m_{x_n})과 변형된 양자값(q_f^*, q_c^*)이 획득되면, 수학식 9에 따라 계산된 거리 기반 확률(m_{x_n})을 변형된 양자값(q_f^*, q_c^*)에 가중 합하여, 양자화값($ø(x_n)$)을 획득하여 출력한다(S80).

[0106] 본 발명에 따른 방법은 컴퓨터에서 실행시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0107] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0108] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0109] 200: 양자화기 210: 정규화부

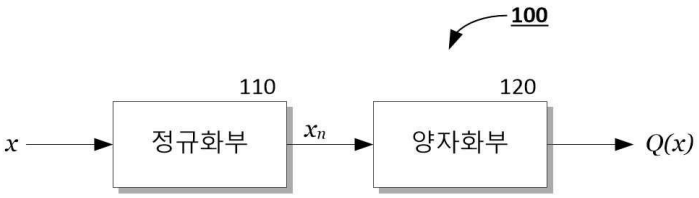
220: 거리 스코어 계산부 230: 확률 분포 기반 양자화부

231: 온도 파라미터 계산부 232: 확률 분포 계산부

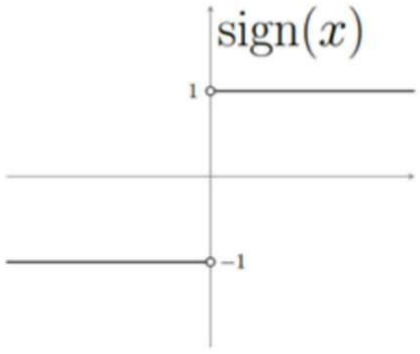
233: 양자화값 획득부

도면

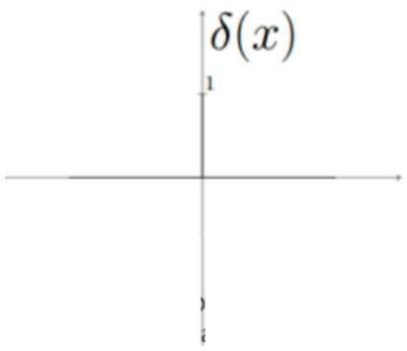
도면1



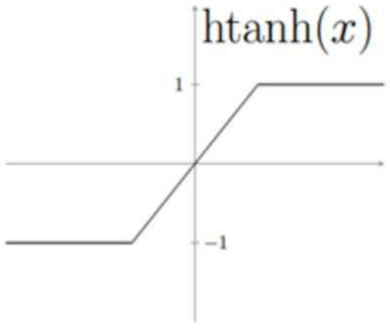
도면2



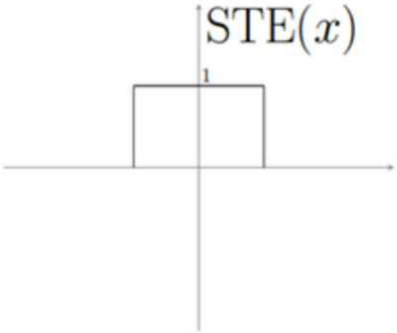
(a)



(b)

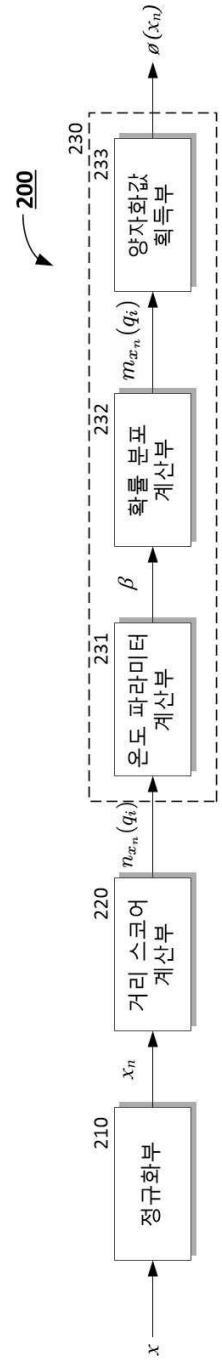


(c)

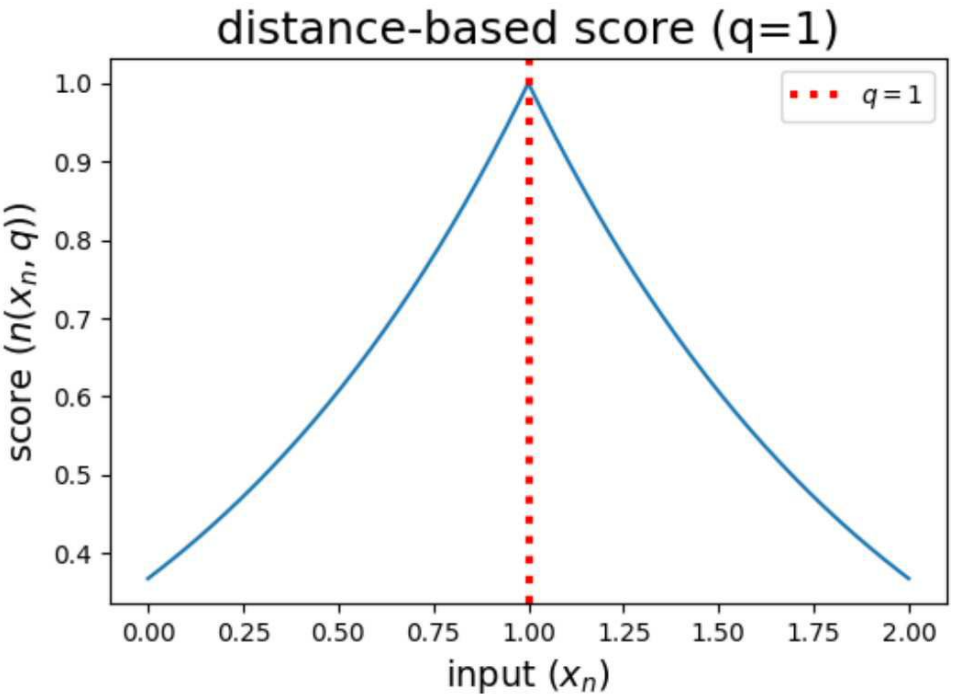


(d)

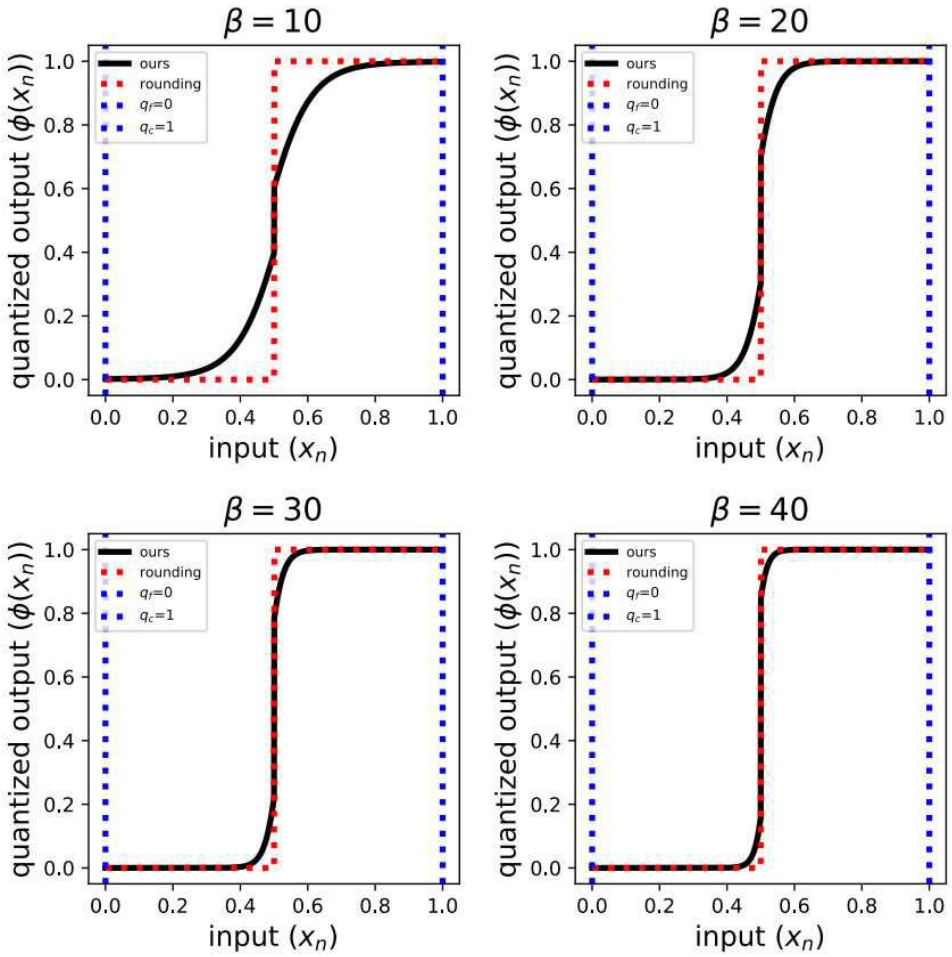
도면3


200
230
233
양자화값 획득부
$\phi(x_n)$
232
확률 분포 계산부
$m_{x_n}(q_i)$
231
온도 파라미터 계산부
β
220
거리 스코어 계산부
$m_{x_n}(q_i)$
210
정규화부
x_n
x

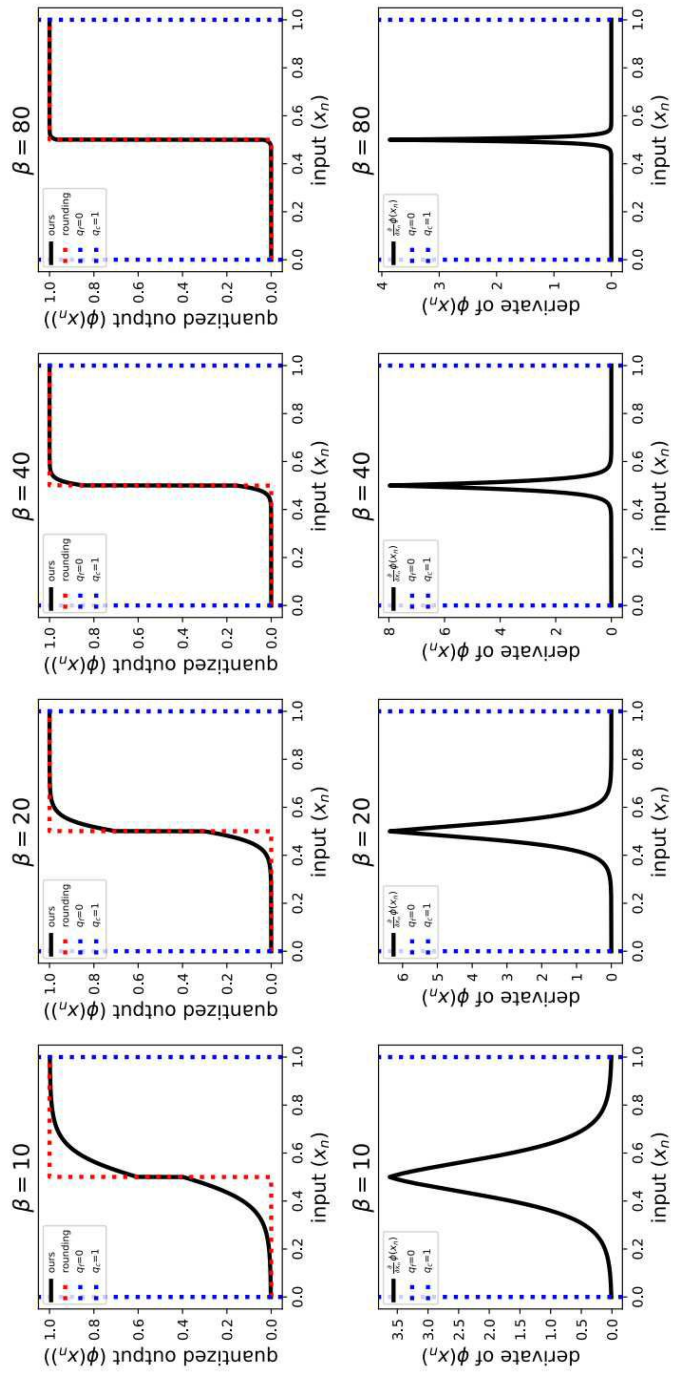
도면4


distance-based score (q=1)
q = 1
score (n(x_n, q))
input (x_n)

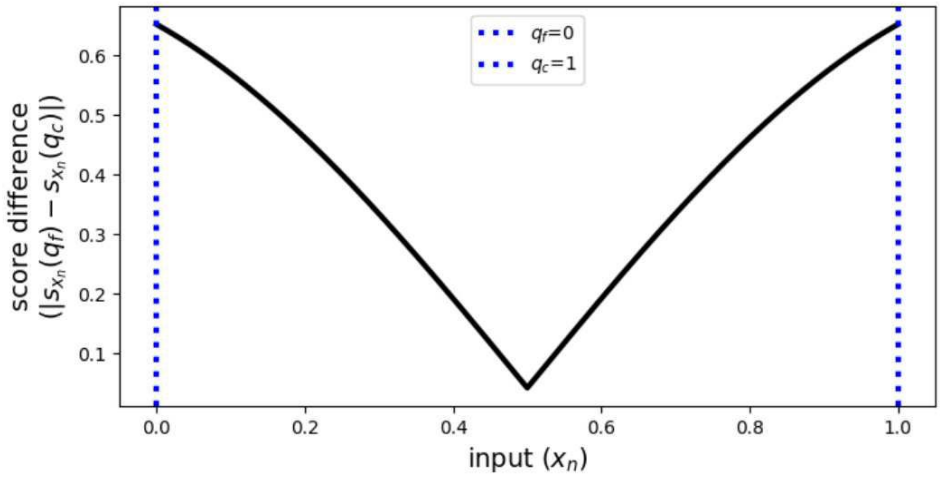
도면5


β = 10
β = 20
β = 30
β = 40
ours
rounding
q_f=0
q_c=1
quantized output (φ(x_n))
input (x_n)

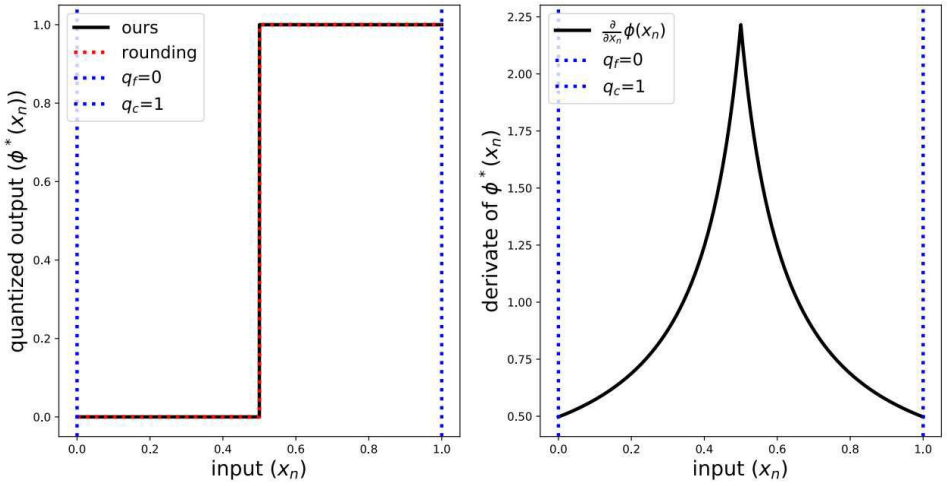
도면6


β = 10
β = 20
β = 40
β = 80
ours
rounding
q_r=0
q_r=1
quantized output ($\phi(x_n)$)
derivate of $\phi(x_n)$
input (x_n)

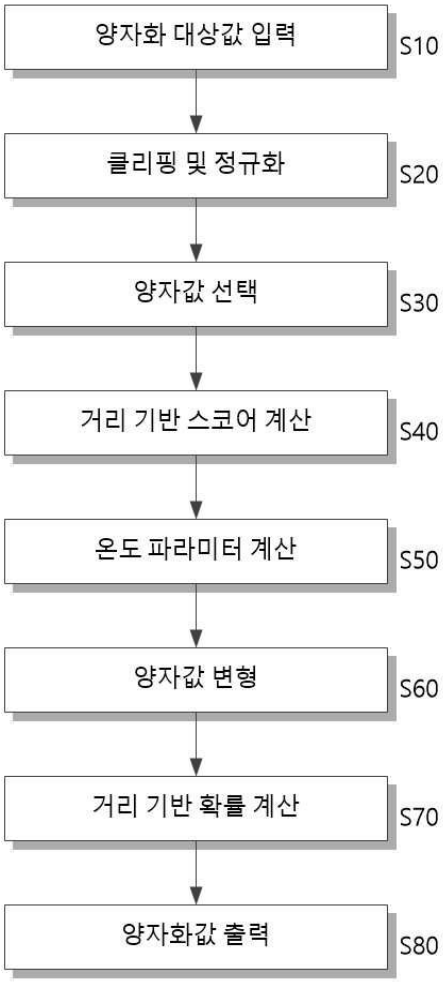
도면7


q_f=0
q_c=1
score difference
$(|s_{x_n}(q_f) - s_{x_n}(q_c)|)$
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0
input (x_n)

도면8


ours
rounding
q_f=0
q_c=1
quantized output ($\phi^*(x_n)$)
0.0
0.2
0.4
0.6
0.8
1.0
input (x_n)
$\frac{\partial}{\partial x_n}\phi(x_n)$
q_f=0
q_c=1
derivate of $\phi^*(x_n)$
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
input (x_n)

도면9


양자화 대상값 입력
S10
클리핑 및 정규화
S20
양자값 선택
S30
거리 기반 스코어 계산
S40
온도 파라미터 계산
S50
양자값 변형
S60
거리 기반 확률 계산
S70
양자화값 출력
S80