

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일

2025년 7월 3일 (03.07.2025)



(10) 국제공개번호

WO 2025/143388 A1

(51) 국제특허분류:

G06N 3/0495 (2023.01)

G06N 3/045 (2023.01)

G06N 3/0442 (2023.01)

(21) 국제출원번호:

PCT/KR2024/006433

(22) 국제출원일:

2024년 5월 13일 (13.05.2024)

(25) 출원언어:

한국어

(26) 공개언어:

한국어

(30) 우선권정보:

10-2023-0194206 2023년 12월 28일 (28.12.2023) KR

(71) 출원인: 이화여자대학교 산학협력단 (EWH A
UNIVERSITY - INDUSTRY COLLABORATION
FOUNDATION) [KR/KR]; 03760 서울특별시 서대문구
이화여대길 52 (KR).

(72) 발명자: 심재형 (SIM, Jae Hyeon); 04187 서울특별시
마포구 만리재로 74, 207동 801호 (KR). 강지의 (KANG,
Ji Eui); 14547 경기도 부천시 원미구 소향로 181, 105동

2402호 (KR). 박지혜 (PARK, Ji Hye); 02512 서울특별시
동대문구 한천로 248, 111동 304호 (KR). 최소은 (CHOI,
So Eun); 15623 경기도 안산시 상록구 감골로 83, 605동
1403호 (KR).

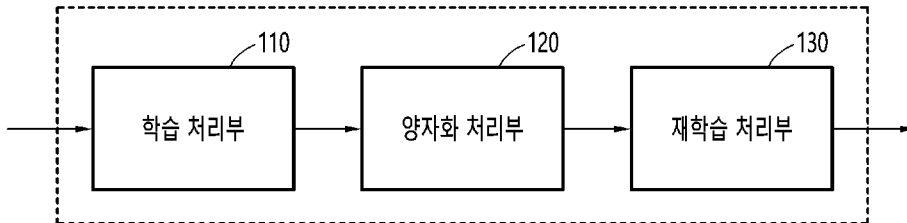
(74) 대리인: 김연권 (KIM, Youn Gwon); 05854 서울특별시
송파구 송파대로 201 B동 1416호 (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의
국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT,
AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH,
CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ,
EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR,
HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH,
KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD,
MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO,
NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW,
SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) Title: APPARATUS AND METHOD FOR PROCESSING QUANTIZATION ARTIFICIAL INTELLIGENCE LEARNING
USING ACCURACY INFORMATION AND SIMILARITY INFORMATION

(54) 발명의 명칭: 정확도 정보 및 유사도 정보를 이용한 양자화 인공지능 학습 처리 장치 및 방법

100



110 ... Learning processing unit

120 ... Quantization processing unit

130 ... Relearning processing unit

(57) Abstract: The present invention relates to an apparatus and method for processing quantization artificial intelligence learning using accuracy information and similarity information, and the apparatus for processing quantization artificial intelligence learning, according to an embodiment of the present invention, may comprise: a learning processing unit that outputs learning data by learning input data by using an artificial intelligence model; a quantization processing unit that outputs quantized data by quantizing the learning data; and a relearning processing unit that relearns the quantized data by considering a ratio between accuracy information and distribution information, which are based on a loss function, with respect to the learning data and the quantized data.

(57) 요약서: 본 발명은 정확도 정보 및 유사도 정보를 이용한 양자화 인공지능 학습 처리 장치 및 방법에 관한 것으로, 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치는 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력하는 학습 처리부, 상기 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력하는 양자화 처리부 및 상기 학습 데이터와 상기 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 상기 양자화 데이터를 재학습하는 재학습 처리부를 포함할 수 있다.

[다음 쪽 계속]



(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의
역내 권리의 보호를 위하여): ARIPO (BW, CV, GH, GM,
KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ,
UG, ZM, ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ,
TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE,
ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC,
ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

공개:

— 국제조사보고서와 함께 (조약 제21조(3))

명세서

발명의 명칭: 정확도 정보 및 유사도 정보를 이용한 양자화 인공지능 학습 처리 장치 및 방법

기술분야

- [1] 본 발명은 정확도 정보 및 유사도 정보를 이용한 양자화 인공지능 학습 처리 장치 및 방법에 관한 것으로, 보다 구체적으로, 인공지능 학습 데이터를 양자화 후 재 학습할 시 학습 데이터와 양자화 데이터 간에 손실에 대한 정확도 정보와 분포에 대한 유사도 정보를 함께 학습하여 정확도와 유사도 간의 비율을 조정함에 따라 양자화 인공지능 학습 결과에 대한 세일리언시(saliency)를 증가시키는 기술에 관한 것이다.
- [2] 본 발명은 2024년도 정부(과학기술정보통신부)의 재원으로 한국연구재단(No.RS-2023-00213548)과 정보통신기획평가원(No.RS-2022-00155966, 인공지능융합혁신인재양성(이화여자대학교), 국가과제고유번호: 1711179344, No.2021-0-02068, 인공지능 혁신 허브 연구 개발)의 지원을 받아 수행된 연구 결과이다.

배경기술

- [3] 명시적으로 프로그래밍하지 않고도 데이터를 학습하고 예측 또는 의사 결정을 내릴 수 있는 고유한 능력을 갖춘 AI(Artificial Intelligence) 기술은 데이터의 복잡성과 양이 증가하는 문제를 해결할 수 있는 유망한 솔루션을 제공한다.
- [4] 데이터 분석 분야에서 LSTM(Long Short-Term Memory)은 인공지능 학습 모델 중 하나인, 순환 신경망(Recurrent Neural Network, RNN) 기법의 하나로, 시계열 데이터를 처리하는데 능숙하기 때문에 주목받고 있다.
- [5] 시계열 데이터는 시간적 상관관계가 강하고 장기간에 걸친 패턴이나 추세를 포함할 수 있고, 대표적으로 실내 환경 데이터가 존재한다.
- [6] 예를 들어, 실내 환경 데이터는 건물 내에서 소비된 에너지와 관련한 데이터로서, 온도, 습도, 점유 수준, 에너지 사용량 등과 같은 다양한 변수 데이터를 고려하여 구성될 수 있다.
- [7] 또한, 실내 환경 데이터는 HVAC(Heating, Ventilation and Air Conditioning)의 제어와 관련된다.
- [8] 시간적 종속성으로 인해 종종 어려움을 겪는 기존 방법과 달리, LSTM은 장기간에 걸쳐 메모리에 정보를 유지함으로써 데이터 시퀀스에서 효과적으로 학습되도록 설계되었다.
- [9] 그러나, LSTM은 실제 애플리케이션에서 실용성을 제한하는 몇가지 단점이 있는데, 연산 복잡성과 리소스 집약적인 특성으로 인해 연산 능력과 메모리가 제한된다는 것이다.

- [10] 이러한 문제를 완화하기 위해서는 양자화(quantization)가 고려 및 사용되고 있지만 근사 오류(approximation error) 도입으로 인해 모델 성능이 저하되는 경우가 많다.
- [11] 다시 말해, 종래의 양자화는 모델에서 사용되는 숫자의 정밀도를 낮추는 작업이 수반되어서, 정보 손실로 이어져 결과적으로 정확도를 낮춘다.
- [12] 인공지능 학습 처리 모델에 대하여 모델의 효율성과 정확성 사이의 절충점을 찾는 것을 에너지 데이터 분석에서 LSTM 네트워크를 활용함에 있어서 중요한 과제로 볼 수 있다.

발명의 상세한 설명

기술적 과제

- [13] 본 발명은 인공지능 학습 데이터를 양자화 후 재 학습할 시 학습 데이터와 양자화 데이터 간에 손실에 대한 정확도 정보와 분포에 대한 유사도 정보를 함께 학습하여 정확도와 유사도 간의 비율을 조정함에 따라 양자화 인공지능 학습 결과에 대한 세일리언시(saliency)를 증가시키는 것을 목적으로 한다.
- [14] 본 발명은 인공지능 학습 처리에서 모델의 효율성과 정확성 사이의 절충점을 확인하여 시간적 의존성과 문맥적 관련성을 효과적으로 확인하여 인공지능 학습 결과를 도출하는 것을 목적으로 한다.
- [15] 본 발명은 LSTM과 어텐션 매커니즘(attention mechanism)에 기반하여 학습된 인공지능 학습 결과의 분포와 인공지능 학습 결과에 대한 양자화 처리후의 분포에 대하여 발산을 사용하여 근접한 근사치를 효과적으로 유지함에 따라 양 분포의 불일치를 최소화하여 인공지능 학습 결과에 대한 효율성과 정확성을 확보하는 것을 목적으로 한다.
- [16] 본 발명은 인공지능 학습 결과에 대한 효율성과 정확성을 확보함에 따라 속도나 정확도를 저하시키지 않으면서 복잡한 다변량 예측 작업을 처리할 수 있음에 시계열 데이터에 기반한 예측이 요구되는 다양한 기술 분야에서 다양한 애플리케이션으로 적용될 수 있는 양자화 인공지능 학습 처리 장치 및 방법을 제공하는 것을 목적으로 한다.

과제 해결 수단

- [17] 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치는 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력하는 학습 처리부, 상기 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력하는 양자화 처리부 및 상기 학습 데이터와 상기 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 상기 양자화 데이터를 재학습하는 재학습 처리부를 포함할 수 있다.
- [18] 상기 재학습 처리부는 상기 학습 데이터와 상기 양자화 데이터의 왜도(skewness)와 첨도(kurtosis)의 차이 값에 기반하여 상기 정확도 정보와 상기 분포

정보에 대한 절대적인 비중을 계산하고, 상기 계산된 절대적인 비중에 기반하여 상기 정확도 정보와 상기 분포 정보에 대한 상대적인 비율을 계산할 수 있다.

- [19] 상기 재학습 처리부는 상기 양자화 데이터의 왜도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 왜도 값의 차이 값 간의 비율에 기반하여 유사도와 관련된 파라미터(x)를 계산하고, 상기 계산된 파라미터(x)에 기반하여 상기 절대적인 비중에 대한 파라미터(m')을 계산하며, 상기 계산된 파라미터(m')에 기반하여 상기 상대적인 비율에 대한 파라미터(m)을 계산할 수 있다.
- [20] 상기 재학습 처리부는 상기 양자화 데이터의 첨도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 첨도 값의 차이 값 간의 비율에 기반하여 정확도와 관련된 파라미터(y)를 계산하고, 상기 계산된 파라미터(y)에 기반하여 상기 절대적인 비중에 대한 파라미터(n')을 계산하며, 상기 계산된 파라미터(n')에 기반하여 상기 상대적인 비율에 대한 파라미터(n)을 계산할 수 있다.
- [21] 상기 재학습 처리부는 상기 손실 함수를 평균제곱오차(mean square error, MSE) 손실 연산과 상대 엔트로피 연산으로 구성하고, 상기 평균제곱오차 손실 연산에 상기 정확도 정보에 대한 상대적인 비율과 관련된 파라미터(n)를 적용하고, 상기 상대 엔트로피 연산에 상기 분포 정보에 대한 상대적인 비율과 관련된 파라미터(m)을 적용한 연산 결과에 따라 상기 양자화 데이터를 재학습할 수 있다.
- [22] 상기 재학습 처리부는 상기 파라미터(m)와 상기 파라미터(n)의 합이 "1"에 근사하도록 상기 양자화 데이터를 재학습하여 상기 학습 데이터와 상기 양자화 데이터 사이의 분포 불일치를 감소시킬 수 있다.
- [23] 상기 재학습 처리부는 상기 손실 함수에서 상기 파라미터(m)의 증가에 따라 유사도에 비중을 증가하고, 상기 파라미터(n)의 증가에 따라 정확도에 비중을 증가하여 상기 양자화 데이터를 재학습할 수 있다.
- [24] 상기 학습 처리부는 상기 인공지능 모델을 RNN(Recurrent Neural Network) 중 하나인 LSTM(Long Short Term Memory)와 어텐션 매커니즘(Attention Mechanism)의 결합 모델로 이용할 수 있다.
- [25] 본 발명의 일실시예에 따르면 학습 처리부에서, 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력하는 단계, 양자화 처리부에서, 상기 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력하는 단계 및 재학습 처리부에서, 상기 학습 데이터와 상기 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 상기 양자화 데이터를 재학습하는 단계를 포함할 수 있다.
- [26] 상기 양자화 데이터를 재학습하는 단계는, 상기 학습 데이터와 상기 양자화 데이터의 왜도(skewness)와 첨도(kurtosis)의 차이 값에 기반하여 상기 양자화 데이터의 왜도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 왜도 값의 차이 값 간의 비율에 기반하여 유사도와 관련된 파라미터(x)를 계산하고, 상기 계산된 파라미터(x)에 기반하여 절대적인 비중에 대한 파라미터(m')을 계산하며, 상기 계산된 파라미터(m')에 기반하여 상대적인 비율에 대한 파라미터(m)을

계산하는 단계 및 상기 양자화 데이터의 침도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 침도 값의 차이 값 간의 비율에 기반하여 정확도와 관련된 파라미터(y)를 계산하고, 상기 계산된 파라미터(y)에 기반하여 상기 절대적인 비율에 대한 파라미터(n')을 계산하며, 상기 계산된 파라미터(n')에 기반하여 상기 상대적인 비율에 대한 파라미터(n)을 계산하는 단계를 포함할 수 있다.

발명의 효과

- [27] 본 발명은 인공지능 학습 데이터를 양자화 후 재 학습할 시 학습 데이터와 양자화 데이터 간에 손실에 대한 정확도 정보와 분포에 대한 유사도 정보를 함께 학습하여 정확도와 유사도 간의 비율을 조정함에 따라 양자화 인공지능 학습 결과에 대한 세일리언시(saliency)를 증가시킬 수 있다.
- [28] 본 발명은 인공지능 학습 처리에서 모델의 효율성과 정확성 사이의 절충점을 확인하여 시간적 의존성과 문맥적 관련성을 효과적으로 확인하여 인공지능 학습 결과를 도출할 수 있다.
- [29] 본 발명은 LSTM과 어텐션 매커니즘에 기반하여 학습된 인공지능 학습 결과의 분포와 인공지능 학습 결과에 대한 양자화 처리후의 분포에 대하여 발산을 사용하여 근접한 근사치를 효과적으로 유지함에 따라 양 분포의 불일치를 최소화하여 인공지능 학습 결과에 대한 효율성과 정확성을 확보할 수 있다.
- [30] 본 발명은 인공지능 학습 결과에 대한 효율성과 정확성을 확보함에 따라 속도나 정확도를 저하시키지 않으면서 복잡한 다변량 예측 작업을 처리할 수 있음에 시계열 데이터에 기반한 예측이 요구되는 다양한 기술 분야에서 다양한 애플리케이션으로 적용될 수 있는 양자화 인공지능 학습 처리 장치 및 방법을 제공할 수 있다.

도면의 간단한 설명

- [31] 도 1은 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치를 설명하는 도면이다.
- [32] 도 2는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 결과를 설명하는 도면이다.
- [33] 도 3 및 도 4는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 성능을 설명하는 도면이다.
- [34] 도 5는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 방법을 설명하는 도면이다.

발명의 실시를 위한 형태

- [35] 이하, 본 문서의 다양한 실시 예들이 첨부된 도면을 참조하여 기재된다.
- [36] 실시 예 및 이에 사용된 용어들은 본 문서에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 해당 실시 예의 다양한 변경, 균등물, 및/또는 대체물을 포함하는 것으로 이해되어야 한다.

- [37] 하기에서 다양한 실시 예들을 설명에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.
- [38] 그리고 후술되는 용어들은 다양한 실시 예들에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [39] 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.
- [40] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.
- [41] 본 문서에서, "A 또는 B" 또는 "A 및/또는 B 중 적어도 하나" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다.
- [42] "제1," "제2," "첫째," 또는 "둘째," 등의 표현들은 해당 구성요소들을, 순서 또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.
- [43] 어떤(예: 제1) 구성요소가 다른(예: 제2) 구성요소에 "(기능적으로 또는 통신적으로) 연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 상기 어떤 구성요소가 상기 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제3 구성요소)를 통하여 연결될 수 있다.
- [44] 본 명세서에서, "~하도록 구성된(또는 설정된)(configured to)"은 상황에 따라, 예를 들면, 하드웨어적 또는 소프트웨어적으로 "~에 적합한," "~하는 능력을 가지는," "~하도록 변경된," "~하도록 만들어진," "~를 할 수 있는," 또는 "~하도록 설계된"과 상호 호환적으로(interchangeably) 사용될 수 있다.
- [45] 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는" 것을 의미할 수 있다.
- [46] 예를 들면, 문구 "A, B, 및 C를 수행하도록 구성된(또는 설정된) 프로세서"는 해당 동작을 수행하기 위한 전용 프로세서(예: 임베디드 프로세서), 또는 메모리 장치에 저장된 하나 이상의 소프트웨어 프로그램들을 실행함으로써, 해당 동작들을 수행할 수 있는 범용 프로세서(예: CPU 또는 application processor)를 의미할 수 있다.
- [47] 또한, '또는'이라는 용어는 배타적 논리합 'exclusive or' 이기보다는 포함적인 논리합 'inclusive or' 를 의미한다.
- [48] 즉, 달리 언급되지 않는 한 또는 문맥으로부터 명확하지 않는 한, 'x가 a 또는 b를 이용한다'라는 표현은 포함적인 자연 순열들(natural inclusive permutations) 중 어느 하나를 의미한다.
- [49] 이하 사용되는 '부', '기' 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어, 또는, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

- [50] 도 1은 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치를 설명하는 도면이다.
- [51] 도 1은 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 구성 요소를 예시하고, 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치는 인공지능 학습 데이터를 양자화 후 재 학습할 시 손실 함수의 정확도 정보와 분포 정보를 함께 학습하여 정확도와 유사도 간의 비율을 조정함에 따라 양자화 인공지능 학습 결과에 대한 세일리언시(saliency)를 증가시키는 기술을 구현한다.
- [52] 도 1을 참고하면, 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치(100)는 학습 처리부(110), 양자화 처리부(120) 및 재학습 처리부(130)를 포함한다.
- [53] 일례로, 학습 처리부(110)는 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력한다.
- [54] 예를 들어, 학습 처리부(110)는 인공지능 모델을 RNN(Recurrent Neural Network) 중 하나인 LSTM(Long Short Term Memory)와 어텐션 매커니즘(Attention Mechanism)의 결합 모델로 이용한다.
- [55] 학습 처리부(110)는 긴 시퀀스 데이터 내의 복잡한 표현을 정확하게 학습하는 인공지능 모델을 이용하여 입력 데이터를 학습한다.
- [56] 예를 들어, 입력 데이터는 에너지 사용량 예측과 같이 추후 예측이 필요한 데이터들을 예측 학습하기 위해 활용되는 시계열 데이터로 구성될 수 있다.
- [57] 본 발명의 일실시예에 따른 양자화 처리부(120)는 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력한다.
- [58] 양자화 처리부(120)는 학습 데이터를 양자화하여 계산 수요와 모델 크기를 줄이기 위해 수치 값의 정밀도를 낮춘 양자화 데이터를 출력한다.
- [59] 여기서, 양자화 데이터는 리소스가 제한된 디바이스에서 분포의 효율성을 도모할 수 있다.
- [60] 다만, 양자화 데이터는 데이터의 미묘하지만 중요한 차이가 제거됨에 따라 정확도가 감소하는 문제점이 일부 존재한다.
- [61] 본 발명의 일실시예에 따르면 재학습 처리부(130)는 학습 데이터와 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 양자화 데이터를 재학습한다.
- [62] 일례로, 재학습 처리부(130)는 학습 데이터와 양자화 데이터의 왜도(skewness)와 첨도(kurtosis)의 차이 값에 기반하여 정확도 정보와 분포 정보에 대한 절대적인 비중을 계산하고, 계산된 절대적인 비중에 기반하여 상기 정확도 정보와 상기 분포 정보에 대한 상대적인 비율을 계산할 수 있다.
- [63] 왜도는 분포 대칭으로서 비대칭도를 나타낼 수 있고, 첨도는 편평한 분포를 나타낸다.
- [64] 본 발명의 일실시예에 따르면 재학습 처리부(130)는 하기 수학적 식 1에 기반하여 유사도 관련 파라미터(x)와 정확도 관련 파라미터(y)를 연산할 수 있다.

[65] [수학식 1]

$$[66] \quad x = \left| \frac{\alpha - \alpha'}{\alpha' + 10^{-7}} \right|$$

$$y = \left| \frac{\beta - \beta'}{\beta' + 10^{-7}} \right|$$

[67] 수학식 1에서, x는 유사도 관련 파라미터를 나타낼 수 있고, y는 정확도 관련 파라미터를 나타낼 수 있으며, α 는 학습데이터의 왜도를 나타낼 수 있고, α' 는 양자화 데이터의 왜도를 나타낼 수 있으며, e는 자연 상수를 나타낼 수 있다.

[68] 학습 데이터는 양자화 이전의 데이터이고, 양자화 데이터는 양자화 이후의 데이터일 수 있다.

[69] 본 발명의 일실시예에 따르면 재학습 처리부(130)는 양자화 데이터의 왜도 값에 대한 학습 데이터의 왜도 값과 상기 양자화 데이터의 왜도 값의 차이 값 간의 비율에 기반하여 유사도와 관련된 파라미터(x)를 계산하고, 계산된 파라미터(x)에 기반하여 절대적인 비중에 대한 파라미터(m')을 계산하며, 계산된 파라미터(m')에 기반하여 상대적인 비율에 대한 파라미터(m)을 계산할 수 있다.

[70] 일례로, 재학습 처리부(130)는 양자화 데이터의 첨도 값에 대한 학습 데이터의 왜도 값과 양자화 데이터의 첨도 값의 차이 값 간의 비율에 기반하여 정확도와 관련된 파라미터(y)를 계산하고, 계산된 파라미터(y)에 기반하여 절대적인 비중에 대한 파라미터(n')을 계산하며, 계산된 파라미터(n')에 기반하여 상대적인 비율에 대한 파라미터(n)을 계산할 수 있다.

[71] 본 발명의 일실시예에 따르면 재학습 처리부(130)는 하기 수학식 2에 기반하여 유사도 관련 파라미터(x)와 정확도 관련 파라미터(y)를 이용하여 파라미터(n')과 파라미터(n)를 연산할 수 있다.

[72] [수학식 2]

$$[73] \quad m' = \frac{x}{\frac{1}{x + 10^{-7}} + \frac{1}{y + 10^{-7}}}$$

$$n' = \frac{y}{\frac{1}{x + 10^{-7}} + \frac{1}{y + 10^{-7}}}$$

$$m = \frac{m'}{m' + n'}$$

$$n = \frac{n'}{m' + n'}$$

[74] 수학식 2에서, m'은 파라미터(x)에 대한 절대적인 비중에 대한 파라미터를 나타낼 수 있고, m은 파라미터(m')에 기반한 상대적인 비율에 대한 파라미터를 나타낼 수 있고, n'은 파라미터(y)에 대한 절대적인 비중에 대한 파라미터를 나타낼 수 있고, n은 파라미터(n')에 기반한 상대적인 비율에 대한 파라미터를 나타낼 수 있다.

- [75] 본 발명의 일실시예에 따르면 재학습 처리부(130)는 손실 함수를 평균제곱오차(mean square error, MSE) 손실 연산과 상대 엔트로피 연산으로 구성하고, 평균제곱오차 손실 연산에 상기 정확도 정보에 대한 상대적인 비율과 관련된 파라미터(n)를 적용하고, 상대 엔트로피 연산에 상기 분포 정보에 대한 상대적인 비율과 관련된 파라미터(m)를 적용한 연산 결과에 따라 양자화 데이터를 재학습할 수 있다.
- [76] 일례로, 재학습 처리부(130)는 하기 수학적 식 3을 이용하여 양자화 데이터를 재학습할 수 있다.
- [77] [수학적 식 3]
- [78]
$$Loss = \frac{n}{N} \sum (y_i - Q(x_i))^2 + m * \lambda [\int p(x) \log(\frac{p(x)}{q(x)}) dx]$$
- [79] 수학적 식 3은 평균제곱오차(mean square error, MSE) 손실을 연산하는 부분과 관련하여 양자화된 출력인 양자화 데이터 Q(xi)와 기존 출력인 학습 데이터 yi 그리고 수학적 식 2에 기반한 파라미터(n)와 총 샘플 수에 해당하는 N의 연산 결과와 상대 엔트로피 연산인 KL 발산 연산에 대하여 파라미터(m)를 적용하여 MSE와 KL 발산 연산 사이의 절충점을 결정하는 하이퍼 파라미터인 λ 의 연산을 나타낼 수 있다.
- [80] 또한, KL 발산 연산에서 p(x)는 양자화 전에 해당하는 학습 데이터의 분포를 나타내고, q(x)는 양자화 이후에 해당하는 양자화 데이터의 분포를 나타낼 수 있다.
- [81] 본 발명의 일실시예에 따른 재학습 처리부(130)는 파라미터(m)와 파라미터(n)의 합이 "1"에 근사하도록 양자화 데이터를 재학습하여 학습 데이터와 양자화 데이터 사이의 분포 불일치를 감소시킬 수 있다.
- [82] 일례로, 재학습 처리부(130)는 손실 함수에서 파라미터(m)의 증가에 따라 유사도에 비중을 증가하고, 파라미터(n)의 증가에 따라 정확도에 비중을 증가하여 양자화 데이터를 재학습할 수 있다.
- [83] 다시 말해, 재학습 처리부(130)는 수학적 식 3에서 파라미터(n)이 증가하면 정확도에 더 많은 비중을 두도록 연산 결과인 재학습 결과를 도출하고, 파라미터(m)이 증가하면 유사도에 더 많은 비중을 두도록 연산 결과인 재학습 결과를 도출함에 파라미터(m)와 파라미터(n)를 정확도와 유사도 간의 비율을 조정하는 값으로 활용한다.
- [84] 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치(100)는 양자화된 데이터가 최초 학습에 대하여 분포를 유지하여 손실 함수에 KL 발산을 통합함에 따라 양자화 프로세스 중에서 모델이 데이터의 중요(salient) 특징을 잘 인식하도록 합니다.
- [85] LSTM과 어텐션 매커니즘의 결합 모델은 Latte로 지칭할 수 있고, 양자화 및 정확도 정보와 분포 정보를 적용한 모델은 Q-Latte로 지칭할 수 있다.
- [86] 따라서, 본 발명은 인공지능 학습 데이터를 양자화 후 재 학습할 시 학습 데이터와 양자화 데이터 간에 손실에 대한 정확도 정보와 분포에 대한 유사도 정보를

함께 학습하여 정확도와 유사도 간의 비율을 조정함에 따라 양자화 인공지능 학습 결과에 대한 세일리언시(saliency)를 증가시킬 수 있다.

[87] 도 2는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 결과를 설명하는 도면이다.

[88] 도 2는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 결과와 함께 인공지능 학습 처리 과정에서 재학습 과정에 따른 학습 결과를 예시한다.

[89] 도 2를 참고하면, 본 발명의 일실시예에 따른 인공지능 학습 처리 장치는 입력 데이터를 학습하여 학습 결과로 학습 데이터(200)를 출력한다.

[90] 일례로, 인공지능 학습 처리 장치는 학습 데이터(200)에 대하여 양자화 처리를 수행하여 양자화 데이터(201)를 출력한다.

[91] 학습 데이터(200)와 양자화 데이터(201)는 왜도(skewness)와 첨도(kurtosis)의 차이가 존재한다.

[92] 인공지능 학습 처리 장치는 왜도와 첨도의 차이 값으로 양자화 데이터의 재학습에 활용할 수 있는 파라미터를 수학적 1 내지 수학적 3을 이용하여 도출한다.

[93] 본 발명의 일실시예에 따른 인공지능 학습 처리 장치는 수학적 1 내지 수학적 3을 이용하여 산출된 파라미터를 이용하여 양자화 데이터(201)를 재학습함에 따라 재학습 데이터(202)를 출력한다.

[94] 학습 데이터(200)와 재학습 데이터(202)는 왜도와 첨도의 값에 대한 차이가 줄어든다.

[95] 따라서, 본 발명은 인공지능 학습 처리에서 모델의 효율성과 정확성 사이의 절충점을 확인하여 시간적 의존성과 문맥적 관련성을 효과적으로 확인하여 인공지능 학습 결과를 도출할 수 있다.

[96] 도 3 및 도 4는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 성능을 설명하는 도면이다.

[97] 도 3은 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 성능과 관련하여 재학습 결과가 양자화를 통해 제거될 수 있는 차이 부분을 포함하고 있음에 따라 모델 성능 저하를 방지할 수 있음을 보여준다.

[98] 도 3을 참고하면, 그래프(300)는 양자화 전에 해당하는 학습 데이터를 나타내고, 그래프(310)는 양자화 이후에 해당하는 양자화 데이터를 나타내며, 그래프(320)는 양자화 데이터의 재학습 후를 나타내는 재학습 데이터를 나타낸다.

[99] 그래프(300)에 대하여 그래프(310)와 그래프(320)를 대비하면 그래프(310)는 그래프(300)로부터 손실된 데이터가 존재함을 확인할 수 있다.

[100] 그래프(300)와 그래프(310)를 대비하면 양자화는 분포를 왜곡하여 중요한 정보의 손실이 발생할 수 있음을 보여준다.

[101] LSTM 모델은 리소스가 제한된 디바이스에 배포하는 데는 더 효율적이지만, 실제 어플리케이션에 효과적으로 적용하기 위해서는 효율성과 정확성 사이에서 상충 관계를 고려하여 그래프(320)와 같이 재학습되어야 한다.

- [102] 반면에, 그래프(320)는 그래프(300)로부터 손실된 데이터가 그래프(310)에 대하여 적음을 확인할 수 있다.
- [103] 도 4는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 성능과 관련하여 재학습 적용 여부에 따른 학습 결과를 예시한다.
- [104] 도 4를 참고하면, 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치의 학습 성능과 관련하여 재학습 처리를 적용하지 않은 결과에 대하여 그래프(400)로 나타내고, 재학습 처리를 적용한 결과에 대하여 그래프(410)로 나타낸다.
- [105] 그래프(400)는 학습데이터(401)에 대하여 양자화를 적용한 제1 양자화 데이터(402) 및 양자화를 적용한 제2 양자화 데이터(403)를 나타낼 수 있다.
- [106] 그래프(410)는 학습데이터(411)에 대하여 양자화를 적용한 제1 양자화 데이터(412) 및 양자화를 적용한 제2 양자화 데이터(413)를 나타낼 수 있다.
- [107] 그래프(400)와 그래프(410)를 대비하면 분포 정보가 상대적으로 차이가 감소된 것을 확인할 수 있다.
- [108] 즉, 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치는 분포 정보를 고려하여 양자화에 따른 데이터 손실 및 차이를 감소한다.
- [109] 예를 들어, 양자화 인공지능 학습 처리 장치는 효율적이고 정확한 딥러닝 모델을 개발하여 성능을 향상시킨다.
- [110] 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 장치는 추론 비용을 크게 줄이면서 뛰어난 성능을 제공할 수 있음에 따라 시계열 예측 작업에 매우 효율적인 솔루션을 제공할 수 있다.
- [111] 본 발명은 양자화 인공지능 학습 처리 장치에 기반하여 장기적인 건물 에너지 예측부터 건물 내 여러 에너지 데이터의 동시 예측에 이르기까지 다양한 유형의 예측 작업에 적용할 수 있는 다재다능한 솔루션을 제공할 수 있다.
- [112] 본 발명은 속도나 정확도를 저하시키지 않으면서 복잡한 다변량 예측 작업을 처리할 수 있는 고성능의 효율적인 예측 모델에 대한 필요성에 따라 양자화 인공지능 학습 처리 장치로 구현할 수 있다.
- [113] 본 발명의 양자화 인공지능 학습 처리 장치는 시계열 예측의 새로운 기준을 제시하고, 강력한 예측 기능이 필수적인 여러 분야에서 다양한 실제 어플리케이션(application)에 적용될 수 있다.
- [114] 따라서, 본 발명은 LSTM과 어텐션 매커니즘에 기반하여 학습된 인공지능 학습 결과의 분포와 인공지능 학습 결과에 대한 양자화 처리후의 분포에 대하여 발산을 사용하여 근접한 근사치를 효과적으로 유지함에 따라 양 분포의 불일치를 최소화하여 인공지능 학습 결과에 대한 효율성과 정확성을 확보할 수 있다.
- [115] 도 5는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 방법을 설명하는 도면이다.
- [116] 도 5는 본 발명의 일실시예에 따른 양자화 인공지능 학습 처리 방법이 인공지능 학습 데이터를 양자화 후 재 학습할 시 손실 함수의 정확도 정보와 분포 정보를

함께 학습하여 정확도와 유사도 간의 비율을 조정하여 양자화 인공지능 학습 결과에 대한 세일리언시(saliency)를 증가시키는 절차를 예시한다.

- [117] 도 5를 참고하면, 단계(S501)에서 본 발명의 일실시에에 따른 양자화 인공지능 학습 처리 방법은 입력 데이터를 학습한다.
- [118] 즉, 본 발명의 일실시에에 따른 양자화 인공지능 학습 처리 방법은 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력할 수 있다.
- [119] 단계(S502)에서 본 발명의 일실시에에 따른 양자화 인공지능 학습 처리 방법은 학습 데이터의 양자화 처리를 수행한다.
- [120] 즉, 본 발명의 일실시에에 따른 양자화 인공지능 학습 처리 방법은 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력할 수 있다.
- [121] 단계(S503)에서 본 발명의 일실시에에 따른 양자화 인공지능 학습 처리 방법은 정확도 정보와 분포 정보 간의 비율을 고려하여 양자화 데이터를 재학습한다.
- [122] 즉, 본 발명의 일실시에에 따른 양자화 인공지능 학습 처리 방법은 학습 데이터와 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 상기 양자화 데이터를 재학습할 수 있다.
- [123] 따라서, 본 발명은 인공지능 학습 결과에 대한 효율성과 정확성을 확보함에 따라 속도나 정확도를 저하시키지 않으면서 복잡한 다변량 예측 작업을 처리할 수 있음에 시계열 데이터에 기반한 예측이 요구되는 다양한 기술 분야에서 다양한 애플리케이션으로 적용될 수 있는 양자화 인공지능 학습 처리 장치 및 방법을 제공할 수 있다.
- [124] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [125] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동

작하도록 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 분포되어서, 분포된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

- [126] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 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.
- [127] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

청구범위

- [청구항 1] 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력하는 학습 처리부;
상기 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력하는 양자화 처리부; 및
상기 학습 데이터와 상기 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 상기 양자화 데이터를 재학습하는 재학습 처리부를 포함하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.
- [청구항 2] 제1항에 있어서,
상기 재학습 처리부는 상기 학습 데이터와 상기 양자화 데이터의 왜도(skewness)와 첨도(kurtosis)의 차이 값에 기반하여 상기 정확도 정보와 상기 분포 정보에 대한 절대적인 비중을 계산하고, 상기 계산된 절대적인 비중을 기반으로 상기 정확도 정보와 상기 분포 정보에 대한 상대적인 비율을 계산하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.
- [청구항 3] 제2항에 있어서,
상기 재학습 처리부는 상기 양자화 데이터의 왜도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 왜도 값의 차이 값 간의 비율에 기반하여 유사도와 관련된 파라미터(x)를 계산하고, 상기 계산된 파라미터(x)에 기반하여 상기 절대적인 비중에 대한 파라미터(m')을 계산하며, 상기 계산된 파라미터(m')에 기반하여 상기 상대적인 비율에 대한 파라미터(m)을 계산하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.
- [청구항 4] 제2항에 있어서,
상기 재학습 처리부는 상기 양자화 데이터의 첨도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 첨도 값의 차이 값 간의 비율에 기반하여 정확도와 관련된 파라미터(y)를 계산하고, 상기 계산된 파라미터(y)에 기반하여 상기 절대적인 비중에 대한 파라미터(n')을 계산하며, 상기 계산된 파라미터(n')에 기반하여 상기 상대적인 비율에 대한 파라미터(n)을 계산하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.
- [청구항 5] 제2항에 있어서,
상기 재학습 처리부는 상기 손실 함수를 평균제곱오차(mean square error, MSE) 손실 연산과 상대 엔트로피 연산으로 구성하고, 상기 평균제곱오차 손실 연산에 상기 정확도 정보에 대한 상대적인 비율과 관련된 파라미터(n)를 적용하고, 상기 상대 엔트로피 연산에 상기 분포 정보에 대한 상대

적인 비율과 관련된 파라미터(m)을 적용한 연산 결과에 따라 상기 양자화 데이터를 재학습하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.

[청구항 6]

제5항에 있어서,
상기 재학습 처리부는 상기 파라미터(m)와 상기 파라미터(n)의 합이 "1"에 근사하도록 상기 양자화 데이터를 재학습하여 상기 학습 데이터와 상기 양자화 데이터 사이의 분포 불일치를 감소시키는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.

[청구항 7]

제5항에 있어서,
상기 재학습 처리부는 상기 손실 함수에서 상기 파라미터(m)의 증가에 따라 유사도에 비중을 증가하고, 상기 파라미터(n)의 증가에 따라 정확도에 비중을 증가하여 상기 양자화 데이터를 재학습하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.

[청구항 8]

제1항에 있어서,
상기 학습 처리부는 상기 인공지능 모델을 RNN(Recurrent Neural Network) 중 하나인 LSTM(Long Short Term Memory)와 어텐션 메커니즘(Attention Mechanism)의 결합 모델로 이용하는 것을 특징으로 하는 양자화 인공지능 학습 처리 장치.

[청구항 9]

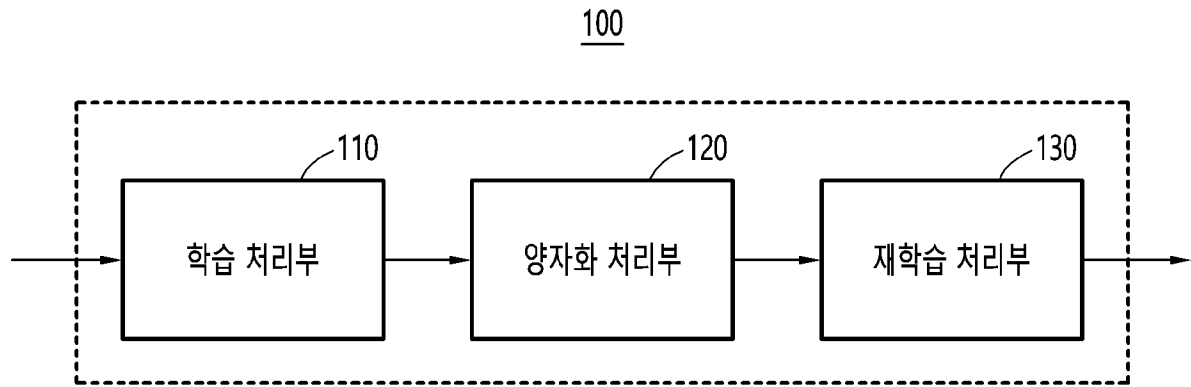
학습 처리부에서, 인공지능 모델을 이용해 입력 데이터를 학습하여 학습 데이터를 출력하는 단계;
양자화 처리부에서, 상기 학습 데이터를 양자화(quantization) 처리하여 양자화 데이터를 출력하는 단계; 및
재학습 처리부에서, 상기 학습 데이터와 상기 양자화 데이터에 대하여 손실 함수에 기반한 정확도 정보와 분포(distribution) 정보 간의 비율을 고려하여 상기 양자화 데이터를 재학습하는 단계를 포함하는 것을 특징으로 하는 양자화 인공지능 학습 처리 방법.

[청구항 10]

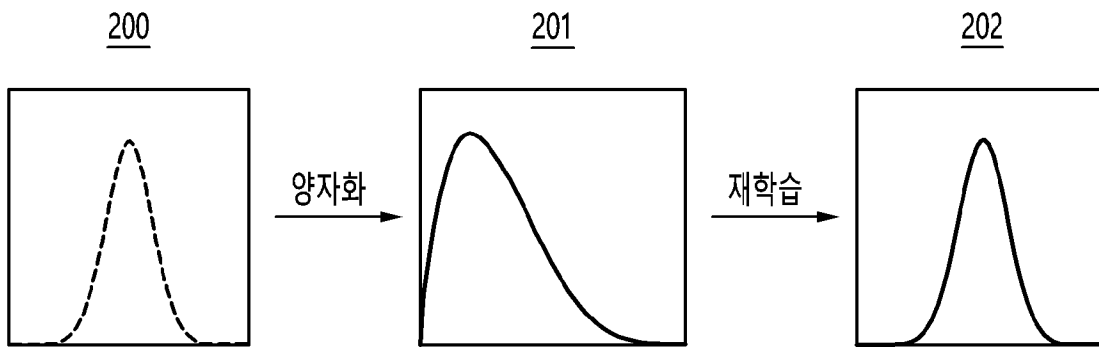
제9항에 있어서,
상기 양자화 데이터를 재학습하는 단계는,
상기 학습 데이터와 상기 양자화 데이터의 왜도(skewness)와 첨도(kurtosis)의 차이 값에 기반하여 상기 양자화 데이터의 왜도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 왜도 값의 차이 값 간의 비율에 기반하여 유사도와 관련된 파라미터(x)를 계산하고, 상기 계산된 파라미터(x)에 기반하여 절대적인 비중에 대한 파라미터(m')을 계산하며, 상기 계산된 파라미터(m')에 기반하여 상대적인 비율에 대한 파라미터(m)을 계산하는 단계; 및

상기 양자화 데이터의 침도 값에 대한 상기 학습 데이터의 왜도 값과 상기 양자화 데이터의 침도 값의 차이 값 간의 비율에 기반하여 정확도와 관련된 파라미터(y)를 계산하고, 상기 계산된 파라미터(y)에 기반하여 상기 절대적인 비중에 대한 파라미터(n')을 계산하며, 상기 계산된 파라미터(n')에 기반하여 상기 상대적인 비율에 대한 파라미터(n)을 계산하는 단계를 포함하는 것을 특징으로 하는 양자화 인공지능 학습 처리 방법.

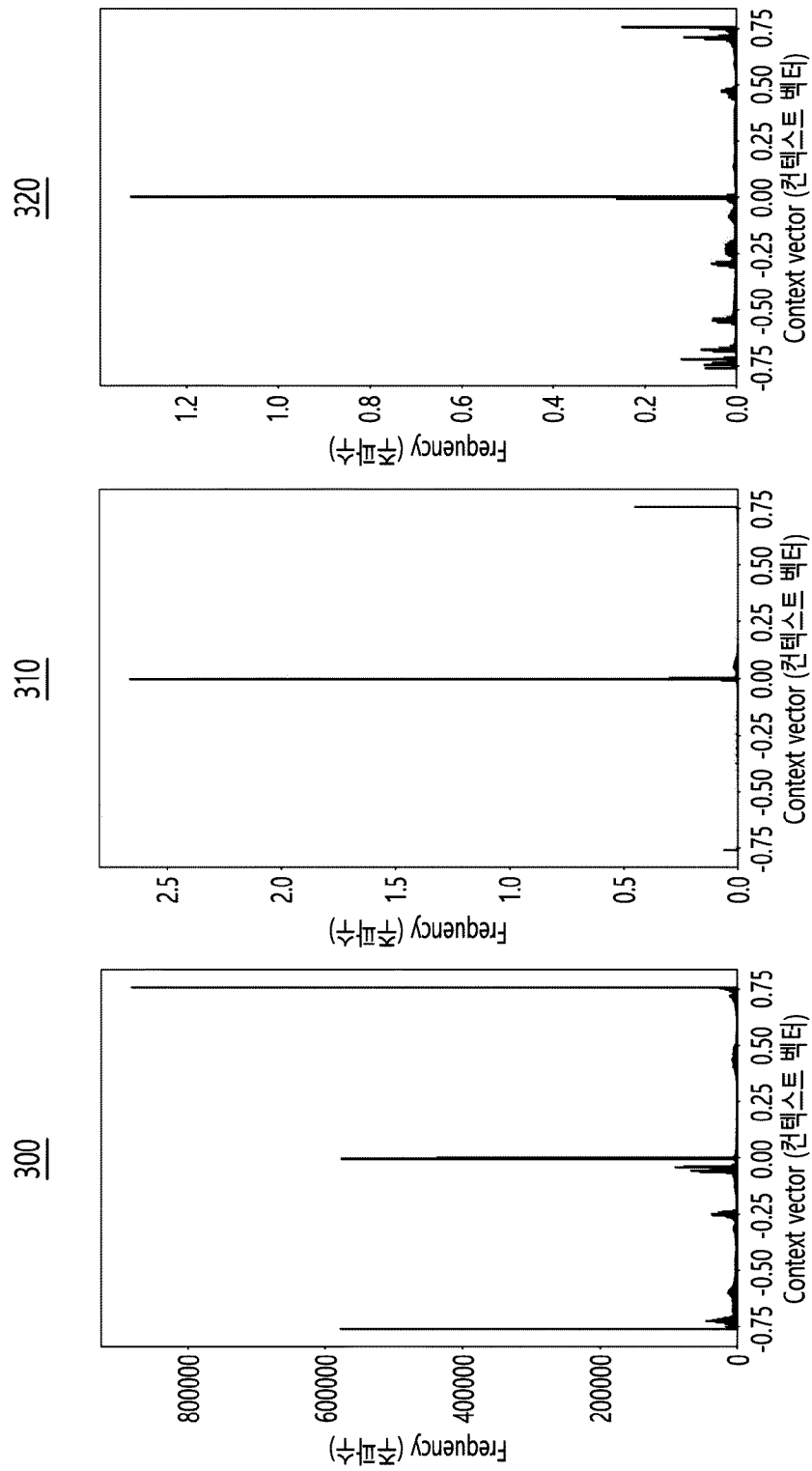
[도1]



[도2]

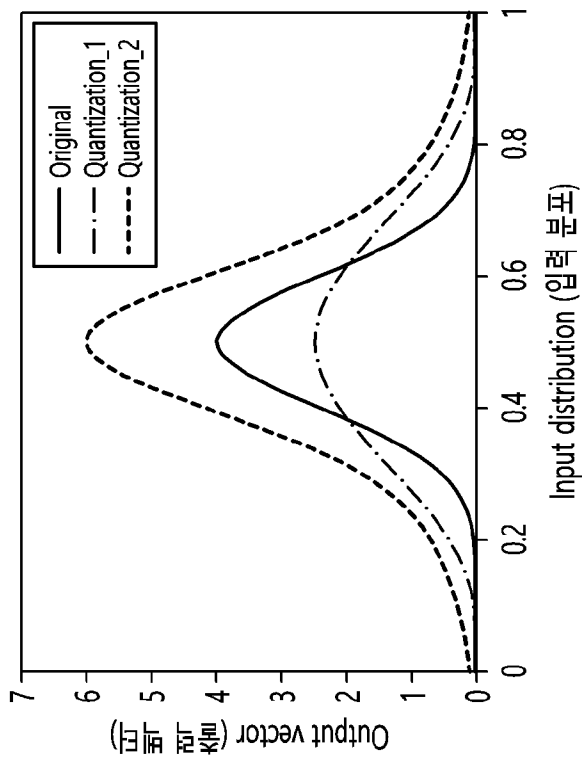


[도3]

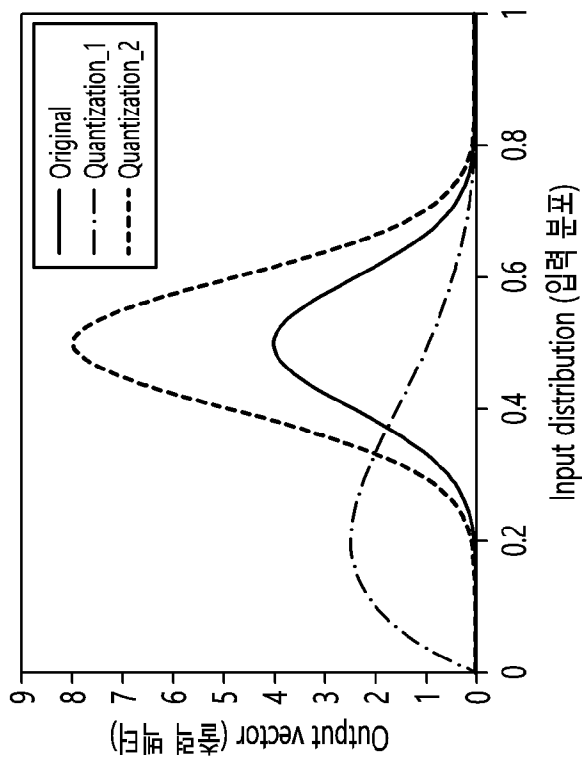


[도4]

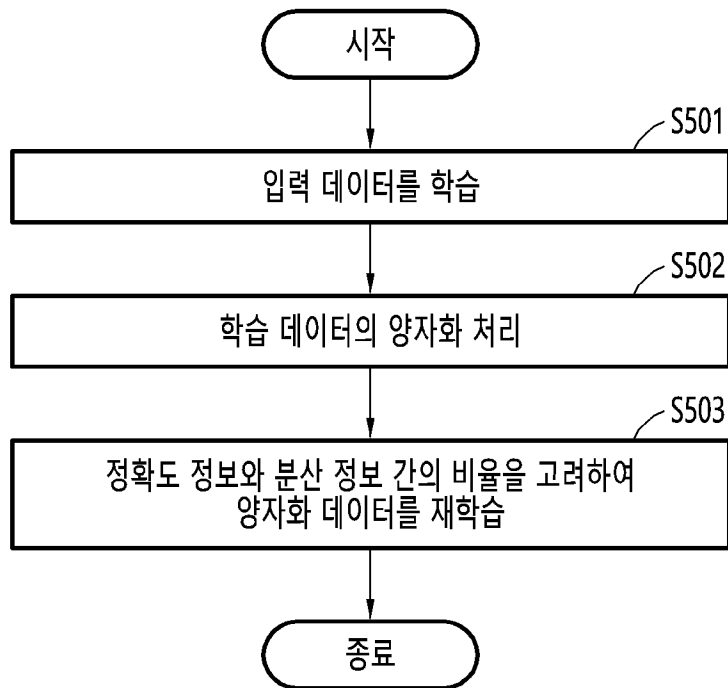
410



400



[도5]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2024/006433

A. CLASSIFICATION OF SUBJECT MATTER G06N 3/0495(2023.01)i; G06N 3/0442(2023.01)i; G06N 3/045(2023.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06N 3/0495(2023.01); G06F 11/34(2006.01); G06F 17/18(2006.01); G06F 18/22(2023.01); G06N 20/00(2019.01); G06N 3/04(2006.01); G06N 3/08(2006.01) Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models: IPC as above Japanese utility models and applications for utility models: IPC as above Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS (KIPO internal) & keywords: 학습 데이터(training data), 손실 함수(loss function), 양자화(quantization), 정확도(accuracy), 분포 정보(distribution information), 왜도(skewness), 첨도(kurtosis)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	CN 112840358 B (BAIDU.COM TIMES TECHNOLOGY (BEIJING) CO., LTD. et al.) 23 May 2023 (2023-05-23) See claim 1.	1,8-9 2-7,10
Y	KR 10-2304900 B1 (KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY) 24 September 2021 (2021-09-24) See paragraphs [0046] and [0062].	1,8-9
Y	CN 117235435 A (4U (BEIJING) TECHNOLOGY CO., LTD.) 15 December 2023 (2023-12-15) See paragraph [0097].	8
A	US 2023-0368077 A1 (INTEL CORPORATION) 16 November 2023 (2023-11-16) See paragraph [0315].	1-10
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 25 September 2024		Date of mailing of the international search report 25 September 2024
Name and mailing address of the ISA/KR Korean Intellectual Property Office Government Complex-Daejeon Building 4, 189 Cheongsaro, Seo-gu, Daejeon 35208 Facsimile No. +82-42-481-8578		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2024/006433

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2022-0109301 A (SAMSUNG ELECTRONICS CO., LTD.) 04 August 2022 (2022-08-04) See claims 1 and 3-4.	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2024/006433

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112840358	B	23 May 2023	CN	112840358	A	25 May 2021
				US	2021-0232890	A1	29 July 2021
				WO	2021-056180	A1	01 April 2021

KR	10-2304900	B1	24 September 2021	None			

CN	117235435	A	15 December 2023	CN	117235435	B	20 February 2024

US	2023-0368077	A1	16 November 2023	None			

KR	10-2022-0109301	A	04 August 2022	CN	112906294	A	04 June 2021
				US	2022-0237513	A1	28 July 2022

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06N 3/0495(2023.01)i; G06N 3/0442(2023.01)i; G06N 3/045(2023.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06N 3/0495(2023.01); G06F 11/34(2006.01); G06F 17/18(2006.01); G06F 18/22(2023.01); G06N 20/00(2019.01);
G06N 3/04(2006.01); G06N 3/08(2006.01)

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: 학습 데이터(training data), 손실 함수(loss function), 양자화(quantization), 정확도(accuracy), 분포 정보(distribution information), 왜도(skewness), 척도(kurtosis)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
Y	CN 112840358 B (BAIDU.COM TIMES TECHNOLOGY (BEIJING) CO., LTD. 등) 2023.05.23 청구항 1	1,8-9
A		2-7,10
Y	KR 10-2304900 B1 (한국과학기술원) 2021.09.24 단락 [0046], [0062]	1,8-9
Y	CN 117235435 A (4U (BEIJING) TECHNOLOGY CO., LTD.) 2023.12.15 단락 [0097]	8
A	US 2023-0368077 A1 (INTEL CORPORATION) 2023.11.16 단락 [0315]	1-10
A	KR 10-2022-0109301 A (삼성전자주식회사) 2022.08.04 청구항 1, 3-4	1-10

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

“A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌

“D” 본 국제출원에서 출원인이 인용한 문헌

“E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후에 공개된 선출원 또는 특허 문헌

“L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌

“O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌

“P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌

“T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌

“X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.

“Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.

“&” 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일

2024년09월25일(25.09.2024)

국제조사보고서 발송일

2024년09월25일(25.09.2024)

ISA/KR의 명칭 및 우편주소

대한민국 특허청
(35208) 대전광역시 서구 청사로 189, 4동 (둔산동,
정부대전청사)

팩스 번호 +82-42-481-8578

심사관

양정록

전화번호 +82-42-481-5709

국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
CN 112840358 B	2023/05/23	CN 112840358 A	2021/05/25
		US 2021-0232890 A1	2021/07/29
		WO 2021-056180 A1	2021/04/01
KR 10-2304900 B1	2021/09/24	없음	
CN 117235435 A	2023/12/15	CN 117235435 B	2024/02/20
US 2023-0368077 A1	2023/11/16	없음	
KR 10-2022-0109301 A	2022/08/04	CN 112906294 A	2021/06/04
		US 2022-0237513 A1	2022/07/28