(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0092335
(43) 공개일자 2025년06월24일

(51) 국제특허분류(Int. Cl.)
G06N 3/0495 (2023.01) *G06N 3/04* (2023.01)
G06N 3/044 (2023.01) *G06N 3/0464* (2023.01)
(52) CPC특허분류
G06N 3/0495 (2023.01)
G06N 3/044 (2023.01)
(21) 출원번호 10-2023-0181695
(22) 출원일자 2023년12월14일
심사청구일자 2023년12월14일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
장익준
경기도 용인시 기흥구 사은로126번길 33, 204동 102호
호 뉴엔 동
경기도 수원시 영통구 매영로415번길 42-22, 101호 (영통동)
(74) 대리인
김연권

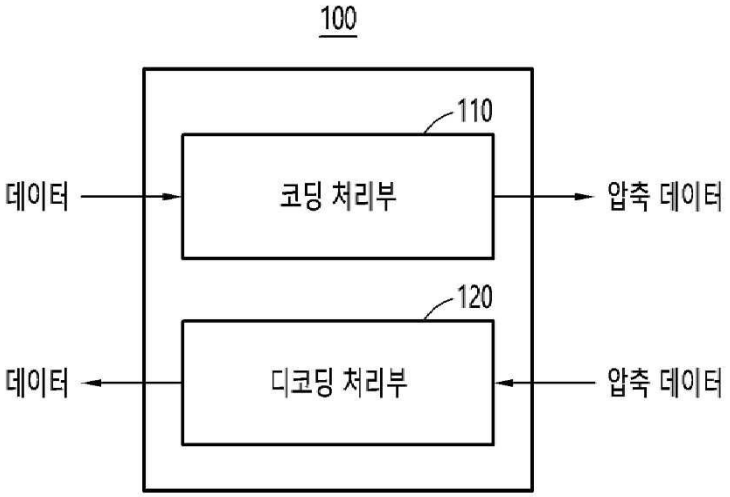
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **인공 신경망의 파라미터 압축 처리 장치 및 방법**

(57) 요 약

본 발명은 인공 신경망의 파라미터가 이상치(outlier)인지 여부를 파라미터의 값에 따라 결정하여 파라미터를 구성하는 MSB(Most Significant Bit)를 제거하거나 LSB(Least Significant Bit)를 잘라내어 파라미터를 압축하는 기술에 관한 것으로, 본 발명의 일실시예에 따른 인공 신경망의 파라미터의 값과 상기 파라미터에서 부호 비트의 존재 여부에 따라 상기 파라미터의 이상치 여부를 판별하고, 상기 파라미터가 비이상치(non-outlier)이면 상기 파라미터의 m에 해당하는 MSB를 제거(remove)하고, 상기 파라미터가 이상치이면 상기 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB를 절단(truncate)하여 파라미터를 압축하는 코딩 처리부 및 상기 압축된 파라미터를 디코딩하는 디코딩 처리부를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/0464 (2023.01)
G06N 3/049 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193316
과제번호 2021-0-00106-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 (혁신법)SW컴퓨팅산업원천기술개발(R&D,정보화)_2023
연구과제명 가속기 최적화 기반 인공신경망 자동생성 기술 및 오픈 서비스 플랫폼 개발(3/4)
기 여 율 1/2
과제수행기관명 서울대학교 산학협력단
연구기간 2021.04.01 ~ 2024.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711193525
과제번호 2021-0-00105-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 (혁신법)SW컴퓨팅산업원천기술개발(R&D,정보화)_2023
연구과제명 엣지 응용 스케일러블 온디바이스 AI 컴퓨팅을 위한 모델 경량화 프레임워크 개발
기 여 율 1/2
과제수행기관명 서울대학교 산학협력단
연구기간 2021.04.01 ~ 2024.12.31

명 세 서

청구범위

청구항 1

인공 신경망의 파라미터의 값과 상기 파라미터에서 부호 비트의 존재 여부에 따라 상기 파라미터의 이상치(outlier) 여부를 판별하고, 상기 파라미터가 비이상치(non-outlier)이면 상기 파라미터의 m에 해당하는 MSB(Most Significant Bit)를 제거(remove)하고, 상기 파라미터가 이상치이면 상기 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB(Least Significant Bit)를 절단(truncate)하여 파라미터를 압축하는 코딩 처리부; 및

상기 압축된 파라미터를 디코딩하는 디코딩 처리부를 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 2

제1항에 있어서,

상기 코딩 처리부는 상기 파라미터를 구성하는 비트의 좌측을 상기 MSB로 설정하고, 우측을 상기 LSB로 설정하며, 상기 파라미터가 이상치이고, 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 l을 상기 좌측에서 상기 부호 비트와 다른 비트가 확인되는 순서에 해당하는 이전의 수로 결정하고, 상기 파라미터에서 부호가 존재하지 않는 경우, 상기 l을 상기 좌측에서 "0"비트와 다른 "1" 비트가 확인되는 순서에 해당하는 이전의 수로 결정하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 3

제1항에 있어서,

상기 코딩 처리부는 상기 파라미터를 구성하는 전체 비트의 수에서 상기 압축된 파라미터의 비트의 수에 해당하는 n을 제외하여 상기 m을 결정하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 4

제3항에 있어서,

상기 코딩 처리부는 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 부호 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 부호 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하고, 상기 파라미터에서 부호 비트가 존재하지 않는 경우 "0" 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하며, 상기 불일치하는 비트가 상기 m에 해당하는 MSB내에서 존재하지 않으면 상기 m에 해당하는 MSB가 제거된 이후부터 상기 압축된 파라미터의 비트로 결정하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 5

제4항에 있어서,

상기 코딩 처리부는 상기 파라미터에서 부호 비트가 존재하는 경우에서 상기 부호 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 부호 비트와 불일치하는 비트 값이 등장하거나 상기 파라미터에서 부호 비트가 존재하지 않는 경우에서 "0" 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하는 비트 값이 등장하면 상기 파라미터를 상기 이상치로 판별하고, 불일치하는 비트가 상기 m에 해당하는 MSB내

에서 존재하지 않으면 상기 파라미터를 상기 비이상치로 판별하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 6

제1항에 있어서,

상기 코딩 처리부는 상기 LSB가 절단되는 수와 관련하여 플래그 비트 값을 설정하고, 상기 설정된 플래그 비트 값을 상기 압축된 파라미터의 일측에 추가하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 7

제1항에 있어서,

상기 코딩 처리부는 상기 파라미터를 포함하는 복수의 파라미터 각각에 대하여 설정되는 플래그 비트 값들에 기반하여 상기 복수의 파라미터를 적어도 하나의 그룹으로 구분하여 압축하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 8

제7항에 있어서,

상기 플래그 비트 값은 상기 인공 신경망의 파라미터와 상기 압축된 파라미터를 구분하는 정보인 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 9

제7항에 있어서,

상기 코딩 처리부는 상기 플래그 비트값들 중 어느 하나의 플래그 비트값에 기반하여 상기 적어도 하나의 그룹에서 하나의 그룹을 구분하여 압축하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 10

제1항에 있어서,

상기 인공 신경망은 DNN(Deep Neural Network), CNN(Convolution neural network), RNN(Recurrent Neural Network) 및 SNN(Spiking Neural Network) 중 적어도 하나를 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 11

제1항에 있어서,

상기 인공 신경망의 파라미터의 값은 8비트로 구성되고, 상기 파라미터를 구성하는 비트의 좌측을 상기 MSB로 설정하고, 우측을 상기 LSB로 설정하며, 상기 부호 비트의 존재 여부에 따라 양수와 음수로 구분하여 "128"의 십진수를 나타내거나 양수로 "255"까지의 십진수를 나타내는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 장치.

청구항 12

코딩 처리부에서, 인공 신경망의 파라미터의 값과 상기 파라미터에서 부호 비트의 존재 여부에 따라 상기 파라미터의 이상치(outlier) 여부를 판별하는 단계;

상기 코딩 처리부에서, 상기 파라미터가 비이상치(non-outlier)이면 상기 파라미터의 m에 해당하는 MSB(Most

Significant Bit)를 제거(remove)하거나 상기 파라미터가 이상치이면 상기 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB(Least Significant Bit)를 절단(truncate)하여 파라미터를 압축하는 단계; 및

디코딩 처리부에서, 상기 압축된 파라미터를 디코딩하는 단계를 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 방법.

청구항 13

제12항에 있어서,

상기 파라미터를 압축하는 단계는,

상기 파라미터를 구성하는 비트의 좌측을 상기 MSB로 설정하고, 우측을 상기 LSB로 설정하며, 상기 파라미터가 이상치이고, 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 l을 상기 좌측에서 상기 부호 비트와 다른 비트가 확인되는 순서에 해당하는 이전의 수로 결정하고, 상기 파라미터에서 부호가 존재하지 않는 경우, 상기 l을 상기 좌측에서 "0"비트와 다른 "1" 비트가 확인되는 순서에 해당하는 이전의 수로 결정하는 단계를 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 방법.

청구항 14

제12항에 있어서,

상기 파라미터를 압축하는 단계는,

상기 파라미터를 구성하는 전체 비트의 수에서 상기 압축된 파라미터의 비트의 수에 해당하는 n을 제외하여 상기 m을 결정하는 단계; 및

상기 파라미터에서 부호 비트가 존재하는 경우, 상기 부호 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 부호 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하고, 상기 파라미터에서 부호 비트가 존재하지 않는 경우 "0" 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하며, 상기 불일치하는 비트가 상기 m에 해당하는 MSB내에서 존재하지 않으면 상기 m에 해당하는 MSB가 제거된 이후부터 상기 압축된 파라미터의 비트로 결정하는 단계를 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 방법.

청구항 15

제12항에 있어서,

상기 코딩 처리부에서, 상기 LSB가 절단되는 수와 관련하여 플래그 비트 값을 설정하고, 상기 설정된 플래그 비트 값을 상기 압축된 파라미터의 일측에 추가하는 단계를 더 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 방법.

청구항 16

제12항에 있어서,

상기 코딩 처리부에서, 상기 파라미터를 포함하는 복수의 파라미터 각각에 대하여 설정되는 플래그 비트 값들에 기반하여 상기 복수의 파라미터를 적어도 하나의 그룹으로 구분하여 압축하는 단계를 더 포함하는 것을 특징으로 하는

인공 신경망의 파라미터 압축 처리 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 신경망의 파라미터 압축 처리 장치 및 방법에 관한 것으로, 보다 구체적으로, 인공 신경망의 파라미터가 이상치(outlier)인지 여부를 파라미터의 값에 따라 결정하여 파라미터를 구성하는 MSB(Most Significant Bit)를 제거하거나 LSB(Least Significant Bit)를 잘라내어 파라미터를 압축하는 기술에 관한 것이다.

배 경 기 술

[0002] DNN(Deep Neural Network) 모델의 크기 증가와 상당한 계산 요구 사항으로 인해 가중치, 활성화 및 중간 결과를 저장하고 검색하기 위하여 상당한 메모리 액세스가 발생한다.

[0003] 생물학적인 신경망과 유사한 인공 신경망(artificial neural network, ANN)에 기반하는 인공지능 기술이 이미지 인지, 음성 인지, 번역, 검색, 딥러닝, 데이터 수집 및 분석, 자율 주행 등과 같은 다양한 분야에서 사용되고 있다.

[0004] 컴퓨팅 장치의 하드웨어가 발전함에 따라, 다수의 은닉층들을 포함하는 DNN이 주로 사용되고 있다.

[0005] 신경망의 학습, 트레이닝, 또는 추론에 있어서, 다수의 파라미터들이 생성되고, 참조되고, 또는 업데이트 될 수 있다.

[0006] 이러한 파라미터들이 증가함에 따라, 파라미터들을 저장하기 위한 메모리 장치의 용량 개수, 집적도 등도 증가하고 있다.

[0007] 따라서, 신경망을 실행하는데 요구되는 메모리 장치의 전력 소모를 줄이는 것이 필요하다.

[0008] DNN 애플리케이션의 주 메모리와 처리 장치 간의 이러한 데이터 이동은 상당한 양의 전력을 소비한다.

[0009] 게다가 프로세싱과 메인 메모리 중 특히, DRAM(Dynamic Random Access Memory) 사이의 성능 격차는 컴퓨팅 분야에서 오랫동안 우려되어 왔다.

[0010] 처리 속도는 수년에 걸쳐 지속적으로 크게 향상되었지만 메모리 속도는 동일한 발전 속도를 따라가지 못했다.

[0011] 이러한 상황으로 인해 프로세서가 종종 DRAM에서 데이터를 가져올 때까지 기다려야 하는 성능 병목(bottleneck) 현상이 발생하여 특히 DNN과 같은 메모리 집약적인 작업의 경우 전체 시스템 성능이 지연되고 방해된다는 문제점이 있다.

[0012] 결과적으로 이러한 문제는 DNN 구현 시 에너지 효율성과 대기 시간 효율성을 방해한다.

[0013] PTQ(Post Training Quantization)는 DNN의 기본 메모리 공간을 줄이는 실용적이고 비용 효율적인 기술이다.

[0014] 그러나 정밀도가 8비트 미만으로 떨어지면 정확도가 눈에 띄게 감소하여 PTQ의 효율성이 제한된다.

[0015] PTQ의 이러한 한계를 극복하기 위해 우리는 정확도 손실을 최소화하면서 8비트 PTQ 모델을 낮은 정밀도로 압축하는 O-2A(Outlier Aware Approximation)라는 새로운 압축 방법을 제시되었다.

[0016] O-2A에서는 매개변수를 이상치와 비이상치로 분류한다.

[0017] 이상치의 중요한 비트는 보존되고, 이상치가 아닌 불필요한 비트는 제거되어 압축 오류가 줄어준다.

[0018] 그럼에도, 불구하고 6비트 정밀도 미만의 O-2A를 사용하여 이상치를 압축하면 상당한 압축 오류가 발생하는 문제점이 존재한다.

[0019] 예를 들어, 6비트 압축에서는 이상치의 LSB(Least Significant Bit)를 제거함으로써 발생하는 오류는 미비하나, 5비트 및 4비트 압축에서는 비트 사용량을 기준으로 파라미터를 비이상치와 이상치로 나눌 때 대상 비트폭(n)이 감소하면 이상치가 증가하는 문제가 존재한다.

[0020] 즉, 6비트 압축보다 압축 중에서 이상치에서 m-LSB가 제거 됨에 따라 이상치가 많을수록 누적 오류가 증가하는 문제점이 존재한다.

[0021] 여기서, m은 압축 대상 비트인 8비트에서 6비트로 압축하기 위한 대상 비트폭(n)을 제외한 2가 된다(m=8-n).

[0022] 더해서, 6비트 압축에서는 각 이상치에서 제거되는 LSB의 수가 m=2인 반면에 5비트는 m=3이며, 4비트는 m=4로 증가하여 압축 오류가 증가하는 문제점이 존재한다.

선행기술문헌

특허문헌

[0023] (특허문헌 0001) 한국등록특허 제10-2399197호, "이상치 인지 근사적 코딩을 수행하는 전자 장치 그리고 그것의 동작 방법"

(특허문헌 0002) 한국공개특허 제10-2023-0143041호, "다중 모드 활성화 함수를 위한 가속기 장치"

(특허문헌 0003) 한국공개특허 제10-2023-0096904호, "희소 행렬 데이터 압축 방법 및 장치"

(특허문헌 0004) 한국공개특허 제10-2019-0128980호, "데이터 압축 및 복원을 위한 전자 장치 및 그 압축 방법"

발명의 내용

해결하려는 과제

[0024] 본 발명은 인공 신경망의 파라미터가 이상치(outlier)인지 여부를 파라미터의 값에 따라 결정하여 파라미터를 구성하는 m-MSB 또는 l-MSB를 제거하거나, m-l-LSB를 잘라내어 파라미터를 압축함에 따라 잘라내는 LSB에 따른 압축 오류를 해소함에 따라 압축 효율성을 향상시키는 것을 목적으로 한다.

[0025] 본 발명은 전체 비트에서 n비트 대상으로 압축하기 위해 이상치와 비이상치로 구분하되, 이상치에 대하여 m-LSB(Least Significant Bit)를 모두 절단하는 것 대신하여 m-MSB(Most Significant Bit)에서 사용되지 않는 m-l 비트와 l개의 LSB를 제거하여 LSB의 절단 길이를 감소시킴에 따라 누적 압축 오류를 해소하면서 압축 효율성을 향상시키는 것을 목적으로 한다.

[0026] 본 발명은 파라미터의 압축에 따라 생성되는 플래그를 공유하여 플래그의 수를 감소시킴에 따라 그룹 파라미터의 압축을 구현하여 개별 플래그 활용에 따른 오버헤드의 발생을 억제하는 것을 목적으로 한다.

과제의 해결 수단

[0027] 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치는 인공 신경망의 파라미터의 값과 상기 파라미터에서 부호 비트의 존재 여부에 따라 상기 파라미터의 이상치(outlier) 여부를 판별하고, 상기 파라미터가 비이상치(non-outlier)이면 상기 파라미터의 m에 해당하는 MSB(Most Significant Bit)를 제거(remove)하고, 상기 파라미터가 이상치이면 상기 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB(Least Significant Bit)를 절단(truncate)하여 파라미터를 압축하는 코딩 처리부 및 기 압축된 파라미터를 디코딩하는 디코딩 처리부를 포함할 수 있다.

[0028] 상기 코딩 처리부는 상기 파라미터를 구성하는 비트의 좌측을 상기 MSB로 설정하고, 우측을 상기 LSB로 설정하며, 상기 파라미터가 이상치이고, 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 l을 상기 좌측에서 상기 부호 비트와 다른 비트가 확인되는 순서에 해당하는 이전의 수로 결정하고, 상기 파라미터에서 부호가 존재하지 않는 경우, 상기 l을 상기 좌측에서 "0"비트와 다른 "1" 비트가 확인되는 순서에 해당하는 이전의 수로 결정할 수 있다.

[0029] 상기 코딩 처리부는 상기 파라미터를 구성하는 전체 비트의 수에서 상기 압축된 파라미터의 비트의 수에 해당하는 n을 제외하여 상기 m을 결정할 수 있다.

[0030] 상기 코딩 처리부는 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 부호 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 부호 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하고, 상기 파라미터에서 부호 비트가 존재하지 않는 경우 "0" 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하며, 상기 불일치하는 비트가 상기 m에 해당하는 MSB내에서 존재하지 않으면 상기 m에 해당하는 MSB가 제거된 이후부터 상기 압축된 파라미터의 비트로 결정할 수 있다.

[0031] 상기 코딩 처리부는 상기 파라미터에서 부호 비트가 존재하는 경우에서 상기 부호 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 부호 비트와 불일치하는 비트 값이 등장하거나 상기 파라미터에서 부호 비트가 존재하지 않는 경우에서 "0" 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하

는 비트 값이 등장하면 상기 파라미터를 상기 이상치로 판별하고, 불일치하는 비트가 상기 m에 해당하는 MSB내에서 존재하지 않으면 상기 파라미터를 상기 비이상치로 판별할 수 있다.

[0032] 상기 코딩 처리부는 상기 LSB가 절단되는 수와 관련하여 플래그 비트 값을 설정하고, 상기 설정된 플래그 비트 값을 상기 압축된 파라미터의 일측에 추가할 수 있다.

[0033] 상기 코딩 처리부는 상기 파라미터를 포함하는 복수의 파라미터 각각에 대하여 설정되는 플래그 비트 값들에 기반하여 상기 복수의 파라미터를 적어도 하나의 그룹으로 구분하여 압축할 수 있다.

[0034] 상기 플래그 비트 값은 상기 인공 신경망의 파라미터와 상기 압축된 파라미터를 구분하는 정보일 수 있다.

[0035] 상기 코딩 처리부는 상기 플래그 비트값들 중 어느 하나의 플래그 비트값에 기반하여 상기 적어도 하나의 그룹에서 하나의 그룹을 구분하여 압축할 수 있다.

[0036] 상기 인공 신경망은 DNN(Deep Neural Network), CNN(Convolution neural network), RNN(Recurrent Neural Network) 및 SNN(Spiking Neural Network) 중 적어도 하나를 포함할 수 있다.

[0037] 상기 인공 신경망의 파라미터의 값은 8비트로 구성되고, 상기 파라미터를 구성하는 비트의 좌측을 상기 MSB로 설정하고, 우측을 상기 LSB로 설정하며, 상기 부호 비트의 존재 여부에 따라 양수와 음수로 구분하여 "128"의 십진수를 나타내거나 양수로 "255"까지의 십진수를 나타낼 수 있다.

[0038] 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 코딩 처리부에서, 인공 신경망의 파라미터의 값과 상기 파라미터에서 부호 비트의 존재 여부에 따라 상기 파라미터의 이상치(outlier) 여부를 판별하는 단계, 코딩 처리부에서, 상기 파라미터가 비이상치(non-outlier)이면 상기 파라미터의 m에 해당하는 MSB(Most Significant Bit)를 제거(remove)하거나 상기 파라미터가 이상치이면 상기 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB(Least Significant Bit)를 절단(truncate)하여 파라미터를 압축하는 단계 및 코딩 처리부에서, 상기 압축된 파라미터를 디코딩하는 단계를 포함할 수 있다.

[0039] 상기 파라미터를 압축하는 단계는, 상기 파라미터를 구성하는 비트의 좌측을 상기 MSB로 설정하고, 우측을 상기 LSB로 설정하며, 상기 파라미터가 이상치이고, 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 l을 상기 좌측에서 상기 부호 비트와 다른 비트가 확인되는 순서에 해당하는 이전의 수로 결정하고, 상기 파라미터에서 부호가 존재하지 않는 경우, 상기 l을 상기 좌측에서 "0"비트와 다른 "1" 비트가 확인되는 순서에 해당하는 이전의 수로 결정하는 단계를 포함할 수 있다.

[0040] 상기 파라미터를 압축하는 단계는, 상기 파라미터를 구성하는 전체 비트의 수에서 상기 압축된 파라미터의 비트의 수에 해당하는 n을 제외하여 상기 m을 결정하는 단계 및 상기 파라미터에서 부호 비트가 존재하는 경우, 상기 부호 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 부호 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하고, 상기 파라미터에서 부호 비트가 존재하지 않는 경우 "0" 비트와 상기 파라미터에서 상기 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하는 비트 값이 등장하는 순서부터 상기 압축된 파라미터의 비트로 결정하며, 상기 불일치하는 비트가 상기 m에 해당하는 MSB내에서 존재하지 않으면 상기 m에 해당하는 MSB가 제거된 이후부터 상기 압축된 파라미터의 비트로 결정하는 단계를 포함할 수 있다.

[0041] 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 상기 코딩 처리부에서, 상기 LSB가 절단되는 수와 관련하여 플래그 비트 값을 설정하고, 상기 설정된 플래그 비트 값을 상기 압축된 파라미터의 일측에 추가하는 단계를 더 포함할 수 있다.

[0042] 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 상기 코딩 처리부에서, 상기 파라미터를 포함하는 복수의 파라미터 각각에 대하여 설정되는 플래그 비트 값들에 기반하여 상기 복수의 파라미터를 적어도 하나의 그룹으로 구분하여 압축하는 단계를 더 포함할 수 있다.

발명의 효과

[0043] 본 발명은 인공 신경망의 파라미터가 이상치(outlier)인지 여부를 파라미터의 값에 따라 결정하여 파라미터를 구성하는 m-MSB 또는 l-MSB를 제거하거나, m-l-LSB를 잘라내어 파라미터를 압축함에 따라 잘라내는 LSB에 따른 압축 오류를 해소함에 따라 압축 효율성을 향상시킬 수 있다.

[0044] 본 발명은 전체 비트에서 n비트 대상으로 압축하기 위해 이상치와 비이상치로 구분하되, 이상치에 대하여 m-LSB(Least Significant Bit)를 모두 절단하는 것 대신하여 m-MSB(Most Significant Bit)에서 사용되지 않는 m-

1 비트와 1개의 LSB를 제거하여 LSB의 절단 길이를 감소시킴에 따라 누적 압축 오류를 해소하면서 압축 효율성을 향상시킬 수 있다.

[0045] 본 발명은 파라미터의 압축에 따라 생성되는 플래그를 공유하여 플래그의 수를 감소시킴에 따라 그룹 파라미터의 압축을 구현하여 개별 플래그 활용에 따른 오버헤드의 발생을 억제할 수 있다.

도면의 간단한 설명

[0046] 도 1은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치를 설명하는 도면이다.

도 2a 및 도 2b는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 구성을 설명하는 도면이다.

도 3은 본 발명의 일실시예에 따른 인공 신경망의 파라미터의 분포들을 예시적으로 설명하는 도면이다.

도 4 내지 도 6은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 파라미터 압축 과정을 설명하는 도면이다.

도 7a 내지 도 7c는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹 압축 코딩을 설명하는 도면이다.

도 8a 내지 도 8c는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹에 적합한 압축 시나리오를 선택하는 프로세스를 설명하는 도면이다.

도 9는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 압축 성능을 설명하는 도면이다.

도 10은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법을 설명하는 도면이다.

도 11은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 프로그램 코드 활용 구조를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0047] 이하, 본 문서의 다양한 실시 예들이 첨부된 도면을 참조하여 기재된다.

[0048] 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.

[0049] 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.

[0050] 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.

[0051] 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.

[0052] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.

[0053] 본 문서에서, "A 또는 B" 또는 "A 및/또는 B 중 적어도 하나" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다.

[0054] "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.

[0055] 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.

[0056] 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.

[0057] 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는"

것을 의미할 수 있다.

[0058] 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.

[0059] 또한, '또는' 이라는 용어는 배타적 논리합 'exclusive or' 이기보다는 포함적인 논리합 'inclusive or' 를 의미한다.

[0060] 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.

[0061] 이하 사용되는 '..부', '..기' 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어, 또는, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0063] 도 1은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치를 설명하는 도면이다.

[0064] 도 1은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 구성 요소를 예시한다.

[0065] 도 1을 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치(100)는 코딩 처리부(110) 및 디코딩 처리부(120)를 포함한다.

[0066] 코딩 처리부(110)는 인공 신경망의 파라미터의 값과 파라미터에서 부호 비트의 존재 여부에 따라 파라미터의 이상치(outlier) 여부를 판별한다.

[0067] 코딩 처리부(110)는 파라미터가 비이상치(non-outlier)이면 파라미터의 m에 해당하는 MSB(Most Significant Bit)를 제거(remove)한다.

[0068] 코딩 처리부(110)는 파라미터가 이상치이면 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB(Least Significant Bit)를 절단(truncate)한다.

[0069] 이에 따라, 코딩 처리부(110)는 파라미터를 압축하고, 한편, 디코딩 처리부(120)는 기 압축된 파라미터를 디코딩한다.

[0070] 코딩 처리부(110)와 디코딩 처리부(120)는 각각 하드웨어 방식 및 소프트웨어 방식 중 적어도 하나의 방식을 이용하여 구현된다.

[0071] 코딩 처리부(110)는 다단계 이상치 인지 근사적 코딩을 수행함에 따라 파라미터를 압축할 수 있다.

[0072] 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치(100)는 외부로부터 제공되는 파라미터를 그대로 메모리 장치에 저장하지 않고 압축된 파라미터를 메모리 장치에 저장할 수 있고, 메모리 장치에 의해 소모되는 전력 또는 에너지를 감소시킬 수 있다.

[0073] 디코딩 처리부(120)는 압축된 파라미터를 디코딩하거나, 압축 해제하거나, 또는 복원할 수 있다.

[0074] 예를 들어, 디코딩 처리부(120)는 디코딩 모듈, 압축 해제 회로, 디코딩 회로로 지칭될 수 있고, 압축된 파라미터와 관련된 데이터를 정상 데이터로 복원하여 출력할 수 있다.

[0075] 코딩 처리부(110)는 파라미터를 구성하는 비트의 좌측을 MSB로 설정하고, 우측을 LSB로 설정한다.

[0076] 코딩 처리부(110)는 파라미터가 이상치이고, 파라미터에서 부호 비트가 존재하는 경우, l을 좌측에서 부호 비트와 다른 비트가 확인되는 순서에 해당하는 이전의 수로 결정한다.

[0077] 코딩 처리부(110)는 파라미터에서 부호가 존재하지 않는 경우, l을 좌측에서 "0"비트와 다른 "1" 비트가 확인되는 순서에 해당하는 이전의 수로 결정할 수 있다.

[0078] 예를 들어, m에 해당하는 MSB는 m-MSB로 표현할 수 있고, l비트를 사용하는 이상치는 레벨 l로 분류할 수 있다.

[0079] 예를 들어, m-MSB의 첫 번째 비트만 필요한 작은 값을 갖는 이상치는 레벨 1로 분류할 수 있다.

[0080] 즉, 코딩 처리부(110)는 m에 해당하는 LSB를 모두 제거하는 대신에 m-MSB에서 사용되지 않는 비트인 l비트와 m-l개의 LSB에 해당하는 두 부분을 제거한다.

[0081] 코딩 처리부(110)는 파라미터를 구성하는 전체 비트의 수에서 압축된 파라미터의 비트의 수에 해당하는 n을 제외하여 m을 결정할 수 있다.

[0082] 예를 들어, 전체 비트가 8비트이고, 압축하고자 하는 비트인 n이 6, 5, 4이면, m은 2, 3, 4가될 수 있다.

[0083] 코딩 처리부(110)는 파라미터에서 부호 비트가 존재하는 경우, 부호 비트와 파라미터에서 m에 해당하는 MSB 내에서 부호 비트와 불일치하는 비트 값이 등장하는 순서부터 압축된 파라미터의 비트로 결정한다.

[0084] 코딩 처리부(110)는 파라미터에서 부호 비트가 존재하지 않는 경우 "0" 비트와 파라미터에서 m에 해당하는 MSB 내에서 "0" 비트와 불일치하는 비트 값이 등장하는 순서부터 압축된 파라미터의 비트로 결정한다.

[0085] 또한, 코딩 처리부(110)는 불일치하는 비트가 m에 해당하는 MSB내에서 존재하지 않으면 m에 해당하는 MSB가 제거된 이후부터 압축된 파라미터의 비트로 결정할 수 있다.

[0086] 코딩 처리부(110)는 파라미터에서 부호 비트가 존재하는 경우에서 부호 비트와 파라미터에서 m에 해당하는 MSB 내에서 부호 비트와 불일치하는 비트 값이 등장하거나 파라미터에서 부호 비트가 존재하지 않는 경우에서 "0" 비트와 파라미터에서 m에 해당하는 MSB내에서 상기 "0" 비트와 불일치하는 비트 값이 등장하면 파라미터를 이상치로 판별한다.

[0087] 코딩 처리부(110)는 불일치하는 비트가 m에 해당하는 MSB내에서 존재하지 않으면 파라미터를 상기 비이상치로 판별할 수 있다.

[0088] 코딩 처리부(110)는 LSB가 절단되는 수와 관련하여 플래그 비트 값을 설정하고, 설정된 플래그 비트 값을 상기 압축된 파라미터의 일측에 추가할 수 있다.

[0089] 코딩 처리부(110)는 파라미터를 포함하는 복수의 파라미터 각각에 대하여 설정되는 플래그 비트 값들에 기반하여 복수의 파라미터를 적어도 하나의 그룹으로 구분하여 압축할 수 있다.

[0090] 예를 들어, 플래그 비트 값은 인공 신경망의 파라미터와 압축된 파라미터를 구분하는 정보일 수 있다.

[0091] 코딩 처리부(110)는 플래그 비트값들 중 어느 하나의 플래그 비트값에 기반하여 적어도 하나의 그룹에서 하나의 그룹을 구분하여 압축할 수 있다.

[0092] 예를 들어, 인공 신경망은 DNN(Deep Neural Network), CNN(Convolution neural network), RNN(Recurrent Neural Network) 및 SNN(Spiking Neural Network) 중 적어도 하나를 포함할 수 있다.

[0093] 따라서, 본 발명은 인공 신경망의 파라미터가 이상치(outlier)인지 여부를 파라미터의 값에 따라 결정하여 파라미터를 구성하는 m-MSB 또는 l-MSB를 제거하거나, m-l-LSB를 잘라내어 파라미터를 압축함에 따라 잘라내는 LSB에 따른 압축 오류를 해소함에 따라 압축 효율성을 향상시킬 수 있다.

[0094] 또한, 본 발명은 전체 비트에서 n비트 대상으로 압축하기 위해 이상치와 비이상치로 구분하되, 이상치에 대하여 m-LSB(Least Significant Bit)를 모두 절단하는 것 대신하여 m-MSB(Most Significant Bit)에서 사용되지 않는 m-l 비트와 l개의 LSB를 제거하여 LSB의 절단 길이를 감소시킴에 따라 누적 압축 오류를 해소하면서 압축 효율성을 향상시킬 수 있다.

[0096] 도 2a 및 도 2b는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 구성을 설명하는 도면이다.

[0097] 도 2a는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 구성과 관련하여 부호가 없는 경우에 8비트 정수 형식에 따른 파라미터를 예시한다.

[0098] 도 2a를 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터(200)는 DNN 시스템에서 활용되는 8비트 정수 형식으로 크기는 MSB와 LSB로 구성되고, MSB는 좌측으로 설정되며, LSB는 우측으로 설정된다.

[0099] 본 발명의 일실시예에 따른 인공 신경망의 파라미터(200)는 부호가 없는 형식(unsigned format)으로, 양수로 "255"까지의 십진수를 나타낼 수 있다.

[0100] 인공 신경망의 파라미터(200)와 관련된 표(201)는 이진수(binary)를 십진수로 변환한 결과를 예시한다.

[0101] 도 2b는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 구성과 관련하여 부호가 있는 경우에 8비트 정수 형식에 따른 파라미터를 예시한다.

[0102] 도 2b를 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터는 DNN 시스템에서 활용되는 8비트 정수 형식으로 크기는 MSB와 LSB로 구성되고, MSB는 좌측으로 설정되며, LSB는 우측으로 설정된다.

[0103] 본 발명의 일실시예에 따른 인공 신경망의 파라미터(210)는 부호가 있는 형식(signed format)으로, 양수로 "127"까지의 십진수를 나타낼 수 있다.

[0104] 인공 신경망의 파라미터(210)와 관련된 표(211)는 이진수(binary)를 십진수로 변환한 결과를 예시한다.

[0105] 본 발명의 일실시예에 따른 인공 신경망의 파라미터(220)는 부호가 있는 형식(signed format)으로, 음수로 "-127"까지의 십진수를 나타낼 수 있다.

[0106] 인공 신경망의 파라미터(220)와 관련된 표(221)는 이진수(binary)를 십진수로 변환한 결과를 예시한다.

[0107] 예를 들어, 인공 신경망의 파라미터의 값은 8비트로 구성되고, 파라미터를 구성하는 비트의 좌측을 MSB로 설정하고, 우측을 LSB로 설정하며, 부호 비트의 존재 여부에 따라 양수와 음수로 구분하여 "128"의 십진수를 나타내거나 양수로 "255"까지의 십진수를 나타낼 수 있다.

[0108] 8비트 정수 형식은 마이크로 컨트롤러, 임베디드 시스템 및 다양한 하위 수준 애플리케이션에서 널리 사용되는 데이터 유형이다.

[0109] 이 형식은 0~255(부호 없음) 또는 -128~127(부호 있음) 범위의 값을 수용할 수 있다.

[0110] 부호 없는 8비트 정수(UINT8)는 간단한 이진 표현을 사용하고, 8비트 시퀀스의 각 비트는 2의 거듭제곱을 나타내며 가장 오른쪽 비트에서 값 1부터 시작하여 각 후속 비트의 값이 두 배가 된다.

[0111] 가장 왼쪽 비트의 값은 128이고, 각 비트에 0 또는 1을 할당하면 0에서 255 사이의 숫자를 나타낼 수 있다.

[0112] 부호 있는 8비트 정수(INT8)는 2의 보수라는 보다 복잡한 표현을 활용하고, 가장 왼쪽 비트는 부호 비트로, 1은 음수를 나타내고 0은 양수를 나타내며, 나머지 7비트는 크기를 나타낼 수 있다.

[0113] 결과적으로 부호 있는 정수는 0부터 127까지의 양수와 -1부터 -128까지의 음수를 나타낼 수 있다.

[0114] 2의 보수 표현에서 음수를 얻기 위해 모든 비트를 반전시킨 다음 결과에 1을 더한다.

[0115] 2의 보수는 부호 없는 정수와 동일한 하드웨어 및 알고리즘을 사용하여 부호 있는 정수에 대한 산술 연산을 가능하게 한다.

[0116] DNN에서 활성화는 일반적으로 양수 값만 생성하는 ReLU 또는 시그모이드와 같은 활성화 함수의 출력을 구현하므로 양수 값을 갖는다.

[0117] 따라서 부호에 1비트를 요구하지 않고 더 높은 범위를 나타낼 수 있는 부호 없는 형식이 사용된다.

[0118] 반대로, DNN 가중치에는 양수 값과 음수 값이 모두 포함될 수 있으므로 부호 있는 형식이 필요하다.

[0119] DNN의 가중치는 뉴런 간의 연결 강도를 나타내며, 뉴런 관계를 효과적으로 포착하려면 양수 값과 음수 값을 모두 나타낼 수 있어야 하고, 결과적으로, 가중치를 표현하기 위해서는 부호 있는 형식이 필수적일 수 있다.

[0121] 도 3은 본 발명의 일실시예에 따른 인공 신경망의 파라미터의 분포들을 예시적으로 설명하는 도면이다.

[0122] 도 3은 본 발명의 일실시예에 따른 인공 신경망의 파라미터의 분포들을 예시한다.

[0123] 도 3을 참고하면, 그래프(300)는 가중치(weight)를 나타내고, 그래프(310)는 활성화(activations)에 대한 경향성을 나타낸다.

[0124] 그래프(300) 및 그래프(310)에 따르면, DNN 모델 영역에서는 대부분의 가중치와 활성화가 0 값 주위에 모이는 경향이 널리 관찰된다.

[0125] 그래프(300)에 따르면 가중치는 비이상치(non-outlier)에 관련되고, 그래프(310)에 따르면 활성화는 이상치와 관련된다.

[0126] 이러한 현상은 작은 파라미터에 대한 정밀도를 유지하면서 큰 파라미터에 대한 양자화 범위를 효율적으로 활용하는 측면에서 과제를 제기한다.

[0127] 이 문제를 해결하기 위해 연구자들은 다양한 기술을 제안했는데, 그 중 하나는 파라미터를 이상치로 분리할 수 있다.

[0128] 이러한 맥락에서 이상치는 절대값이 상당히 큰 소수의 파라미터에 해당하는 반면, 비이상치는 절대값이 작은 대부분의 파라미터로 구성될 수 있다.

[0129] 이 두 그룹을 분리함으로써 연구자는 뚜렷한 양자화 전략을 구현할 수 있으며, 이상치에 대한 양자화 범위를 보다 효율적으로 사용하고 이상치가 아닌 항목에 대해 더 높은 정밀도를 얻을 수 있다.

[0130] 본 발명은 해당 값(치)을 표현하는 데 필요한 비트폭을 기준으로 파라미터를 이상치와 비이상치로 분류하는 분류 기술을 사용할 수 있다.

[0131] 이상치가 아닌 항목은 값이 작기 때문에 낮은 비트폭을 사용하여 정확하게 표현할 수 있으므로 해당 값(치)을 표시하는데 특정 수의 MSB가 사용되지 않는다.

[0132] n의 압축 목표 비트폭을 달성하기 위해 이상치와 비이상치를 나누어 m MSB(m = 8 - n)를 고정한 상태에서 이상치가 n 비트로 처리될 수 있도록 할 수 있다.

[0133] 그래프(300) 및 그래프(310)에서는 n이 "6"임에 대한 이상치와 비이상치 분할을 보여주며, 비이상치는 2-MSB로 고정되어 있고, 이상치는 가중치의 17.65%와 활성화의 0.135%만을 설명할 수 있다.

[0135] 도 4 내지 도 6은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 파라미터 압축 과정을 설명하는 도면이다.

[0136] 도 4는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 파라미터 압축 과정과 관련하여 6비트 압축 과정을 예시한다.

[0137] 도 4를 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치는 6비트 압축 과정(400)을 수행하고, 6비트 압축 과정(400)은 다양한 파라미터 압축 연산과 관련된 실시예를 포함한다.

[0138] 6비트 압축 과정(400)은 양수(positive number)와 음수(negative number)를 나타내는 부호가 존재하는 포맷(signed format)과 부호가 존재하지 않는 포맷(unsigned format)으로 구분된다.

[0139] 제1 실시예(410)는 부호가 "0"이고, m에 해당하는 MSB는 "00"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 부호가 나머지 비트와 내려오면서, 플래그 비트는 "00"이 추가된다.

[0140] 제2 실시예(411)는 부호가 "1"이고, m에 해당하는 MSB는 "11"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 부호가 나머지 비트와 내려오면서, 플래그 비트는 "00"이 추가된다.

[0141] 제3 실시예(420)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "00"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 나머지 비트가 내려오면서, 플래그 비트는 "00"이 추가된다.

[0142] 제1 실시예(410), 제2 실시예(411) 및 제3 실시예(420)는 비이상치에 해당된다.

[0143] 제4 실시예(412)는 부호가 "0"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, 1에 해당하는 MSB(여기서, "0")가 제거되고, m-1에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0144] 제5 실시예(413)는 부호가 "1"이고, m에 해당하는 MSB는 "10"임에 이상치로 판단되어서, 1에 해당하는 MSB(여기서, "1")가 제거되고, m-1에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0145] 제6 실시예(421)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "01"임에 비이상치로 판단되어서, 1에 해당하는 MSB(여기서, "0")가 제거되고, m-1에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0146] 제4 실시예(412), 제5 실시예(413) 및 제6 실시예(421)는 제1 레벨의 이상치에 해당된다.

[0147] 제7 실시예(414)는 부호가 "0"이고, m에 해당하는 MSB는 "10"임에 이상치로 판단되어서, 1에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와

내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "10"이 추가된다.

[0148] 제8 실시예(415)는 부호가 "1"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "10"이 추가된다.

[0149] 제9 실시예(422)는 부호가 없음에 따라 첫 비트가 "1"이고, m에 해당하는 MSB는 "10"임에 비이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "10"이 추가된다.

[0150] 제7 실시예(415), 제8 실시예(416) 및 제9 실시예(422)는 제2 레벨의 이상치에 해당된다.

[0151] 도 5는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 파라미터 압축 과정과 관련하여 5비트 압축 과정을 예시한다.

[0152] 도 5를 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치는 5비트 압축 과정(500)을 수행하고, 5비트 압축 과정(500)은 다양한 파라미터 압축 연산과 관련된 실시예를 포함한다.

[0153] 5비트 압축 과정(500)은 양수(positive number)와 음수(negative number)를 나타내는 부호가 존재하는 포맷(signed format)과 부호가 존재하지 않는 포맷(unsigned format)으로 구분된다.

[0154] 제1 실시예(510)는 부호가 "0"이고, m에 해당하는 MSB는 "000"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 부호가 나머지 비트와 내려오면서, 플래그 비트는 "00"이 추가된다.

[0155] 제2 실시예(511)는 부호가 "1"이고, m에 해당하는 MSB는 "111"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 부호가 나머지 비트와 내려오면서, 플래그 비트는 "00"이 추가된다.

[0156] 제3 실시예(520)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "000"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 나머지 비트가 내려오면서, 플래그 비트는 "00"이 추가된다.

[0157] 제1 실시예(510), 제2 실시예(511) 및 제3 실시예(520)는 비이상치에 해당된다.

[0158] 제4 실시예(512)는 부호가 "0"이고, m에 해당하는 MSB는 "001"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "00")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0159] 제5 실시예(513)는 부호가 "1"이고, m에 해당하는 MSB는 "110"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "11")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0160] 제6 실시예(521)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "001"임에 비이상치로 판단되어서, l에 해당하는 MSB(여기서, "00")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0161] 제4 실시예(512), 제5 실시예(513) 및 제6 실시예(521)는 제1 레벨의 이상치에 해당된다.

[0162] 제7 실시예(514)는 부호가 "0"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "0")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0163] 제8 실시예(515)는 부호가 "1"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "1")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0164] 제9 실시예(522)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "01"임에 비이상치로 판단되어서, l에 해당하는 MSB(여기서, "0")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0165] 제7 실시예(515), 제8 실시예(516) 및 제9 실시예(522)는 제2 레벨의 이상치에 해당된다.

[0166] 제10 실시예(516)는 부호가 "0"이고, m에 해당하는 MSB는 "10"임에 이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오

고, 절단된 LSB의 수와 관련하여 플래그 비트는 "11"이 추가된다.

[0167] 제11 실시예(517)는 부호가 "1"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "11"이 추가된다.

[0168] 제12 실시예(523)는 부호가 없음에 따라 첫 비트가 "1"이고, m에 해당하는 MSB는 "10"임에 비이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "11"이 추가된다.

[0169] 제10 실시예(516), 제11 실시예(517) 및 제12 실시예(523)는 제3 레벨의 이상치에 해당된다.

[0170] 도 6은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 파라미터 압축 과정과 관련하여 4비트 압축 과정을 예시한다.

[0171] 도 6을 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치는 4비트 압축 과정(600)을 수행하고, 4비트 압축 과정(600)은 다양한 파라미터 압축 연산과 관련된 실시예를 포함한다.

[0172] 4비트 압축 과정(600)은 양수(positive number)와 음수(negative number)를 나타내는 부호가 존재하는 포맷(signed format)과 부호가 존재하지 않는 포맷(unsigned format)으로 구분된다.

[0173] 제1 실시예(610)는 부호가 "0"이고, m에 해당하는 MSB는 "0000"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 부호가 나머지 비트와 내려오면서, 플래그 비트는 "000"이 추가된다.

[0174] 제2 실시예(611)는 부호가 "1"이고, m에 해당하는 MSB는 "1111"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 부호가 나머지 비트와 내려오면서, 플래그 비트는 "000"이 추가된다.

[0175] 제3 실시예(620)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "0000"임에 비이상치로 판단되어서, m에 해당하는 MSB가 제거되고, 나머지 비트가 내려오면서, 플래그 비트는 "000"이 추가된다.

[0176] 제1 실시예(610), 제2 실시예(611) 및 제3 실시예(620)는 비이상치에 해당된다.

[0177] 제4 실시예(612)는 부호가 "0"이고, m에 해당하는 MSB는 "0001"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "000")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "001"이 추가된다.

[0178] 제5 실시예(613)는 부호가 "1"이고, m에 해당하는 MSB는 "1110"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "111")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "01"이 추가된다.

[0179] 제6 실시예(621)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "001"임에 비이상치로 판단되어서, l에 해당하는 MSB(여기서, "000")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "001"이 추가된다.

[0180] 제4 실시예(612), 제5 실시예(613) 및 제6 실시예(621)는 제1 레벨의 이상치에 해당된다.

[0181] 제7 실시예(614)는 부호가 "0"이고, m에 해당하는 MSB는 "001"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "00")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "010"이 추가된다.

[0182] 제8 실시예(615)는 부호가 "1"이고, m에 해당하는 MSB는 "110"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "11")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "010"이 추가된다.

[0183] 제9 실시예(622)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "01"임에 비이상치로 판단되어서, l에 해당하는 MSB(여기서, "00")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "010"이 추가된다.

[0184] 제7 실시예(615), 제8 실시예(616) 및 제9 실시예(622)는 제2 레벨의 이상치에 해당된다.

[0185] 제10 실시예(617)는 부호가 "0"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "0")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와

관련하여 플래그 비트는 "011"이 추가된다.

[0186] 제11 실시예(618)는 부호가 "1"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB(여기서, "1")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "011"이 추가된다.

[0187] 제12 실시예(623)는 부호가 없음에 따라 첫 비트가 "0"이고, m에 해당하는 MSB는 "01"임에 비이상치로 판단되어서, l에 해당하는 MSB(여기서, "0")가 제거되고, m-l에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "011"이 추가된다.

[0188] 제10 실시예(617), 제11 실시예(618) 및 제12 실시예(623)는 제3 레벨의 이상치에 해당된다.

[0189] 제13 실시예(618)는 부호가 "0"이고, m에 해당하는 MSB는 "10"임에 이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "100"이 추가된다.

[0190] 제14 실시예(619)는 부호가 "1"이고, m에 해당하는 MSB는 "01"임에 이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "100"이 추가된다.

[0191] 제15 실시예(624)는 부호가 없음에 따라 첫 비트가 "1"이고, m에 해당하는 MSB는 "10"임에 비이상치로 판단되어서, l에 해당하는 MSB가 존재하지 않음에 제거되는 MSB가 존재하지 않고, m에 해당하는 LSB가 절단되면서, 부호가 나머지 비트와 내려오고, 절단된 LSB의 수와 관련하여 플래그 비트는 "100"이 추가된다.

[0192] 제13 실시예(618), 제14 실시예(619) 및 제15 실시예(624)는 제4 레벨의 이상치에 해당된다.

[0193] 도 4 내지 도 6의 6비트 압축 과정(400), 5비트 압축 과정(500) 및 4비트 압축 과정(600)에 따르면 파라미터를 구성하는 부호 비트의 존재 여부에 따라 부호 비트와 m-MSB를 구성하는 비트 간에 비교에 기반하여 이상치 및 비이상치를 판별하고, 비이상치이면 m-MSB를 제거하고, 이상치이면 m-MSB를 구성하는 비트 중에서 제거 가능한 비트에 해당하는 l-MSB를 제거하고, m에서 l의 수를 제외한 수만큼 LSB를 절단하여 파라미터를 압축함에 따라 압축 에러의 발생을 억제함을 확인할 수 있다.

[0194] m에서 l의 수를 제외한 수만큼 LSB를 제거함에 따라 절단되는 LSB의 수가 감소함에 따른 압축 효율성 및 압축 에러가 감소하는 것을 표 1을 통해서 확인할 수 있다.

[0195] 정확도와 관련하여 모든 파라미터에서 LSB를 완전히 자르면 압축 오류로 인해 정확도가 심각하게 저하되는 문제와 관련하여 더 짧은 길이의 LSB를 제거하더라도 정확도가 크게 향상됨을 확인할 수 있다.

[0196] 따라서, 본 발명은 5비트 압축과 4비트 압축에서도 합리적인 정확도를 제공함을 확인할 수 있다.

[0198] 도 7a 내지 도 7c는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹 압축 코딩을 설명하는 도면이다.

[0199] 도 7a는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹 압축 코딩과 관련하여 6비트 압축의 그룹 코딩에 대하여 개별 코딩(700)을 예시한다.

[0200] 도 7b는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹 압축 코딩과 관련하여 6비트 압축의 그룹 코딩에 대하여 2 그룹 코딩(710)을 예시한다.

[0201] 도 7c는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹 압축 코딩과 관련하여 6비트 압축의 그룹 코딩에 대하여 4 그룹 코딩(720)을 예시한다.

[0202] 개별 코딩(700)은 압축된 각 파라미터에 대하여 플래그 비트를 표시하여 오버 헤드가 발생될 수 있다.

[0203] 2 그룹 코딩(710)은 두 개의 파라미터가 모두 플래그를 공유하므로 플래그 수가 2개로 줄어들어서 오버헤드가 줄어든다.

[0204] 마찬 가지로, 4 그룹 코딩(720)은 플래그가 4개의 파라미터 간에 공유되므로 오버헤드가 더 줄어든다.

[0205] 각 파라미터에 대하여 개별 압축 시나리오를 적용하는 것은 어려움에 따라 전체 그룹에 대한 압축 시나리오를

결정하기 위한 우선순위 전략이 필요하다.

[0206] g 값이 증가할수록 이상치가 갖는 영향력도 커지고, 이는 그룹 g의 구성원 수가 증가하면 압축 오류가 증가할 가능성도 존재함을 나타낸다.

[0207] 즉, g 값이 감소하도록 그룹의 수가 적게 설정되어서 사용되는 플래그 수가 감소하면 오버헤드가 감소한다.

[0208] 일반적으로는 파라미터의 작은 부분에 불과한 MSB의 이상치를 보호하기 위해 압축 우선순위를 지정하고, 이상치 압축 시나리오를 이용하여 그룹의 모든 파라미터가 압축될 수 있다.

[0209] 따라서, 본 발명은 파라미터의 압축에 따라 생성되는 플래그를 공유하여 플래그의 수를 감소시킴에 따라 그룹 파라미터의 압축을 구현하여 개별 플래그 활용에 따른 오버헤드의 발생을 억제할 수 있다.

[0211] 도 8a 내지 도 8c는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 그룹에 적합한 압축 시나리오를 선택하는 프로세스를 설명하는 도면이다.

[0212] 도 8a를 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치는 인공 신경망의 파라미터와 관련하여 6비트 압축(800)과 5비트 압축(801)에 대한 적합한 압축 시나리오를 선택하기 위한 프로세스의 초기 과정을 예시한다.

[0213] 처음에는 모든 파라미터의 m-MSB가 OR 게이트(gate)를 사용하여 결합되어 있는 것을 확인할 수 있다.

[0214] 도 8b를 참고하면, 6비트 압축과 관련하여 압축 시나리오를 선택하기 위한 프로세스(810)를 나타낸다.

[0215] 도 8c를 참고하면, 5비트 압축과 관련하여 압축 시나리오를 선택하기 위한 프로세스(820)를 나타낸다.

[0216] 다시 말해, 프로세스(810) 및 프로세스(820)는 그룹에 적합한 압축 시나리오를 선택하는 프로세스를 보여준다.

[0217] 프로세스(810)는 파라미터를 구성하는 MSB에서 이상치 판단에 사용되는 F1과 F2에서 비이상치, 제1 레벨의 이상치 및 제2 레벨의 이상치에 따른 값과 각 값에 따른 오류 압축 시나리오의 요약을 예시한다.

[0218] 프로세스(810)는 도 4에서 설명된 6비트 압축과 관련하여 압축 시나리오와 관련될 수 있다.

[0219] 프로세스(820)는 파라미터를 구성하는 MSB에서 이상치 판단에 사용되는 F1 내지 F3에서 비이상치, 제1 레벨의 이상치, 제2 레벨의 이상치 및 제3 레벨의 이상치에 따른 값과 각 값에 따른 오류 압축 시나리오의 요약을 예시한다.

[0220] m 비트는 O-2A 또는 mO-2A에서와 동일한 방법을 사용하여 그룹에 대한 적절한 압축 시나리오를 결정하는데 사용된다.

[0221] 프로세스(820)는 도 5에서 설명된 5비트 압축과 관련하여 압축 시나리오와 관련될 수 있다.

[0222] 선택한 압축 시나리오는 이후 그룹의 모든 파라미터에 적용되고, mO-2A의 각 시나리오에 대한 평균 압축 오류는 이상치가 높을수록 압축 오류가 더 커진다는 점을 강조한다.

[0223] 그룹 코딩을 사용한다는 것은 이상치가 동일한 그룹에 있는 주변 파라미터의 양자화 오류에 영향을 미칠 수 있음을 나타낸다.

[0225] 도 9는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 압축 성능을 설명하는 도면이다.

[0226] 도 9는 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 압축 성능에 대한 시뮬레이션 결과를 예시한다.

[0227] 도 9를 참고하면, 그래프(900) 내지 그래프(940)는 다양한 모델에 대하여 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 파라미터 압축 방식을 적용하는 경우에 성능을 나타낸다.

[0228] 그래프(900)는 VGG16 모델에 해당되고, 그래프(910)는 ResNet18 모델에 해당되며, 그래프(920)는 ResNet50 모델에 해당되고, 그래프(930)는 MobileNet_v1에 해당되며, 그래프(940)는 MobileNet_V2에 해당된다.

[0229] VGG16, ResNet18, ResNet50과 같은 모델에서와 같이 가중치와 활성화가 모두 압축되면 모델 크기가 20%에서 50%

로 줄어드는 것을 확인할 수 있다.

[0230] 활성화만 압축되는 MobileNet_v1과 MobileNet_v2의 경우와 관련하여, MobileNetV1의 경우 모델 크기가 12~30% 이고, MobileNetV2의 경우 15~35% 감소하는 것을 확인할 수 있다.

[0231] 종래 기술에 해당하는 이상치 인지 근사적(Outlier Aware Approximation, O-2A) 코딩과 본 발명에 해당되는 다단계 수준의 이상치 인지 근사적(Multiple Level Outlier Aware Approximation, mO-2A) 코딩에 대하여 다양한 압축 기술에 대한 네트워크 정확도와 함께 파라미터의 유효 비트폭(effective bitwidth)를 하기 표 1과 같이 정리할 수 있다.

[0232] 표 1에 따르면, 전체 LSB 잘림의 경우 모델 압축과 관련된 비용이 없으므로 파라미터 비트폭은 절대 n 비트일 수 있다.

[0233] 그러나, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법에서는 압축된 파라미터의 n 비트 외에도 g 파라미터가 공유하는 플래그 비트가 있다.

[0234] 이러한 플래그 비트는 g를 증가시켜 크게 줄일 수 있는데, g가 16일 때 플래그 비트는 일반적으로 약 0.2비트임을 확인 할 수 있다.

[0235] [표 1]

bitwidth	Truncation LSBs	O2A				Multiple Levels O-2A			
		Group 4	Group 8	Group 12	Group 16	Group 4	Group 8	Group 12	Group 16
Effective bitwidth									
6-bit	6	6.25	6.13	6.08	6.06	6.50	6.25	6.17	6.13
5-bit	5	5.25	5.13	5.08	5.06	5.50	5.25	5.17	5.13
4-bit	4	4.25	4.13	4.08	4.06	4.75	4.38	4.25	4.19
Accuracy of VGG16								FP32: 73.35, Int8: 73.36	
6-bit	**72.64**	**73.22**	**73.16**	**73.20**	**73.06**	**73.36**	**73.29**	**73.27**	**73.20**
5-bit	70.60	**72.87**	**72.85**	**72.75**	**72.73**	**73.28**	**73.11**	**73.13**	**73.04**
4-bit	37.93	63.21	57.73	55.72	52.19	**72.62**	72.30	72.13	72.38
Accuracy of ResNet18								FP32: 69.76, Int8: 69.62	
6-bit	68.75	**69.45**	**69.51**	**69.39**	**69.42**	**69.54**	**69.56**	**69.48**	**69.51**
5-bit	65.25	68.26	67.65	67.61	67.23	**69.30**	**69.28**	**69.11**	**69.13**
4-bit	34.68	49.22	40.66	37.52	37.91	68.25	67.64	67.55	67.08
Accuracy of ResNet34								FP32: 73.30, Int8: 73.25	
6-bit	72.42	**73.04**	**73.03**	**72.97**	**72.98**	**73.14**	**73.13**	**73.07**	**73.07**
5-bit	69.71	72.24	71.74	71.45	71.44	**73.01**	**72.71**	**72.58**	**72.83**
4-bit	55.67	63.16	60.68	59.31	58.54	72.16	71.93	71.62	71.41
Accuracy of ResNet50								FP32: 76.15, Int8: 76.12	
6-bit	74.94	**76.02**	**75.98**	**76.02**	**76.03**	**76.03**	**76.06**	**75.95**	**75.99**
5-bit	70.94	74.80	74.62	74.26	74.18	**75.80**	**75.77**	**75.71**	**75.59**
4-bit	49.45	61.28	56.31	53.89	52.35	74.79	74.48	74.12	73.88
Accuracy of ResNet101								FP32: 77.37, Int8: 77.27	
6-bit	75.52	**77.17**	**77.12**	**77.08**	**77.05**	**77.18**	**77.24**	**77.19**	**77.12**
5-bit	70.10	76.02	75.54	75.53	75.39	**77.08**	**76.92**	**76.60**	**76.75**
4-bit	24.88	62.72	55.71	50.44	47.07	76.41	75.83	75.65	74.89
Accuracy of MobileNet_v1 (only activation are compressed)								FP32: 70.66, Int8: 70.46	
6-bit	64.68	**70.45**	**70.35**	**70.42**	**70.25**	**70.41**	**70.47**	**70.46**	**70.44**
5-bit	41.45	67.23	66.02	66.46	65.45	**70.18**	**70.02**	**69.96**	**70.01**
4-bit	1.92	15.51	6.64	9.06	5.33	69.16	68.91	68.61	68.47
Accuracy of MobileNet_v2 (only activation are compressed)								FP32: 71.87, Int8: 71.22	
6-bit	61.54	70.94	70.90	70.79	70.82	**71.17**	**71.28**	**71.15**	70.98
5-bit	24.84	64.23	63.15	62.32	61.71	70.41	70.33	70.21	70.19
4-bit	0.46	8.10	10.37	8.10	6.40	68.36	67.86	67.11	66.95

[0236]

[0237] 이는 파라미터 비트에 비해 중요하지 않으며 전체 비트의 약 2%만 차지할 수 있음을 보여준다.

[0239] 도 10은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법을 설명하는 도면이다.

[0240] 도 10은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법이 보다 공격적인 압축을 달성하기 위해 이상치를 여러 수준으로 나누고 압축 오류를 최소화하는 다단계 수준의 이상치 인지 근사적(Multiple Level Outlier Aware Approximation, mO-2A) 코딩을 수행하는 절차를 예시한다.

[0241] 도 10을 참고하면, 단계(S1001)에서 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 파라미터의 값에 대한 이상치 여부를 판별한다.

[0242] 즉, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 인공 신경망의 파라미터의 값과 파라미터에서 부호 비트의 존재 여부에 따라 파라미터의 이상치(outlier) 여부를 판별할 수 있다.

[0243] 단계(S1001)에서 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 이상치로 판별하면 단계

(S1003)으로 진행하고, 비이상치로 판별하면 단계(S1002)로 진행한다.

[0244] 단계(S1002)에서 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 파라미터의 m에 해당하는 MSB를 제거한다.

[0245] 즉, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 파라미터가 비이상치(non-outlier)로 판별하면, 파라미터의 m에 해당하는 MSB를 제거한다.

[0246] 단계(S1003)에서 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB를 절단한다.

[0247] 즉, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 방법은 파라미터가 이상치(outlier)이면 파라미터의 l에 해당하는 MSB를 제거하면서 m에서 l을 제외한 수에 해당하는 LSB를 절단하여 파라미터를 압축할 수 있다.

[0248] 여기서, l은 파라미터에서 부호 비트가 존재하는 경우, MSB 측에서 부호 비트와 다른 비트가 확인되는 순서에 해당하는 이전의 수로 결정된다.

[0249] 한편, l은 파라미터에서 부호 비트가 존재하지 않는 경우, MSB 측에서 "0"비트와 다른 "1" 비트가 확인되는 순서에 해당하는 이전의 수로 결정될 수 있다.

[0251] 도 11은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 프로그램 코드 활용 구조를 설명하는 도면이다.

[0252] 도 11은 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치의 프로그램 코드 활용 구조를 예시한다.

[0253] 도 11을 참고하면, 본 발명의 일실시예에 따른 인공 신경망의 파라미터 압축 처리 장치(1100)는 프로세서(1110) 및 메모리(1120)로 구성되고, 메모리(1120)내에는 코딩 프로그램 코드가 포함된다.

[0254] 코딩 프로그램 코드는 도 1에서 설명된 인공 신경망의 파라미터 압축 처리 장치(100)의 동작과 관련하여 도 10에서 설명된 인공 신경망의 파라미터 압축 처리 방법을 지원한다.

[0255] 프로세서(1110)는 하드웨어 장치로서 메모리(1120)에 저장된 코딩 프로그램 코드를 실행할 수 있다.

[0256] 예를 들어, 프로세서(1110)는 CPU(Central Processing Unit), ISP(Image Signal Processing Unit), DSP(Digital Signal Processing Unit), GPU(Graphics Processing Unit), TPU(Tensor Processing Unit), VPU(Vision Processing Unit), 및 NPU(Neural Processing Unit) 등의 동종 또는 이종의 다수의 코어들 그리고 다수의 코어들에 의해 공유되는 캐시 메모리 등을 포함할 수 있다.

[0257] 프로세서(1110)는 코딩 프로그램 코드의 명령어들에 대한 페치(fetch), 실행, 데이터 요청, 데이터 저장 등과 같은 프로세싱 연산들, 산술 연산들 등을 실행할 수 있다.

[0258] 메모리(1120)는 프로세서(1110)에 의해 실행 가능한 코딩 프로그램 코드를 저장하는 비일시적 컴퓨터 판독가능 매체, 즉 하드웨어 장치일 수 있다.

[0259] 예를 들어, 메모리(1120)에 저장되고 프로세서(1110)에 의해 실행되는 코딩 프로그램 코드는 도 1에서 전술한 코딩 처리부(110) 및 디코딩 처리부(120)에 의해 수행되는 다단계 이상치 인지 근사적 코딩(multiple levels outlier aware approximation, mO-2A)에 관한 명령어들을 포함할 수 있다.

[0260] 예를 들어, 프로세서(1110)는 메모리(1120)에 로드된 코딩 프로그램 코드를 실행함으로써 파라미터를 압축하거나 압축된 파라미터를 압축 해제할 수 있고 도 10의 단계(S1001) 내지 단계(S1004)를 실행할 수 있다.

[0261] 다른 예를 들어, 메모리(1120)는 프로세서(1110)에 의해 실행 가능한 인공 신경망 프로그램 코드를 더 저장할 수도 있다.

[0262] 프로세서(1110)는 메모리(11120)에 로드된 인공 신경망 프로그램 코드를 실행함으로써 인공 신경망을 실행할 수 있다.

[0263] 예를 들어, 인공 신경망은 DNN(Deep Neural Network), CNN(Convolution neural network),

RNN(RecurrentNeural Network), SNN(Spiking Neural Network) 등을 포함할 수 있으나 이에 한정되지 않는다.

[0265] 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.

[0266] 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.

[0267] 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.

[0268] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

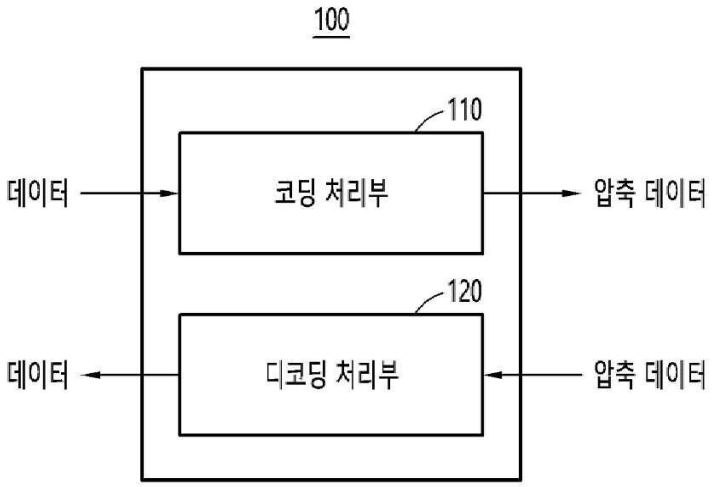
부호의 설명

[0269] 100: 인공 신경망의 파라미터 압축 처리 장치

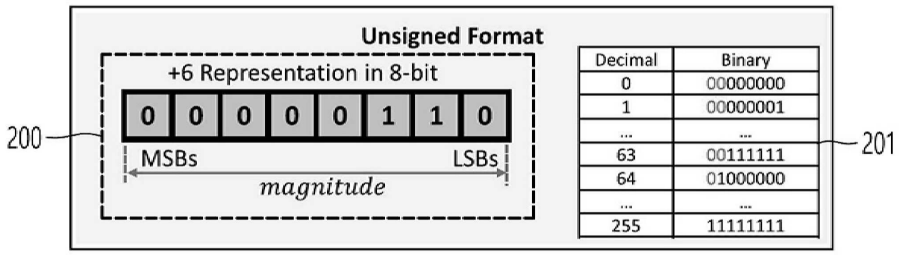
110: 코딩 처리부 120: 디코딩 처리부

도면

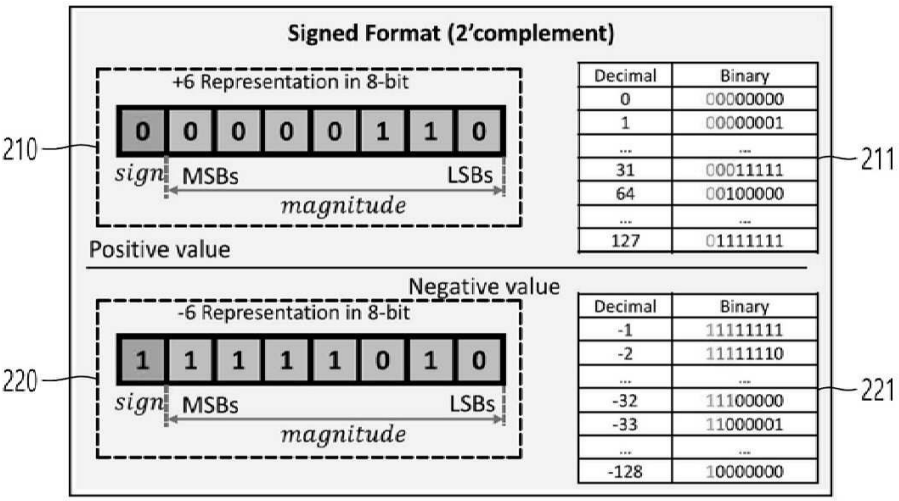
도면1



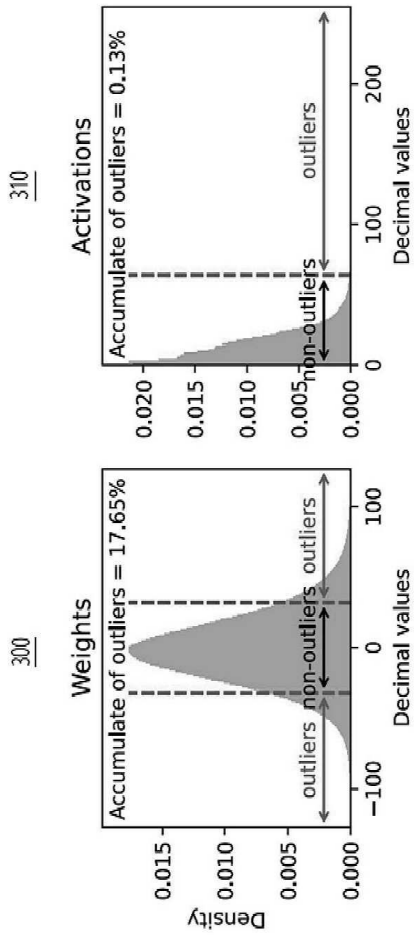
도면2a



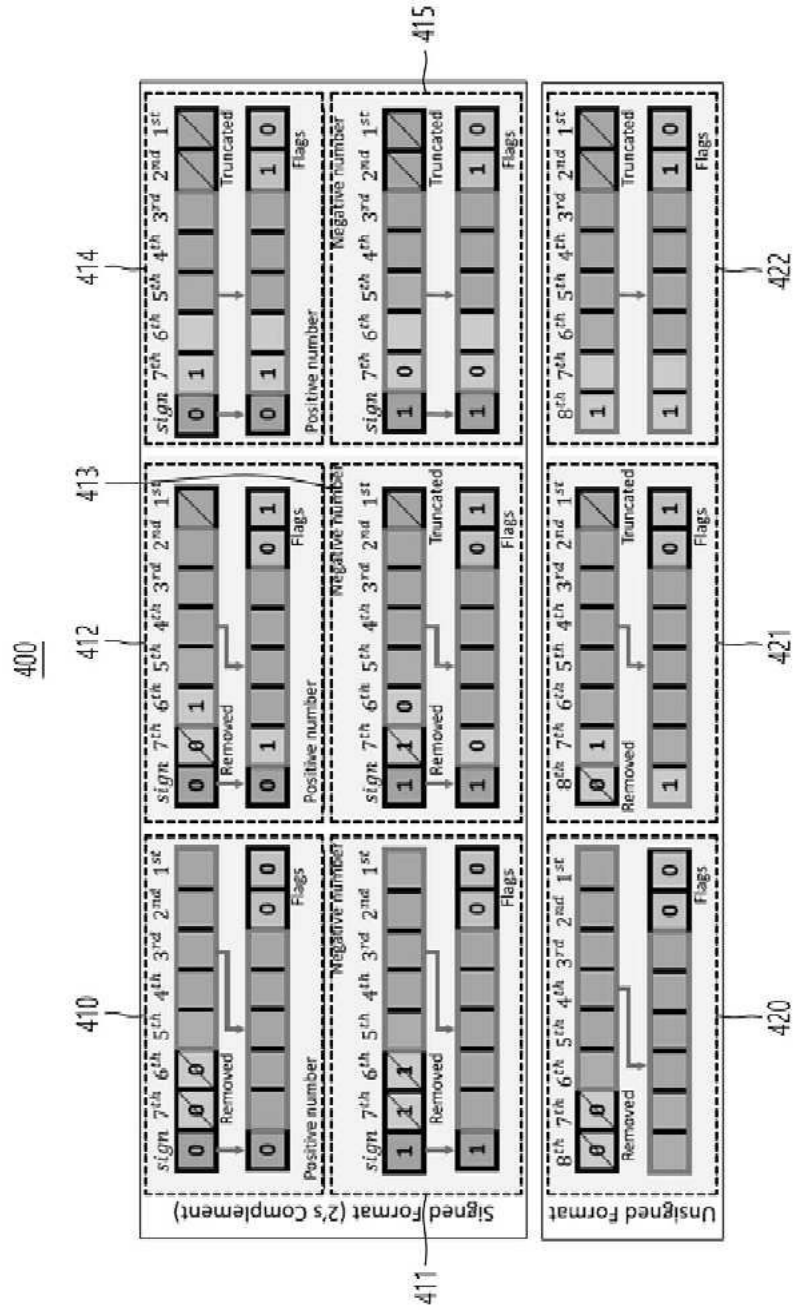
도면2b



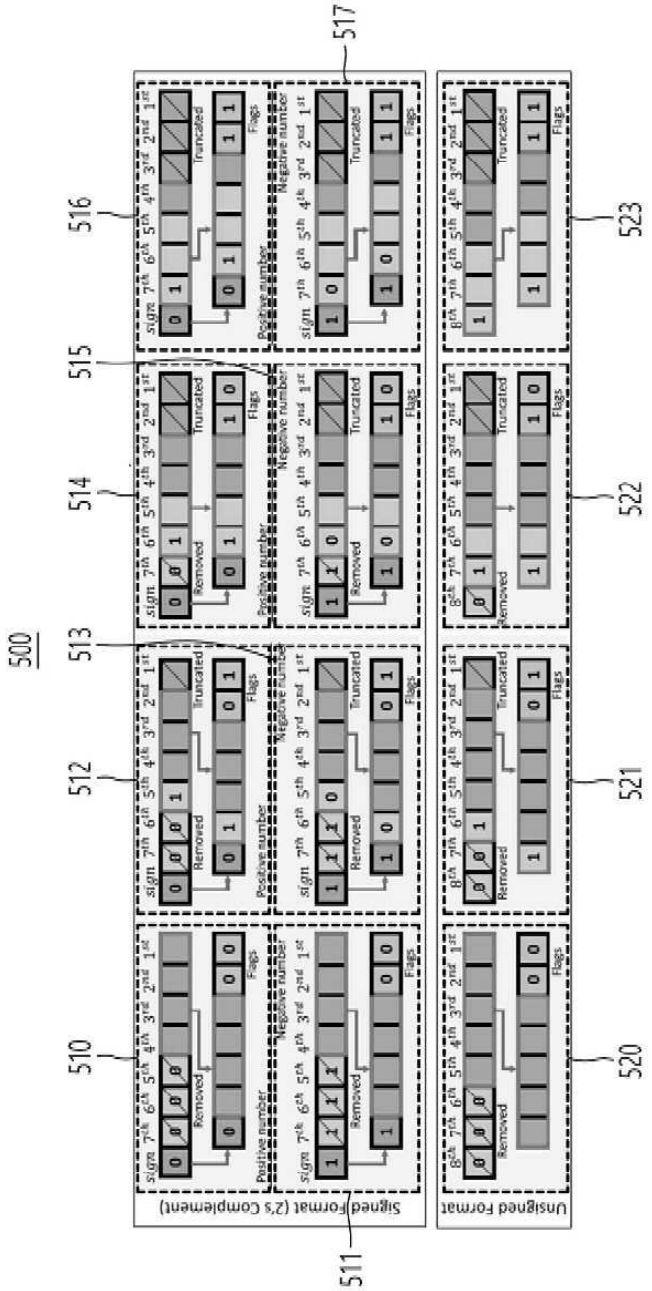
도면3


300
Weights
Accumulate of outliers = 17.65%
outliers
non-outliers
outliers
Density
0.015
0.010
0.005
0.000
−100
0
100
Decimal values
310
Activations
Accumulate of outliers = 0.13%
outliers
non-outliers
0.020
0.015
0.010
0.005
0.000
0
100
200
Decimal values

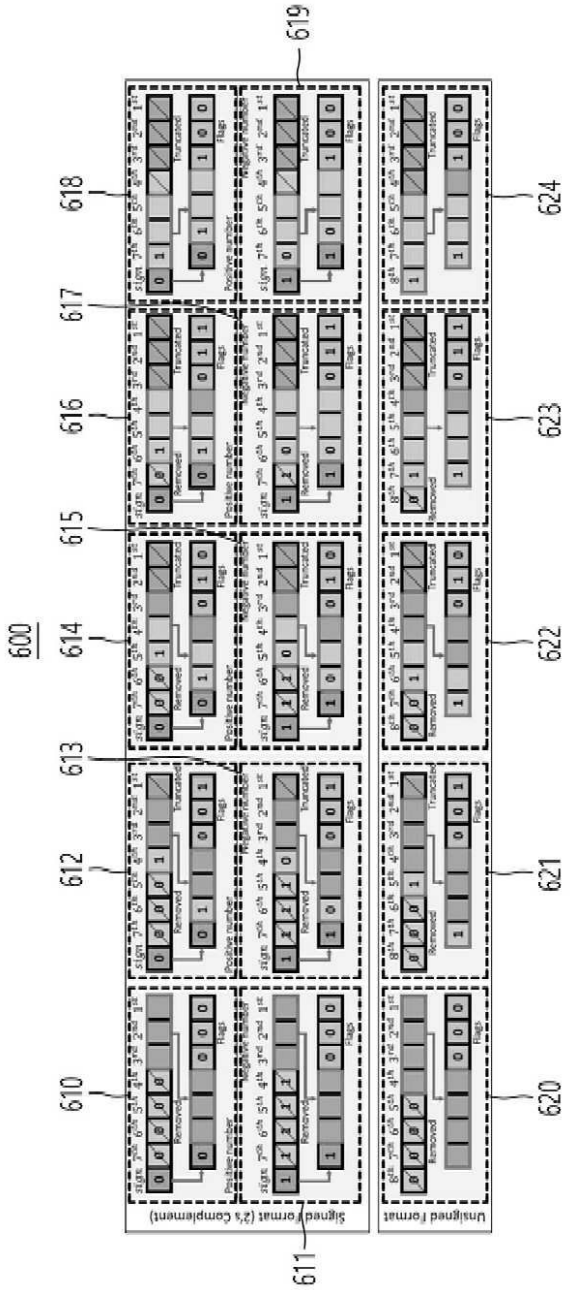
도면4


400
410
411
412
413
414
415
420
421
422
Signed Format (2's Complement)
Unsigned Format
Positive number
Negative number
Removed
Truncated
Flags
sign

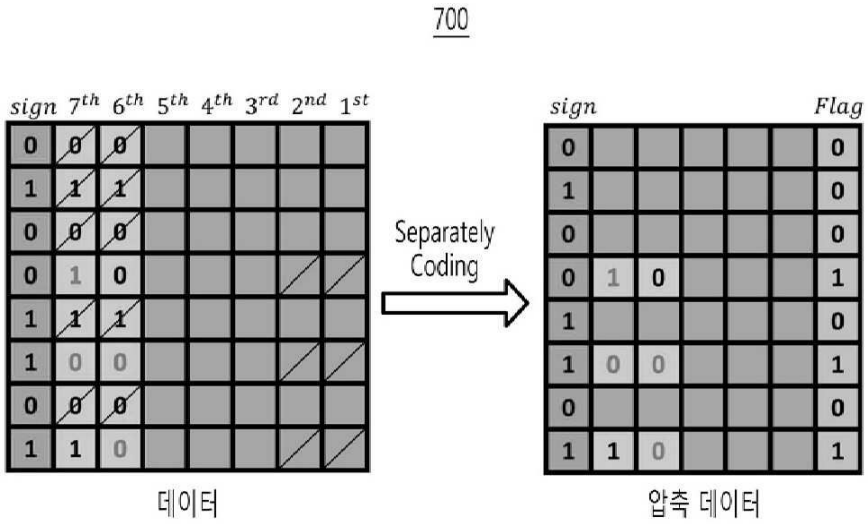
도면5


500
510
511
512
513
514
515
516
517
520
521
522
523
Signed Format (2's Complement)
Unsigned Format
Positive number
Negative number
Removed
Truncated
Flags

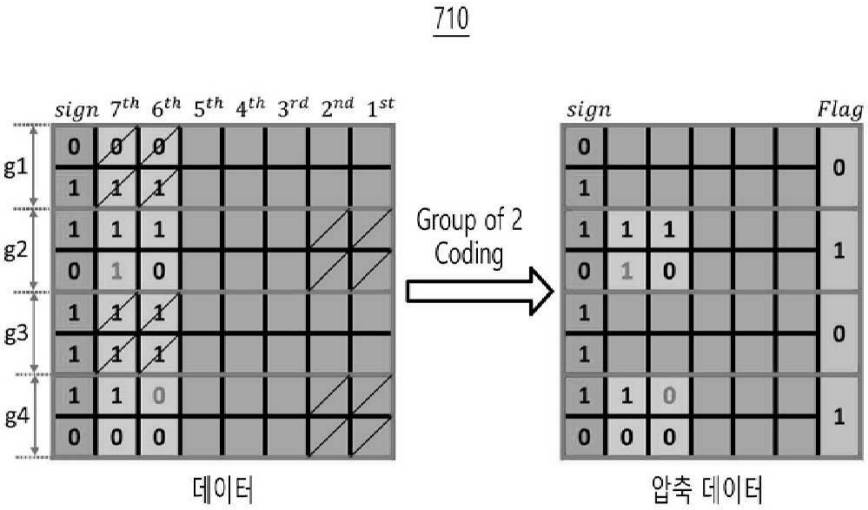
도면6


600
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
Signed format (2's complement)
Unsigned format

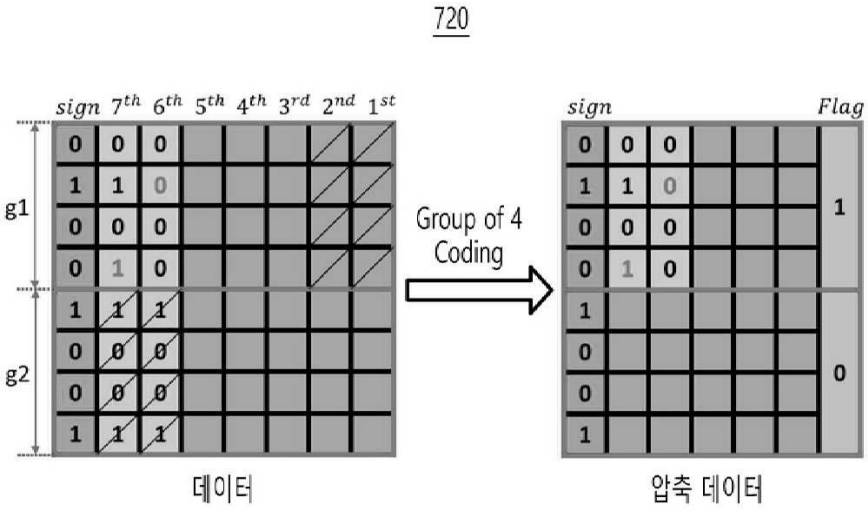
도면7a



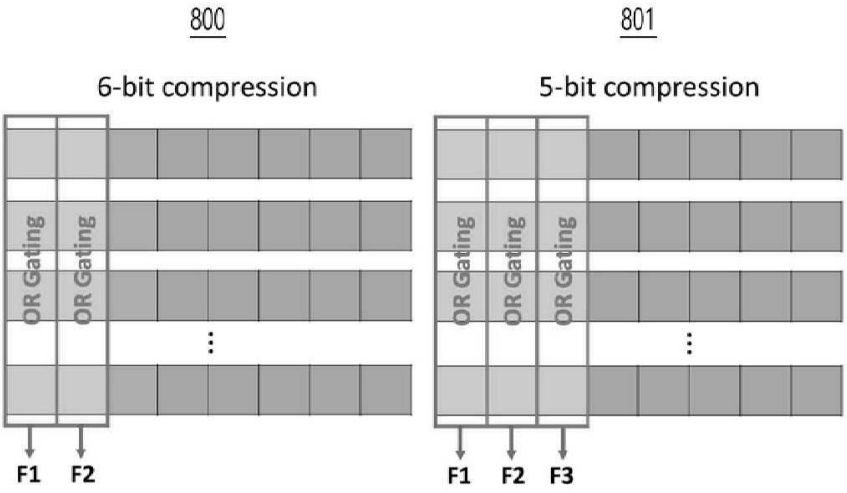
도면7b



도면7c

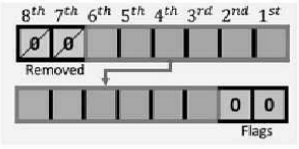
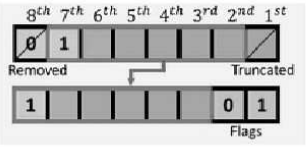
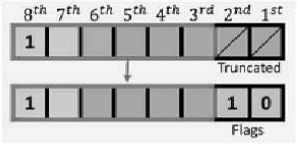


도면8a



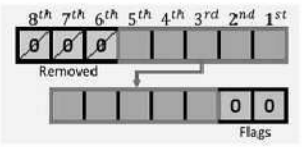
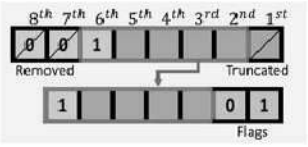
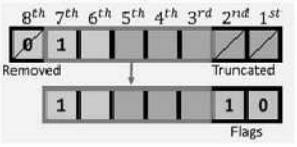
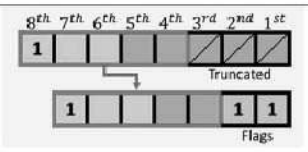
도면8b

810

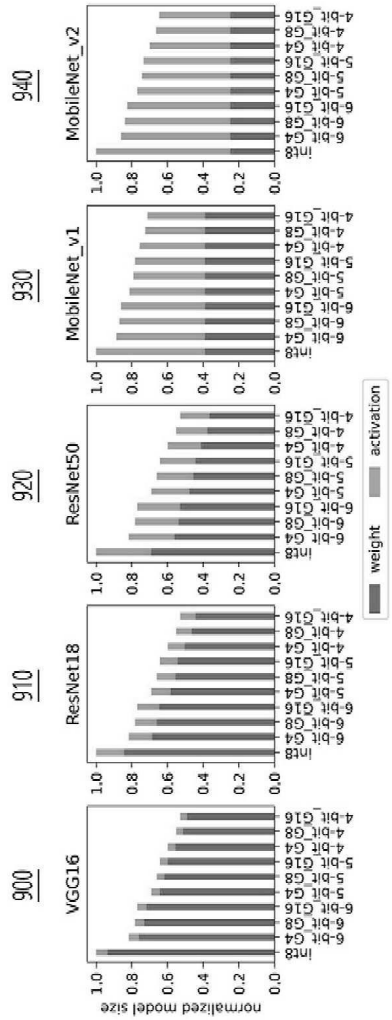
F1, F2	6-bit compression	Summary
0, 0		- Non-outlier - Shift left 2-bit - $e_{max} = \Delta/2$
0, 1		- Outlier level 1 - Shift left 1-bit - Truncate 1-LSB - $e_{max} = \Delta$
1, x		- Outlier level 2 - Truncate 2-LSB - $e_{max} = 2\Delta$

도면8c

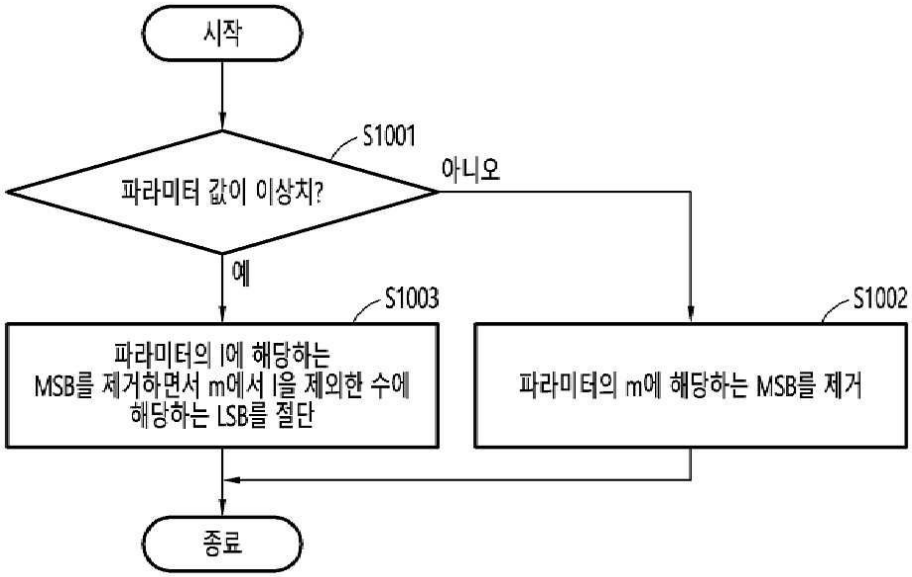
820

F1, F2, F3	5-bit compression	Summary
0, 0, 0		- Non-outlier - Shift left 3-bit - $e_{max} = \Delta/2$
0, 0, 1		- Outlier level 1 - Shift left 2-bit - Truncate 1-LSB - $e_{max} = \Delta$
0, 1, x		- Outlier level 2 - Shift left 1-bit - Truncate 2-LSB - $e_{max} = 2\Delta$
1, x, x		- Outlier level 3 - Truncate 3-LSB - $e_{max} = 4\Delta$

도면9


940
MobileNet_v2
930
MobileNet_v1
920
ResNet50
910
ResNet18
900
VGG16
int8
6-bit_G4
6-bit_G8
6-bit_G16
5-bit_G4
5-bit_G8
5-bit_G16
4-bit_G4
4-bit_G8
4-bit_G16
1.0
0.8
0.6
0.4
0.2
0.0
normalized model size
weight
activation

도면10


시작
S1001
파라미터 값이 이상치?
아니오
예
S1003
파라미터의 l에 해당하는
MSB를 제거하면서 m에서 l을 제외한 수에
해당하는 LSB를 절단
S1002
파라미터의 m에 해당하는 MSB를 제거
종료

도면11

