(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0136476
(43) 공개일자 2021년11월17일

(51) 국제특허분류(Int. Cl.)
H04N 19/96 (2014.01) *G06N 3/063* (2006.01)
H04N 19/103 (2014.01) *H04N 19/124* (2014.01)
H04N 19/184 (2014.01)
(52) CPC특허분류
H04N 19/96 (2015.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2020-0054770
(22) 출원일자 2020년05월07일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
전성호
경기도 화성시 동탄대로12길 64, 1833동 1403호(오산동, 동탄2신도시 금강펜테리움 센트럴파크Ⅰ)
박준석
경기도 화성시 동탄기흥로 393-15, 1502동 1001호(오산동, 동탄역 반도유보라 아이비파크 5.0)
(뒷면에 계속)
(74) 대리인
리앤목특허법인

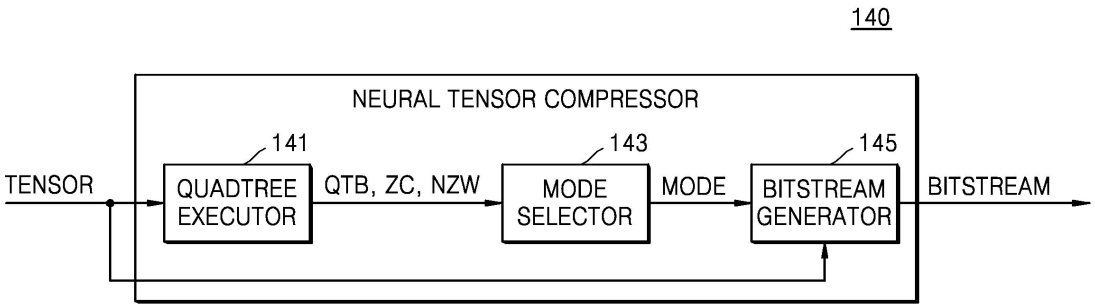
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **쿼드 트리 방법의 파라미터들을 이용하여 압축하는 장치 및 방법**

(57) 요 약

본 개시의 기술적 사상에 따른 장치는 복수의 셀들을 포함하는 텐서(tensor)를 압축하기 위한 것으로서, 상기 텐서에 포함된 논-제로(non-zero) 셀을 탐색하는 쿼드 트리(quad tree)를 생성하고, 상기 쿼드 트리로부터 적어도 하나의 파라미터 값을 추출하도록 구성된 쿼드 트리 생성기, 상기 적어도 하나의 파라미터에 기초하여 압축 모드를 결정하도록 구성된 모드 선택기 및 상기 압축 모드에 기초하여 상기 텐서를 압축함으로써 비트스트림을 생성하도록 구성된 비트스트림 생성기를 포함하는 장치.

대 표 도



(52) CPC특허분류
H04N 19/103 (2015.01)
H04N 19/124 (2015.01)
H04N 19/184 (2015.01)
(72) 발명자
서상민
서울특별시 서초구 효령로2길 30, 603호(방배동, 방배엔스위트)
이현수
경기도 화성시 동탄반석로 277, 122동 1402호(석우동, 동탄예당마을 우미린제일풍경채)
장혁재
경기도 수원시 영통구 영통로154번길 56, 102동 1901호(망포동, 한양수자인 에듀파크)
정경아
서울특별시 송파구 문정로 83, 129동 1804호(문정동, 문정래미안아파트)

명 세 서

청구범위

청구항 1

복수의 셀들을 포함하는 텐서(tensor)를 압축하기 위한 장치로서,

상기 텐서에 포함된 논-제로(non-zero) 셀을 탐색하는 쿼드 트리(quad tree)를 생성하고, 상기 쿼드 트리로부터 적어도 하나의 파라미터 값을 추출하도록 구성된 쿼드 트리 생성기;

상기 적어도 하나의 파라미터에 기초하여 압축 모드를 결정하도록 구성된 모드 선택기; 및

상기 압축 모드에 기초하여 상기 텐서를 압축함으로써 비트스트림을 생성하도록 구성된 비트스트림 생성기를 포함하는 장치.

청구항 2

제1항에 있어서,

상기 적어도 하나의 파라미터는,

쿼드 트리 결과로서 생성되고, 상기 논-제로 셀의 위치 정보가 표현된 비트들의 총 수에 상응하는 제1 파라미터;

논-제로 셀 중 가장 큰 값을 가지는 셀이 이진수로 표현될 때, 최하위 비트부터 0이 아닌 가장 큰 자리수의 비트까지의 비트 수에 상응하는 제2 파라미터; 및

상기 복수의 셀들 중 제로 셀들의 수에 상응하는 제3 파라미터를 포함하는 것을 특징으로 하는 장치.

청구항 3

제2항에 있어서,

상기 모드 선택기는,

상기 제1 파라미터 값이 복수의 셀들 수 이하인 경우, 쿼드 트리 방식으로 상기 텐서를 압축하는 제1 압축 모드를 상기 압축 모드로 선택하는 것을 특징으로 하는 장치.

청구항 4

제2항에 있어서,

상기 모드 선택기는,

상기 제1 파라미터 값이 상기 복수의 셀들 수 초과이고, 상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값이 상기 복수의 셀들 수 초과인 경우,

상기 논-제로 셀을 '1'로, 상기 제로 셀을 '0'으로 간주하는 제로 비트맵 방식으로 상기 텐서를 압축하는 제2 압축 모드를 상기 압축 모드로 선택하는 것을 특징으로 하는 장치.

청구항 5

제2항에 있어서,

상기 모드 선택기는,

상기 제1 파라미터 값이 상기 복수의 셀들 수 초과이고, 상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값이 상기 복수의 셀들 수 이하인 경우,

상기 복수의 셀 중 가장 큰 셀 값의 비트 폭(bitwidth)에 기초하여 상기 텐서를 압축하는 고정 길이 방식으로 상기 텐서를 압축하는 제3 압축 모드를 상기 압축 모드로 선택하는 것을 특징으로 하는 장치.

청구항 6

제1항에 있어서,

상기 텐서는 4의 M승(M은 자연수)개의 셀을 포함하는 것을 특징으로 하는 장치.

청구항 7

제1항에 있어서,

상기 텐서는 피처맵(Feature Map) 및 웨이트(Weight) 중 적어도 하나를 포함하는 것을 특징으로 하는 장치.

청구항 8

제1항에 있어서,

상기 비트스트림 생성기는, 상기 압축 모드에 대응되는 적어도 하나의 저장 영역 중 하나로 상기 비트스트림을 출력하는 것을 특징으로 하는 장치.

청구항 9

뉴럴 네트워크를 이용하여 입력 데이터에 대한 연산을 수행함으로써 복수의 셀들을 포함하는 텐서를 생성하도록 구성된 연산 회로; 및

상기 텐서를 압축함으로써 비트스트림을 출력하도록 구성된 뉴럴 텐서(Neural Tensor) 압축기를 포함하고,

상기 뉴럴 텐서 압축기는,

상기 텐서에 포함된 논-제로(non-zero) 셀을 찾기 위해 반복적 공간 분할 방식의 쿼드 트리(quad tree)를 생성하고, 상기 쿼드 트리로부터 적어도 하나의 파라미터 값을 추출하고, 상기 적어도 하나의 파라미터에 기초하여 상기 비트스트림의 압축 모드를 결정하는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 10

제9항에 있어서,

쿼드 트리 결과로서 생성되고, 상기 논-제로 셀의 위치 정보가 표현된 비트들의 총 수에 상응하는 제1 파라미터;

논-제로 셀 중 가장 큰 값을 가지는 셀이 이진수로 표현될 때, 최하위 비트부터 0이 아닌 가장 큰 자리수의 비트까지의 비트 수에 상응하는 제2 파라미터; 및

상기 복수의 셀들 중 제로 셀들의 수에 상응하는 제3 파라미터를 포함하는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 11

제10항에 있어서,

상기 뉴럴 텐서 압축기는,

상기 제1 파라미터 값이 상기 복수의 셀들 수 이하일 때, 쿼드 트리 방식을 적용하여 상기 텐서를 압축하는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 12

제10항에 있어서,

상기 뉴럴 텐서 압축기는,

상기 제1 파라미터 값이 상기 복수의 셀들 수보다 큼에 응답하여, 상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값이 상기 복수의 셀들 수 보다 클 때, 상기 논-제로 셀을 '1'로, 상기 제로 셀을 '0'으로 간주하는 제로 비트맵 방식을 적용하여 비트스트림을 압축하는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 13

제10항에 있어서,

상기 제1 파라미터 값이 상기 복수의 셀들 수보다 큼에 응답하여, 상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값이 상기 복수의 셀들 수 이하일 때, 상기 복수의 셀 중 가장 큰 셀 값의 비트 폭(bitwidth)에 기초하여 상기 텐서를 압축하는 고정 길이 방식을 적용하여 비트스트림을 압축하는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 14

제9항에 있어서,

상기 텐서는 4의 M승(M은 자연수)개의 셀을 포함하는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 15

제9항에 있어서,

상기 비트스트림은, 상기 압축 모드에 대응되어 마련된 적어도 하나의 저장 영역 중 하나로 출력되는 것을 특징으로 하는 뉴럴 네트워크 프로세서.

청구항 16

피처맵(Feature Map)과 웨이트(Weight)에 대한 사칙 연산이 반복된 결과인 텐서를 수신하는 단계;

상기 텐서에 포함된 복수의 셀 중 제로(zero) 셀을 압축하기 위해 상기 텐서를 반복적으로 공간 분할한 결과, 적어도 하나의 파라미터를 추출하는 단계;

상기 적어도 하나의 파라미터에 기초하여 압축 모드를 결정하는 단계; 및

상기 압축 모드에 기초하여, 비트스트림을 출력하는 단계를 포함하는 방법.

청구항 17

제16항에 있어서,

상기 적어도 하나의 파라미터를 추출하는 단계는,

쿼드 트리 결과로서 생성되고, 상기 논-제로 셀의 위치 정보가 표현된 비트들의 총 수에 상응하는 제1 파라미터를 추출하는 단계;

논-제로 셀 중 가장 큰 값을 가지는 셀이 이진수로 표현될 때, 최하위 비트부터 0이 아닌 가장 큰 자리수의 비트까지의 비트 수에 상응하는 제2 파라미터를 추출하는 단계; 및

상기 복수의 셀들 중 제로 셀들의 수에 상응하는 제3 파라미터를 추출하는 단계를 포함하는 방법.

청구항 18

제17항에 있어서,

상기 압축 모드를 결정하는 단계는,

상기 제1 파라미터 값과 상기 복수의 셀들 수를 비교하는 단계; 및

상기 제1 파라미터 값이 상기 복수의 셀들 수보다 큰 경우, 상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값을 상기 복수의 셀들 수와 비교하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 19

제18항에 있어서,

상기 제1 파라미터 값이 상기 복수의 셀들 수 이하인 경우, 상기 텐서를 쿼드 트리 방식으로 압축하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 20

제18항에 있어서,

상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값이 상기 복수의 셀들 수 보다 클 때, 상기 논-제로 셀을 '1'로, 제로 셀을 '0'으로 간주하는 제로 비트맵 방식으로 상기 텐서를 압축하는 단계; 및

상기 제2 파라미터 값과 상기 제3 파라미터 값을 곱한 값이 상기 복수의 셀들 수 이하일 때, 상기 복수의 셀 중 가장 큰 셀 값의 비트 폭에 기초하여 상기 텐서를 압축하는 고정 길이 방식으로 상기 텐서를 압축하는 단계를 더 포함하는 것을 특징으로 하는 방법.

발명의 설명

기 술 분 야

[0001] 본 개시의 기술적 사상은 데이터를 압축하는 장치 및 방법에 관한 것으로서, 구체적으로는 뉴럴 네트워크를 이용하는 쿼드 트리 방법의 파라미터들을 이용하여 텐서를 압축하는 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 뉴럴 네트워크(Neural Network)는 생물학적 뇌를 모델링한 컴퓨터 과학적 아키텍처(computational architecture)로 구현된다. 뉴럴 네트워크 프로세서는 큰 입력 데이터에 대해 많은 양의 연산을 수행하므로, 데이터의 빠른 처리, 저장, 독출이 요구된다.

[0003] 뉴럴 네트워크 구조에 텐서(Tensor) 개념이 이용된다. 텐서는 벡터의 일반화된 표현 방식으로, 하나의 텐서에는 복수의 웨이트(Weight)와 피처맵(Feature map)이 포함될 수 있다. 뉴럴 네트워크는 연산, 저장 및/또는 압축의 기본 처리 단위로 텐서를 이용할 수 있다.

발명의 내용

해결하려는 과제

[0004] 본 개시의 기술적 사상이 해결하고자 하는 과제는 텐서를 효율적으로 압축하는 뉴럴 텐서 압축기, 및 이를 포함하는 뉴럴 네트워크 프로세서, 및 이의 동작 방법을 제공하는 데 있다.

[0005] 본 개시의 기술적 사상이 해결하고자 하는 또 다른 과제는 데이터의 특성을 고려한 양자화 방법을 제공하는데 있다.

과제의 해결 수단

[0006] 상기와 같은 목적을 달성하기 위하여, 본 개시의 기술적 사상의 일 측면에 따른 장치는 복수의 셀들을 포함하는 텐서(tensor)를 압축하기 위한 것으로서, 상기 텐서에 포함된 논-제로(non-zero) 셀을 탐색하는 쿼드 트리(quad tree)를 생성하고, 상기 쿼드 트리로부터 적어도 하나의 파라미터 값을 추출하도록 구성된 쿼드 트리 생성기, 상기 적어도 하나의 파라미터에 기초하여 압축 모드를 결정하도록 구성된 모드 선택기 및 상기 압축 모드에 기초하여 상기 텐서를 압축함으로써 비트스트림을 생성하도록 구성된 비트스트림 생성기를 포함할 수 있다.

[0007] 본 개시의 기술적 사상의 일 측면에 따른 뉴럴 네트워크 프로세서는 뉴럴 네트워크를 이용하여 입력 데이터에 대한 연산을 수행함으로써 복수의 셀들을 포함하는 텐서를 생성하도록 구성된 연산 회로 및 상기 텐서에 대한 압축을 수행함으로써 비트스트림을 출력하도록 구성된 압축기를 포함하고, 상기 압축기는, 상기 텐서에 포함된 논-제로(non-zero) 셀을 찾기 위해 반복적 공간 분할 방식의 쿼드 트리(quad tree)를 생성하고, 상기 쿼드 트리로부터 적어도 하나의 파라미터 값을 추출하고, 상기 적어도 하나의 파라미터에 기초하여 상기 비트스트림의 압축 모드를 결정하는 것을 특징으로 할 수 있다.

[0008] 본 개시의 기술적 사상의 일 측면에 따른 데이터를 압축하는 방법은 피처맵(Feature Map)과 웨이트(Weight)에 대한 사칙 연산이 반복된 결과인 텐서를 수신하는 단계, 상기 텐서에 포함된 복수의 셀 중 제로(zero) 셀을 압축하기 위해 상기 텐서를 반복적으로 공간 분할한 결과, 적어도 하나의 파라미터를 추출하는 단계, 상기 적어도 하나의 파라미터에 기초하여 압축 모드를 결정하는 단계 및 상기 압축 모드에 기초하여, 비트스트림을 출력하는 단계를 포함할 수 있다.

[0009] 본 개시의 기술적 사상의 일 측면에 따른 동작 방법은 뉴럴 네트워크를 이용하여 피처맵(Feature Map)과 웨이트(Weight)에 대한 연산을 수행하는 뉴럴 네트워크 프로세서에 있어서, 연산 결과, 복수 개의 셀들을 포함하는 텐서를 수신하는 단계, 최대 셀 값에 기초하여 양자화 범위를 설정하는 단계, 제1 양자화 범위에 포함된 셀을 양자화하지 않고, 상기 제1 양자화 범위에 포함되지 않은 셀을 양자화하는 단계, 양자화된 텐서에 대해 쿼드 트리(quad tree) 데이터 구조를 적용함으로써 복수의 파라미터들을 추출하는 단계, 및 상기 복수의 파라미터들에 기초하여 상기 쿼드 트리 기반 비트스트림의 생성 여부를 결정하는 단계를 포함할 수 있다.

발명의 효과

[0010] 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기를 포함하는 뉴럴 네트워크 프로세서는 제로 값을 가지는 셀이 적은 특성을 가지는 텐서를 적응적으로 압축함으로써, 압축 효율을 증대시킬 수 있다.

[0011] 또한, 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기는 이미 생성된 파라미터를 이용하여 압축 모드를 판단하기 때문에 구현이 간단하다.

[0012] 또한, 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기는 텐서에 포함된 데이터의 특성을 고려하여 일부에 대해서만 양자함으로써 데이터 손실은 최소화하되, 압축 효율은 극대화할 수 있다.

도면의 간단한 설명

[0013] 도 1은 본 개시의 예시적 실시예에 따른 외부 메모리 및 뉴럴 네트워크 프로세서를 도시하는 블록도이다.

도 2는 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 블록도이다.

도 3a 및 도 3b는 본 개시의 예시적 실시예에 따른 쿼드 트리 생성기에서 수행되는 쿼드 트리 기반 압축 방법을 나타내는 도면이다.

도 4는 본 개시의 예시적 실시예에 따른 압축 모드를 결정하는 방법을 나타내는 흐름도이다.

도 5는 본 개시의 예시적 실시예에 따른 비트스트림의 구조를 도시하는 도면이다.

도 6은 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

도 7은 본 개시의 예시적 실시예에 따른 압축기의 동작 방법을 나타내는 흐름도이다.

도 8은 뉴럴 네트워크를 설명하기 위한 도면이다.

도 9는 본 개시의 예시적 실시예에 따른 뉴럴 네트워크의 컨볼루션 연산을 설명하기 위한 도면이다.

도 10은 본 개시의 예시적 실시예에 따른 양자화기를 더 포함하는 뉴럴 텐서 압축기의 블록도이다.

도 11은 본 개시의 예시적 실시예에 따른 셀 값에 따른 셀 분포를 도시하는 그래프이다.

도 12는 본 개시의 예시적 실시예에 따른 양자화기의 동작 방법을 나타내는 흐름도이다.

도 13은 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

도 14는 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

도 15는 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

도 16은 본 개시의 예시적 실시예에 따른 전자 시스템을 나타내는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0014] 이하, 첨부한 도면을 참조하여 본 개시의 실시 예에 대해 상세히 설명한다. 도 1은 본 개시의 예시적 실시예에 따른 뉴럴 네트워크 프로세서(100) 및 외부 메모리(300)를 포함하는 전자 장치(10) 를 도시하는 블록도이다.

[0015] 전자 장치(10)는 뉴럴 네트워크를 기초로 입력 데이터를 실시간으로 분석하여 유효한 정보를 추출하고, 추출된 정보를 기초로 상황을 판단하거나 전자 장치(10)에 포함된 적어도 하나의 구성을 제어할 수 있다. 예를 들어, 전자 장치(10)는 드론(drone), 첨단 운전자 보조 시스템(Advanced Drivers Assistance System; ADAS), 로봇 장치, 스마트 TV, 스마트 폰, 의료 장치, 모바일 장치, 영상 표시 장치, 계측 장치, IoT(Internet of Things) 장치 등에 적용될 수 있으며, 이외에도 다양한 종류의 전자 장치로서 이용될 수 있다..

[0016] 전자 장치(10)는 뉴럴 네트워크 프로세서(100) 및 외부 메모리(300)를 포함할 수 있다. 그러나, 이에 제한되지 않고, 적어도 하나의 IP(Intellectual Property) 블럭을 더 포함할 수 있다. 예를 들어, 전자 장치(10)는 뉴럴 네트워크 프로세서(100), 외부 메모리(300) 뿐만 아니라, 스토리지, 센서 등 뉴럴 네트워크 프로세서(100)의 처리가 요구되는 적어도 하나의 IP 블럭을 더 포함할 수 있다.

[0017] 뉴럴 네트워크 프로세서(100)는 뉴럴 네트워크를 생성하거나, 뉴럴 네트워크를 훈련(train, 또는 학습(learn))하거나, 수신되는 입력 데이터를 기초로 연산을 수행하고, 수행 결과를 기초로 정보 신호(information signal)를 생성하거나, 뉴럴 네트워크를 재훈련(retrain)할 수 있다. 뉴럴 네트워크의 모델들은 GoogleNet, AlexNet, VGG Network 등과 같은 CNN(Convolution Neural Network), R-CNN(Region with Convolution Neural Network), RPN(Region Proposal Network), RNN(Recurrent Neural Network), S-DNN(Stacking-based deep Neural Network), S-SDNN(State-Space Dynamic Neural Network), Deconvolution Network, DBN(Deep Belief Network), RBM(Restrcted Boltzman Machine), Fully Convolutional Network, LSTM(Long Short-Term Memory) Network, Classification Network 등 다양한 종류의 모델들을 포함할 수 있으나 이에 제한되지는 않는다. 뉴럴 네트워크 프로세서(100)는 뉴럴 네트워크의 모델들에 따른 연산을 수행하기 위한 하나 이상의 프로세서를 포함할 수 있다.

[0018] 뉴럴 네트워크 프로세서(100)는 뉴럴 네트워크의 모델들에 대응되는 프로그램들을 저장하기 위한 별도의 메모리를 내부 메모리로서 포함할 수도 있다. 뉴럴 네트워크 프로세서(100)는 뉴럴 네트워크 처리 장치(neural network processing device), 뉴럴 네트워크 집적 회로(neural network integrated circuit) 또는 뉴럴 네트워크 처리 유닛(Neural network Processing Unit; 이하, NPU) 등으로 달리 호칭될 수 있다.

[0019] 예시적인 실시예에 따라, 뉴럴 네트워크 프로세서(100)가 생성하는 정보 신호는 음성 인식 신호, 사물 인식 신호, 영상 인식 신호, 생체 정보 인식 신호 등과 같은 다양한 종류의 인식 신호들 중 적어도 하나를 포함할 수 있다.

[0020] 예시적 실시예에서, 뉴럴 네트워크 프로세서(100)는 비디오 스트림에 포함되는 프레임 데이터를 입력 데이터로서 수신하고, 프레임 데이터로부터 프레임 데이터가 나타내는 이미지에 포함된 사물에 대한 인식 신호를 생성할 수 있다. 예를 들어, 뉴럴 네트워크 프로세서(100)는 카메라로부터 제공되는 프레임 데이터인 입력 데이터에 기초하여, 안면 인식 신호를 생성할 수 있다.

[0021] 예시적 실시예에서, 뉴럴 네트워크 프로세서(100)는 오디오 스트림에 포함되는 주파수 데이터를 입력 데이터로 수신하고, 주파수 데이터로부터 추출되는 음성에 대한 음석 인식 신호를 생성할 수 있다. 또 다른 예로, 하지만, 이에 제한되는 것은 아니며, 뉴럴 네트워크 프로세서(100)는 다양한 종류의 입력 데이터를 수신할 수 있고, 입력 데이터에 따른 인식 신호를 생성할 수 있다.

[0022] 뉴럴 네트워크의 연산 특성상 제로(zero) 값을 가지는 데이터가 많이 발생하기 때문에, 뉴럴 네트워크 프로세서(100)는 제로 값을 가지는 데이터를 솎아냄으로써 데이터를 압축할 수 있다.

[0023] 본 개시의 예시적 실시예에 따르면, 뉴럴 네트워크 프로세서(100)는 컨볼루션 연산에 이용되는 입력 피처맵 데이터(Input featuremap data)에 포함되는 복수의 셀들 중, 데이터 값으로서 '0'을 갖는 제로(zero) 셀을 제거하고, 데이터 값으로서 '0'을 갖지 않는 논-제로(non-zero) 셀의 데이터 값 및 논-제로 셀의 위치 정보를 이용하여 데이터를 압축할 수 있다. 뉴럴 네트워크 프로세서(100)는 데이터를 압축시킴으로써 데이터의 처리, 저장, 로딩, 독출의 속도를 향상 시킬 수 있다. 또한, 뉴럴 네트워크 프로세서(100)는 압축된 데이터를 외부 메모리(300)에 저장하거나, 외부 메모리(300)로부터 압축된 데이터를 로딩함으로써 데이터 입출력 속도를 증가시킬 수 있다.

[0024] 뉴럴 네트워크 프로세서(100)는 뉴럴 텐서 압축 해제기(110), 내부 메모리(120), 연산 회로(130), 뉴럴 텐서 압축기(140)를 포함할 수 있다.

[0025] 뉴럴 텐서 압축 해제기(110)는 외부 메모리(300)에 압축된 형태로 저장된 데이터를 로드하고, 데이터 압축을 해제할 수 있다. 예시적인 실시예에서, 뉴럴 텐서 압축 해제기(110)는 뉴럴 텐서 압축기(140)와 데이터를 압축했던 방식의 역순으로 데이터를 압축 해제할 수 있다.

[0026] 예시적인 실시예에서, 뉴럴 텐서 압축 해제기(110)는 데이터가 저장되어 있던 외부 메모리(300) 내의 메모리 주소를 참조함으로써, 데이터가 압축된 압축 알고리즘을 판단할 수 있고, 판단된 압축 알고리즘에 기초하여 압축된 데이터를 해제할 수 있다. 외부 메모리(300)는 압축 모드에 대응한 저장 영역을 포함할 수 있다. 예를 들어,

외부 메모리(300)에는 제1 압축 모드에 대응하는 제1 저장 영역, 제2 압축 모드에 대응하는 제2 저장 영역 및 제3 압축 모드에 대응하는 제3 저장 영역을 포함할 수 있다. 뉴럴 텐서 압축 해제기(110)는 데이터를 로딩하는 저장 영역(즉 저장 영역의 메모리 주소)으로부터 압축 모드를 판단할 수 있고, 압축 모드에 따른 디코딩 방식을 적용할 수 있다. 본 개시의 기술적 사상에 따르면, 저장 영역에 따라 데이터를 압축 해제할 수 있으므로, 저장된 비트스트림은 압축 모드에 대한 비트 정보를 포함하지 않을 수 있다. 압축 해제된 데이터는 내부 메모리(120)에 임시로 저장될 수 있다.

[0027] 내부 메모리(120)는 압축 해제된 데이터를 임시로 저장할 수 있거나, 연산 회로(130)에서 출력된 연산 결과(예를 들어, 텐서)를 임시로 저장할 수 있다.

[0028] 내부 메모리(120)는 뉴럴 네트워크 프로세서(100) 내에서의 빠른 데이터 처리를 위해 사용중인 데이터들을 임시로 보관할 수 있다. 뉴럴 네트워크 프로세서(100)와, 뉴럴 네트워크 프로세서(100) 외부에 구비된 외부 메모리(300) 간의 데이터 처리 대역폭(bandwidth)에는 한계가 있으므로, 뉴럴 네트워크 프로세서(100)는 내부 메모리(120)를 별도로 구비하여 빠른 데이터 처리를 도모할 수 있다. 예시적인 실시예에서, 내부 메모리(120)는 외부 메모리(300)에 비해 처리 속도가 빠르고 안정성이 높을 수 있지만, 이에 제한되지는 않는다. 예를 들어, 내부 메모리(120)는 SRAM(Static Ramdom Access Ram)을 포함할 수 있고, 외부 메모리(120)는 DRAM(Dynamic Random Access Memory), SDRAM(Synchronous Dynamic Random Access Memory)를 포함할 수 있다.

[0029] 연산 회로(130)는 내부 메모리(120)로부터 입력 피처맵 및 웨이트(Weight)를 포함하는 입력 데이터를 수신할 수 있다. 연산 회로(130)는 수신된 입력 피처맵 및 웨이트를 이용해 컨볼루션 연산을 수행함으로써 텐서를 생성할 수 있다. 텐서는 피처맵 및 웨이트를 포함할 수 있다.

[0030] 연산 회로(130)는 입력 피처맵과 웨이트에 대한 사칙 연산을 반복적으로 수행할 수 있다. 연산 회로(130)는 곱셈과 나눗셈, 덧셈, 뺄셈, 및 논리 연산을 수행할 수 있고, MAC(Multiplier-ACcumulator)이라고 지칭될 수 있다. 연산 회로(130)는 입력 피처맵과 웨이트에 대한 사칙 연산의 조합으로 복잡한 수학적 계산(예를 들면 미분, 적분)을 해결할 수 있다.

[0031] 뉴럴 텐서 압축기(140)는 연산 회로(130)에서 출력된 연산 결과를 내부 메모리(120)로부터 로딩할 수 있다. 연산 회로(130)에서 출력된 연산 결과는 텐서(tensor)로 지칭될 수 있다. 텐서는 벡터(vector)의 일반화된 표현일 수 있고, 복수의 셀들을 포함할 수 있다. 예시적인 실시예에서, 복수 개의 셀들이 매트릭스 형태로 배열됨으로써 피처맵을 구성할 수 있고, 피처맵은 뉴럴 네트워크의 깊이(depth)에 따라 여러 개 존재할 수 있다. 뉴럴 네트워크 프로세서(100)는 텐서 단위로 데이터를 처리할 수 있다. 뉴럴 텐서 압축기(140)는 텐서를 압축하고, 압축 결과를 외부 메모리(300)에 저장할 수 있다.

[0032] 예시적인 실시예에서, 뉴럴 텐서 압축기(140)는 압축 모드에 대응되는 저장 영역에 생성된 비트스트림을 출력할 수 있다. 외부 메모리(300)는 압축 모드에 대응한 저장 영역을 포함할 수 있음은 전술한 바와 같다. 예를 들어, 외부 메모리(300)에는 제1 압축 모드에 대응하는 제1 저장 영역, 제2 압축 모드에 대응하는 제2 저장 영역 및 제3 압축 모드에 대응하는 제3 저장 영역을 포함할 수 있다. 예를 들어, 제1 압축 모드에 기반한 비트스트림은 제1 저장 영역에 출력될 수 있고, 제2 압축 모드에 기반한 비트스트림은 제2 저장 영역에 출력될 수 있으며, 제3 압축 모드에 기반한 비트스트림은 제3 저장 영역에 출력될 수 있다. 본 개시의 기술적 사상에 따르면, 압축 모드에 대응하여 데이터를 다른 공간에 저장할 수 있으므로, 비트스트림은 압축 모드에 대한 비트 정보를 포함하지 않을 수 있다.

[0033] 뉴럴 텐서 압축 해제기(110), 연산 회로(130) 및 뉴럴 텐서 압축기(140)는 로직 회로를 포함하는 하드웨어와 같은 처리 회로로서 구현될 수 있거나, 압축 동작을 수행하는 소프트 웨어를 실행하는 프로세서와 같이 하드웨어와 소프트웨어의 조합으로 구현 될 수 있다. 특히, 처리 회로는 중앙 처리 장치(Central Procesing Unit; CPU), 산술 및 논리 연산, 비트 쉬프트 등을 수행하는 ALU(Arithmetic Logic Unit), DSP(Digital Signal Processor), 마이크로프로세서(microprocessor), ASIC(Application Specific Integrated Circuit) 등으로 구현될 수 있으나, 이에 제한되지 않는다.

[0034] 본 명세서에서 설명의 편의상, 텐서에 포함되는 복수의 셀들 중 데이터 값으로서 '0'을 갖는 셀을 제로 셀, 복수의 셀들 중 데이터 값으로서 '0'이 아닌 값을 갖는 셀을 논-제로 셀로 칭할 수 있다. 뉴럴 네트워크 연산 특성상, 텐서에 존재하는 제로 셀의 비율은 높을 수 있다.

[0035] 도 2는 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 블록도이다.

[0036] 뉴럴 텐서 압축기(140)는 쿼드 트리 생성기(141), 모드 선택기(143) 및 비트스트림 생성기(145)를 포함할 수 있

다.

[0037] 쿼드 트리 생성기(141)는 텐서에 포함된 논-제로 셀을 탐색하기 위한 쿼드 트리를 생성할 수 있다. 쿼드 트리(Quad tree)는 공간을 4개의 정사각형으로 계층적으로 분할하는 자료 구조일 수 있다. 예시적인 실시예에서, 쿼드 트리 생성기(141)는 텐서를 반복적으로 공간적 4 등분함으로써 논-제로 셀이 존재하지 않는 영역을 0으로 지정할 수 있고, 논-제로 셀이 존재하는 영역을 1로 지정할 수 있다. 0으로 지정된 영역에 포함된 셀은 모두 "0"의 셀 값을 가질 수 있어 압축될 수 있고, 1로 지정된 영역에 포함된 셀은 재차 공간적 분할되어 다음 하위 계층에서 논-제로 셀을 탐색할 수 있다. 쿼드 트리 생성기(141)는 텐서에 대한 쿼드 트리를 생성함으로써, 논-제로 셀의 최대 비트 폭(bitwidth)에 관한 정보를 나타내는 최대 비트 데이터, 논-제로 셀의 위치 정보를 나타내는 셀 위치 데이터, 및 논-제로 셀의 값을 나타내는 논-제로 데이터를 포함하는 적어도 하나의 텐서 데이터를 생성할 수 있다.

[0038] 쿼드 트리 생성기(141)는 논-제로 버퍼(미도시)를 구비할 수 있다. 논-제로 버퍼는 입력되는 텐서에 포함된 논-제로 셀을 버퍼링할 수 있다.

[0039] 쿼드 트리 생성기(141)는 생성된 쿼드 트리로부터 적어도 하나의 파라미터를 추출할 수 있다. 예시적인 실시예에서, 쿼드 트리 생성기(141)는 쿼드 트리 압축이 수행된 결과 생성되는 비트들의 수에 상응하는 '쿼드 트리 비트 수(QTB)' 파라미터, , 논-제로 셀의 최대 비트 폭(Non-zero bitWidth)에 상응하는 논-제로 셀 비트 폭(NZW)' 파라미터, 텐서에 포함된 제로 셀 개수(Zero Count)에 상응하는 '제로 개수(ZC)'을 추출할 수 있다.

[0040] 예시적인 실시예에서, 쿼드 트리 비트 수(QTB)는 쿼드 트리 결과로서 생성되고, 논-제로 셀의 위치 정보가 표현된 비트들의 총 수에 상응할 수 있다. 또한, 논-제로 비트 폭(NZW)은 논-제로 셀 중 가장 큰 값을 가지는 셀을 이진수로 표현될 때, 최하위 비트부터 0이 아닌 가장 큰 자리수의 비트까지의 비트 수에 상응할 수 있다. 또한, 제로 개수(zero count)는 상기 복수의 셀들 중 제로 셀들의 수에 상응할 수 있다.

[0041] 텐서로부터 쿼드 트리가 생성되는 방법 및 파라미터가 추출되는 방법은 도 3에서 상세히 설명된다.

[0042] 모드 선택기(143)는 쿼드 트리 비트 수(QTB), 제로 개수(ZC), 논-제로 비트 폭 (NZW) 중 적어도 하나에 기초하여 압축 모드를 선택할 수 있다. 예시적인 실시예에서, 모드 선택기(143)는 쿼드 트리 비트 수(QTB)에 기초하여 쿼드 트리 압축 모드를 선택할지 여부를 판단할 수 있다. 예시적인 실시예에서, 모드 선택기(143)는 제로 개수(ZC) 및 논-제로 비트 폭(NZW)에 기초하여, 제로 비트맵(zero bitmap) 압축 모드 또는 고정 길이(fixed length) 압축 모드 중 어느 하나를 선택할 수 있다. 모드 선택기(143)는 압축 모드를 지시하는 모드 신호(MODE)를 출력할 수 있다. 파라미터에 기초한 압축 모드 선택 방법은 도 4에서 상세히 설명된다.

[0043] 비트스트림 생성기(145)는 모드 신호(MODE)에 기초하여 텐서(TENSOR)를 압축하고, 압축 결과를 비트스트림(bitstream)으로 출력할 수 있다. 예시적인 실시예에서, 비트스트림 생성기(145)는 모드 신호(MODE)에 기초하여 비트스트림을 생성할 수 있다. 모드 신호(MODE)는 쿼드 트리 압축 모드를 지시하는 신호, 비트맵(zero bitmap) 압축 모드 신호, 또는 고정 길이(fixed length) 압축 모드 신호 중 적어도 어느 하나일 수 있다.

[0044] 비트스트림 생성기(145)는 쿼드 트리 압축 모드를 지시하는 모드 신호(MODE)에 기초하여, 제공된 텐서(TENSOR)를 쿼드 트리에 기반해 압축할 수 있다. 그러나, 이에 제한되지 않고, 비트스트림 생성기(145)는 쿼드 트리 압축 모드를 지시하는 모드 신호(MODE)에 기초하여, 쿼드 트리 생성기(141)에서 사용된 결과를 이용하여 쿼드 트리를 다시 생성하지 않고도 텐서(TENSOR)를 압축할 수 있다.

[0045] 뉴럴 네트워크의 컨볼루션 연산에 음수를 0으로 변환하는 ReLU(Rectified Linear Unit) 연산이 적용되는 경우가 많기 때문에, 피처맵에 제로 셀이 많이 분포할 수 있다. 피처맵에 논-제로 셀이 발생하더라도 특정 영역에 모여 있다면, 압축 효율은 크게 저하되지 않을 수 있다. 그러나, 피처맵에 논-제로 셀이 산발적으로 분포된다면, 쿼드 트리 압축을 한 데이터 크기가 비압축 데이터 크기보다 더 큰 오버헤드(head) 현상이 발생할 수 있다.

[0046] 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기(140) 및 뉴럴 텐서 압축기(140)를 포함하는 뉴럴 네트워크 프로세서(100)는 제로 값을 가지는 셀이 적은 특성을 가지는 텐서에 대해 적응적으로 압축함으로써, 제로 셀을 고려하지 않은 압축에 비해 압축 효율을 증대시킬 수 있다. 또한, 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기(140)는, 텐서를 쿼드 트리 기반 압축할 때 추출되는 파라미터를 이용하여 압축 모드를 결정할 수 있으므로, 압축 모드를 판단하는 속도가 빨라 압축 속도를 향상시킬 수 있고, 구현이 용이할 수 있다.

[0047] 또한, 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기(140)는 텐서에 포함된 복수의 셀들이 낮은 셀 값 영역에 주로 분포하는 특성을 고려하여 복수의 셀들 중 일부에 대해서만 양자함으로써 데이터 손실은 최소화하되, 압

축 효율은 극대화할 수 있다.

[0048] 도 3a 및 도 3b은 본 개시의 예시적 실시예에 따른 쿼드 트리 생성기에서 수행되는 쿼드 트리 기반 압축 방법을 나타내는 도면이다. 도 3a 및 도 3b는 도 2와 함께 참조된다.

[0049] 도 3a를 참조하면, 텐서는 복수의 피처맵(FM)을 포함할 수 있고, 피처맵은 셀(CELL)의 집합인 셀 그룹(CG)이 매트릭스 형태로 배열됨으로서 생성될 수 있다. 예를 들어, 셀(CELL) 4개를 2X2 매트릭스 형태로 배열한 집합을 셀 그룹(CG)으로 명명할 수 있고, 셀 그룹(CG) 4개를 2X2 매트릭스 형태로 배열한 집합을 피처맵(FM)으로 명명할 수 있으며, 피처맵(FM) 4개의 집합을 텐서라고 명명할 수 있다. 그러나 이에 제한되지 않고, 셀 그룹(CG)은 셀(CELL)이 4X4 매트릭스 형태로 배열될 수 있고, 피처맵(FM)은 셀 그룹(CG)이 4X4 매트릭스 형태로 배열될 수 있는 등 그 형태는 다양할 수 있다.

[0050] 셀(CELL)은 데이터를 표현하는 최소 단위일 수 있고, 셀 값은 16진수로 표현될 수 있다. 예시적인 실시예에서, 하나의 셀(CELL)은 2 개의 16진수를 표현할 수 있으므로 셀(CELL)이 표현하는 데이터는 8 비트(bit)(즉, 1바이트(Byte)로 구성될 수 있다. 그러나 이에 제한되지 않고, 셀(CELL)이 표현하는 데이터는 10비트 또는 12비트 등 데이터 표현 형식에 따라 달리 구성될 수 있다.

[0051] 셀 그룹(CG)은 인접한 셀(CELL) 4개가 2X2 매트릭스 형태로 배열될 수 있고, 셀 그룹(CG)의 크기는 4바이트일 수 있다. 예시적인 실시예에서, 셀 그룹(CG)은 피처맵(FM)이 1번 쿼드 트리 분할될 때 생성되는 피처맵(FM)의 서브 영역을 의미할 수 있다.

[0052] 피처맵(FM)은 복수의 셀들이 4X4 매트릭스 형태로 배열됨으로서 생성될 수 있다. 피처맵(FM) 4개의 배열로서 텐서가 형성될 수 있다. 그러나, 이에 제한되지 않고, 다양한 개수의 피처맵(FM)의 배열로서 하나의 텐서가 형성될 수 있다.

[0053] 예시적인 실시예에 따르면, 하나의 텐서에 포함된 셀들의 수는 뉴럴 네트워크의 깊이에 의존적일 수 있다. 예를 들어, 뉴럴 네트워크의 깊이(depth)가 3이라면, 셀(CELL)의 개수는 너비(width), 높이(height) 및 채널(channel) 방향으로 각각 3개의 방향축으로 4개씩의 셀을 가진 결과인 4^3=64일 수 있다. 이 경우, 피처맵(FM)은 너비 및 높이가 4개인 4X4 매트릭스 형태로 배열될 수 있으며, 피처맵(FM)의 개수는 4개로서 채널 방향의 셀 개수와 동일할 수 있다.

[0054] 설명의 편의를 위해 하나의 텐서에 64개의 셀들이 포함된 것으로 예시되나, 이에 제한되지 않는다. 예시적인 실시예에서, 뉴럴 네트워크의 깊이(depth)를 M으로 가정하면, 하나의 텐서에 포함된 셀들의 개수는 $N = 4^M$ 이다. 예를 들어, 하나의 텐서가 가지는 셀들의 개수는 256개일 수 있고, 뉴럴 네트워크의 깊이가 5라면 하나의 텐서가 가지는 셀들의 개수는 1024개일 수 있다.

[0055] 쿼드 트리 생성기(도 2, 140)는 쿼드 트리 방식에 기반하여 제1 텐서(TENSOR1)를 압축할 수 있다. 쿼드 트리 생성기(141)는 쿼드 트리 압축을 위해 제1 텐서(TENSOR1)를 행(row) 방향으로 탐색할 수 있다.

[0056] 쿼드 트리 생성기(141)는 제1 텐서(TENSOR1)를 쿼드 트리 기반으로 압축한 길이인 제1 압축 길이(LENGTH1)를 1이라고 판단할 수 있다. 예시적인 실시예에서, 쿼드 트리 생성기(141)는 논-제로 버퍼(미도시)를 구비할 수 있다. 논-제로 버퍼는 입력되는 텐서에 포함된 논-제로 셀을 버퍼링할 수 있다. 예시적인 실시예에 따라, 쿼드 트리 생성기(141)는 쿼드 트리 압축을 위해 제1 텐서(TENSOR1)을 행 방향으로 탐색한 결과, 버퍼링된 논-제로 셀이 존재하지 않음을 확인할 수 있다. 제1 텐서(TENSOR1)에 논-제로 셀이 존재하지 않음(즉, 모두 제로 셀)을 표현하기 위해 텐서에 할당될 수 있는 최소 길이는 1 바이트일 수 있다. 결과적으로, 제1 압축 길이(LENGTH1)가 1인 것은 논-제로 셀의 부존재를 의미하기 위해 1 바이트가 할당된 것이라고 이해될 수 있다.

[0057] 예시적인 실시예에 따라, 제1 압축 길이(LENGTH1)에 할당된 데이터의 크기는 6비트일 수 있다. 텐서에 4^3=64개의 셀들이 포함될 수 있고, 하나의 셀은 1 바이트의 크기를 가질 수 있기 때문에, 텐서의 길이는 모든 셀이 논-제로 셀인 경우 최대 64(=2^6)바이트일 수 있다.

[0058] 결과적으로, 1바이트부터 64바이트까지의 정보는 6비트로 표현될 수 있고, 제1 압축 길이(LENGTH1)는 이진수 $000000_{(2)}$ 로 표현될 수 있다. 1바이트는 8비트로 구성되기 때문에, 나머지 자리수 2개는 제로-패딩(zero-padding)될수 있다.

[0059] 제1 텐서(TENSOR1)에 포함된 논-제로 셀이 존재하지 않으므로, 제1 텐서에 포함된 가장 큰 논-제로 셀의 비트

길이인 제1 논-제로 비트 폭(NZW1) 및 논-제로 셀 값인 제1 논-제로 값(NZV1)은 모두 0일 수 있다. 제1 압축 길이(LENGTH1)가 1이라는 정보는 제1 텐서(TENSOR1)에 논-제로 셀이 포함되지 않는다는 정보와 상응하기 때문에, 제1 논-제로 비트 폭(NZW1) 및 논-제로 셀 값인 제1 논-제로 값(NZV1)의 비트들은 압축된 데이터에 포함되지 않을 수 있다.

[0060] 도 3b를 참조하면, 제2 텐서(TENSOR2)는 4개의 피처맵(FM1, FM2, FM3, FM4)을 포함할 수 있다. 쿼드 트리 생성기(도 2, 140)는 쿼드 트리 압축을 위해 제2 텐서(TENSOR2)를 행(row) 방향으로 탐색할 수 있다. 논-제로 버퍼는 제2 텐서(TENSOR2) 중 논-제로 셀을 버퍼링 할 수 있다. 버퍼링 결과, 제2 텐서(TENSOR2)에 포함된 논-제로 셀 최대 값은 16진수 "0E" (2진수 $00001110_{(2)}$) 이므로, 논-제로 비트 폭은 4일 수 있다.

[0061] 예시적인 실시예에서, 제2 텐서(TENSOR2) 중 제1 피처맵(FM1), 제2 피처맵(FM2) 및 제4 피처맵(FM4)의 셀 값은 모두 0이고, 제3 피처맵(FM3)만이 논-제로 셀을 가질 수 있다(쿼드 트리: $0010_{(2)}$). 제1 피처맵(FM1), 제2 피처맵(FM2) 및 제4 피처맵(FM4)은 도 3a와 유사하게 각각 1바이트 ($00000000_{(2)}$)로 압축될 수 있다.

[0062] 예시적인 실시예에서, 논-제로 셀을 포함하는 제3 피처맵(FM3)에 대해 쿼드 트리를 적용할 수 있다. 쿼드 트리 적용 결과, 제3 피처맵(FM3)은 가운데를 기준으로 4등분된 결과 상하좌우로 구획될 수 있고, 구획된 각각은 제1 셀 그룹(CG1), 제2 셀 그룹(CG2), 제3 셀 그룹(CG3) 및 제4 셀 그룹(CG4)일 수 있다. 제1 셀 그룹(CG1) 및 제2 셀 그룹(CG2)은 논-제로 셀이 존재하지 않고, 제3 셀 그룹(CG3) 및 제4 셀 그룹(CG4)은 논-제로 셀이 존재한다(쿼드 트리: $0011_{(2)}$). 논-제로 셀이 존재하지 않는 제1 셀 그룹(CG1) 및 제2 셀 그룹(CG2)은 압축은 종료되고, 논-제로 셀이 존재하는 제3 셀 그룹(CG3)(0E) 및 제4 셀 그룹(CG4)(06)에 대해 쿼드 트리가 한번 더 적용된다. 제3 셀 그룹(CG3)의 좌상단 셀만 논-제로이고(쿼드 트리: $1000_{(2)}$), 제4 셀 그룹(CG4)의 좌하단 셀만 논-제로이다(쿼드 트리: $0010_{(2)}$). 제3 셀 그룹(CG3) 및 제4 셀 그룹(CG4)의 논-제로 셀에 도달했기 때문에, 쿼드 트리 압축은 종료된다. 제3 셀 그룹(CG3)의 논-제로 값은 16진수 "0E" 이므로, 2진수로 변환하면 $1110_{(2)}$ 이다. 제4 셀 그룹(CG4)의 논-제로 값은 16진수 "06"이므로, 2진수로 변환하면 $0110_{(2)}$ 이다.

[0063] 본 개시의 기술적 사상에 따른 쿼드 트리 생성기(141)는 쿼드 트리 압축 결과, 복수의 파라미터를 가지는 쿼드 트리를 생성할 수 있다. 복수의 파라미터는 텐서를 쿼드 트리 기반 압축한 길이인 압축 길이(LENGTH), 텐서에 포함된 가장 큰 값을 가지는 논-제로 셀의 비트 폭인 논-제로 비트 폭(NZW), 텐서에 포함된 복수의 셀들 중 제로-셀의 개수인 제로 개수(ZC; Zero Count), 및 논-제로 셀 값인 논-제로 값(NZV), 및 쿼드 트리 압축이 수행된 결과 생성되는 비트들의 수인 쿼드 트리 비트 수(QTB)를 포함할 수 있으나, 이에 제한되지 않는다.

[0064] 쿼드 트리 생성기(141)는 텐서에 대해 쿼드 트리 기반 압축을 적용하면서, 적어도 하나의 파라미터들을 추출할 수 있다.

[0065] 예시적인 실시예에서, 쿼드 트리 생성기(141)는 쿼드 트리 압축 과정에서 생성된 비트들의 합이 33 비트임을 확인할수 있다. 33 비트는 4바이트 공간으로 표현될 수 없기 때문에, 쿼드 트리 생성기(141)는 제2 압축 길이(LENGTH2)를 5바이트라고 판단할 수 있다($000100_{(2)}$). 5바이트는 40 비트이므로, 33비트 외의 잔여 7비트는 제로-패딩될 수 있다.

[0066] 예시적인 실시예에서, 쿼드 트리 생성기(141)는 논-제로 버퍼에서 버퍼링된 셀에 기초하여, 제2 논-제로 비트 폭(NZW2)은 4비트라고 판단할 수 있다.

[0067] 예시적인 실시예에서, 쿼드 트리 생성기(141)는 쿼드 트리 압축 결과 4 비트가 4개 생성되었으므로, 쿼드 트리 비트 수(QTB)를 16이라고 결정할 수 있다. 쿼드 트리 비트 수(QTB)는 12~84의 범위에서 존재할 수 있다.

[0068] 예시적인 실시예에서, 쿼드 트리 생성기(141)는 제로 개수(ZC)를 추출할 수 있다. 제로 개수(ZC)는 텐서에 포함된 복수의 셀들 수에서 논-제로 버퍼에 버퍼링된 논-제로를 감산함으로써 추출될 수 있다.

[0069] 그러나, 논-제로 버퍼에 제한되지 않고, 제로 개수(ZC)는 쿼드 트리 압축 결과 생성된 비트들로부터 역산될 수도 있다. 예를 들어, 쿼드 트리 압축 결과 생성된 비트들 중 최상위 계층의 비트들은 "$0010_{(2)}$"이다. 최상위 계층의 비트들은 피처맵의 논-제로 여부와 상응하므로, "0"은 1개의 피처맵에 포함된 16개의 셀이 모두 제로 셀이라는 것을 판단할 수 있다. 즉, 3개의 "0"은 16X3=48개의 제로 개수를 의미할 수 있다. 마찬가지로, 두번째 계층의 비트들은 "$0011_{(2)}$"이고, "0"은 2개인데, 두번째 계층은 한번 쿼드 트리가 적용되었으므로, 셀 그룹 4개의 셀이 제로 셀임을 의미할 수 있다. 즉, 2개의 "0"은 제로 개수가 4X2=8개임을 의미할 수 있다. 마찬가지로,

세번째 계층의 "0"은 6개인데, 세번째 계층은 두번 쿼드 트리가 적용되었으므로, 6개의 "0"은 셀 6개가 제로 셀임을 의미할 수 있다. 결과적으로, 총 제로 개수(ZC)는 48+8+6=62임이 역산될 수 있다.

[0070] 도 4는 본 개시의 예시적 실시예에 따른 압축 모드를 결정하는 방법을 나타내는 흐름도이다. 도 2가 함께 참조된다.

[0071] 쿼드 트리 방법에 기반하여 텐서를 압축하는 경우, 적어도 하나의 파라미터들을 가지는 쿼드 트리가 생성될 수 있다. 쿼드 트리의 파라미터는 텐서에 포함된 가장 큰 값을 가지는 논-제로 셀의 비트 폭인 논-제로 비트 폭(NZW), 텐서에 포함된 복수의 셀들 중 제로-셀의 개수인 제로 개수(ZC), 및 쿼드 트리 압축이 수행된 결과 생성되는 비트들의 수인 쿼드 트리 비트 수(QTB)를 포함할 수 있으나, 이에 제한되지 않는다.

[0072] 도 4를 참조하면, 단계 S11에서, 모드 선택기(도 2, 143)는 쿼드 트리 비트 수(QTB)와 텐서에 포함된 복수의 셀들 수(N)를 비교할 수 있다(S11). 복수의 셀들 수(N)는 뉴럴 네트워크의 깊이 M에 의존적일 수 있고, $N=2^M$ 을만족할 수 있다.

[0073] 쿼드 트리 비트 수(QTB)가 복수의 셀들 수(N)보다 작거나 같은 경우, 쿼드 트리 방법에 기반하여 텐서가 압축될 수 있다(S12).

[0074] 단계 S13에서, 쿼드 트리 비트 수(QTB)가 복수의 셀들 수(N)보다 큰 경우, 모드 선택기(143)는 논-제로 비트 폭(NZW)과 제로 개수(ZC)의 곱을 복수의 셀들 수(N)와 비교할 수 있다(S13).

[0075] 논-제로 비트 폭(NZW)과 제로 개수(ZC)의 곱이 복수의 셀들 수(N)보다 크다면, 제로 비트맵(ZERO BITMAP) 방법에 기반하여 텐서가 압축될 수 있다(S14). 제로 비트맵 방법은 논-제로 셀을 '1'로, 제로 셀을 '0'으로 간주하여 논-제로 셀과 제로 셀의 위치 정보를 담은 프리픽스(Prefix) 테이블을 압축에 이용하는 방법이다.

[0076] 논-제로 비트 폭(NZW)과 제로 개수(ZC)의 곱이 복수의 셀들 수(N)보다 작거나 같다면, 고정 길이(FIXED LENGTH) 방법에 기반하여 텐서가 압축될 수 있다(S15). 고정 길이 방법은 프리픽스 테이블을 이용하지 않고, 복수의 셀들 중 가장 큰 셀 값을 가지는 셀의 비트 폭으로 셀들의 길이를 고정하는 방법이다.

[0077] 본 개시의 기술적 사상에 따라, 논-제로 비트 폭(NZW)과 제로 개수(ZC)의 곱을 복수의 셀들 수(N)와 비교하는 것이 필요하다. 다음의 수학식들이 참조된다.

수학식 1

[0078] $$(\mathrm{N} + \mathrm{NZW} \times \mathrm{NZC}) > (NZW \times N)$$

[0079] 좌항은 제로 비트맵 방법에 상응할 수 있고, 우항은 고정 길이 방법에 상응할 수 있다. 부등호의 방향은 고정 길이 방법의 압축 효율이 더 좋은 경우를 의미할 수 있다.

[0080] 상기 수학식 1을 다음과 같이 정리할 수 있다.

수학식 2

[0081] $$\mathrm{N} + \mathrm{NZW} \times (\mathrm{N} - \mathrm{ZC}) > NZW \times N$$

[0082] 상기 수학식 2를 다음과 같이 정리할 수 있다.

수학식 3

[0083] $$\mathrm{N} > NZW \times ZC$$

[0084] 결론적으로, 상기 수학식 3에 따르면, 모드 선택기(143)는 논-제로 비트 폭(NZW)과 제로 개수(ZC)의 곱과 복수의 셀들 수(N)를 비교하고, 복수의 셀들 수(N)가 더 크면 고정 길이 방법의 압축 효율이 더 좋은지 판단할 수 있다.

[0085] 쿼드 트리 방법, 제로 비트맵 방법 및 고정 길이 방법이 적용된 압축 결과 생성되는 비트스트림은 도 5에서 후술된다.

[0086] 도 5는 본 개시의 예시적 실시예에 따른 비트스트림의 구조를 도시하는 도면이다. 도 2가 함께 참조된다.

[0087] 도 5의 (a)를 참조하면, 비트스트림 생성기(145)는 제로 셀을 8 비트 크기를 가지는 비트스트림으로 생성할 수 있다. 하나의 셀은 16진수 2비트의 정보를 담고 있으므로, 하나의 셀은 8비트의 크기를 가진다. 제로 셀의 셀 값은 "0"이고, 10진수 0을 16진수로 변환하면 "0x00"으로 표현된다.

[0088] 도 5의 (b)를 참조하면, 비트스트림 생성기(145)는 압축되지 않은 텐서를 8X(N+1)의 크기를 가지는 비트스트림으로 생성할 수 있다. 텐서에는 N개의 셀들이 포함되어 있고, 셀 하나는 8비트의 크기를 가지므로 8N개의 셀에 대한 비압축 비트스트림이 생성될 수 있고, 추가적으로 셀의 최대값을 의미하는 8비트가 비트스트림 가장 앞에 위치할 수 있다. 예를 들어, 셀의 최대값이 "2" 라면, 16진수로 "0x02"로 표현될 수 있고, 2진수로 "$00000010_{(2)}$"가 비트스트림 가장 앞 8비트에 위치할 수 있다.

[0089] 도 5의 (c)를 참조하면, 비트스트림 생성기(145)는 모드 선택기(143)에 의해 선택된 쿼드 트리 방법에 상응하는 비트스트림을 생성할 수 있다. 예시적인 실시예에 따라, 도 5의 (c)의 비트스트림은 쿼드 트리 비트 수(QTB)가 복수의 셀들 수(N)보다 같거나 작은 경우에 생성될 수 있다.

[0090] 비트스트림 생성기(145)는 텐서에 포함된 복수의 셀들 수(N)에 대한 비트들을 비트스트림 가장 앞단 6비트에 위치시킬 수 있다. 예를 들어, N=64인 경우, 6비트 크기의 공간에 이진수 "$111111_{(2)}$" (="$63_{(10)}$")가 표현될 수 있다.

[0091] 비트스트림 생성기(145)는 그 다음 위치의 1비트 공간에 쿼드 트리 적용 여부를 확인할 수 있는 비트 "1" 을 기입할 수 있다.

[0092] 비트스트림 생성기(145)는 그 다음 위치에 논-제로 비트 폭(NZW)에 대한 비트들을 위치시킬 수 있다. 셀은 8비트 정보를 가지기 때문에, 최대 논-제로 비트 폭(NZW) 역시 8비트일 수 있다. 논 제로 비트 폭(NZW)이 1일 경우, 비트로서 "$000_{(2)}$" (=0=NZW-1)이 비트 공간에 기입될 수 있고, 논 제로 비트 폭(NZW)이 8일 경우, 비트로서 "$111_{(2)}$"(=7=NZW-1)이 기입될 수 있다.

[0093] 비트스트림 생성기(145)는 그 다음 위치의 비트 공간에 쿼드 트리 비트 수(QTB)에 대한 비트들을 기입할 수 있다. 예시적인 실시예에서 N=64일 경우, 쿼드 트리 압축 결과 생성되는 비트들의 총 수 인 쿼드 트리 비트 수(QTB)는 12~84 범위 내에 분포할 수 있는데, 도 5의 (c)는 모드 선택기(143)에 의해 복수의 셀들 수(N=64)보다 작거나 같은 경우에 쿼드 트리 압축이 수행되도록 결정되었기 때문에, 쿼드 트리 비트 수(QTB)는 12~64의 분포 범위를 가질 수 있다.

[0094] 비트스트림 생성기(145)는 그 다음 위치의 비트 공간에 k 개의 논-제로 값(NZV)들에 대한 비트들을 기입할 수 있다. 논-제로 비트 폭(NZW)은 셀 중 가장 큰 값을 가지는 셀이므로, 논-제로 값(NZV)들 각각은 모두 논-제로 비트 폭(NZW)에 상응하는 비트 수만으로도 논-제로 값(NZV)을 표현할 수 있다.

[0095] 비트스트림 생성기(145)는 그 다음 위치의 비트 공간에 바이트 단위를 맞추기 위한 제로-패딩을 수행할 수 있다. 바이트 단위를 맞추기 위한 패딩이므로, 제로-패딩될 수 있는 비트 수는 0에서 7 사이이다.

[0096] 도 5의 (d)를 참조하면, 비트스트림 생성기(145)는 모드 선택기(143)에 의해 선택된 제로 비트맵 방법에 상응하는 비트스트림을 생성할 수 있다. 예시적인 실시예에 따라, 도 5의 (d)의 비트스트림은 쿼드 트리 비트 수(QTB) 및 논-제로 비트 폭과 제로 개수의 곱(NZWXZC)이 모두 복수의 셀들 수(N)보다 큰 경우에 생성될 수 있다. 도 5의 (c)와 중복되는 비트 공간에 대한 설명은 생략한다.

[0097] 도 5의 (d)가 도시하는 비트스트림은 도 5의 (c)가 도시하는 비트스트림와 비교해서 쿼드 트리 비트 수(QTB) 대신 프리픽스 테이블에 대한 비트가 기입된 것, 및 가변 길이 적용 여부를 확인할 수 있는 비트가 기입된 것이 비트스트림 길이의 차이를 유발한다. 그 외에, 도 5의 (d)가 도시하는 비트스트림은 도 5의 (c)의 비트스트림에 비해 쿼드 트리 적용 여부 확인 비트가 "0"이 기입된 차이가 있다. 프리픽스 테이블은 논-제로 셀을 '1'로, 제로 셀을 '0'으로 간주하여 논-제로 셀과 제로 셀의 위치 정보를 담은 64 비트의 정보이다.

[0098] 예시적인 실시예에 따르면, 쿼드 트리 비트 수(QTB)가 64 비트를 초과한다면, 쿼드 트리 비트 수(QTB)는 최대 84비트일 수 있기 때문에, 항상 64비트를 가지는 프리픽스 테이블을 사용하는 제로 비트맵 방법의 압축 효율이

더 우수할 수 있다.

[0099] 도 5의 (e)를 참조하면, 비트스트림 생성기(145)는 모드 선택기(143)에 의해 선택된 고정 길이(fixed length) 방법에 상응하는 비트스트림을 생성할 수 있다. 예시적인 실시예에 따라, 도 5의 (e)의 비트스트림은, 쿼드 트리 비트 수(QTB) 는 복수의 셀들 수(N)보다 크되, 논-제로 비트 폭과 제로 개수의 곱(NZW X ZC)은 복수의 셀들 수(N)보다 작거나 같은 경우에 생성될 수 있다. 도 5의 (c), (d)와 중복되는 비트 공간에 대한 설명은 생략한다.

[0100] 도 5의 (e)가 도시하는 비트스트림은 도 5의 (d)가 도시하는 비트스트림와 비교해서 프리픽스 테이블 및 N 개의 논-제로 값(NZV)을 이용하지 않고, 고정된 길이의 값을 이용하는 것, 및 가변 길이 적용 여부를 확인할 수 있는 비트가 "1"인 것에 차이가 있다.

[0101] 예시적인 실시예에 따르면 N=64일 때, 대부분이 논-제로 셀인 경우, 위치 정보를 줄 필요가 없으므로, 64개의 셀 중 가장 큰 값만을 논-제로 비트 폭(NZW)으로 결정하고, 고정 길이 압축을 수행할 수 있다. 논-제로 비트 폭과 제로 개수의 곱(NZW X ZC)은 복수의 셀들 수(N)보다 작거나 같다면, 프리픽스 테이블이 가지는 64비트는 오버헤드일 수 있기 때문에, 고정 길이 방식이 더 유리할 수 있다.

[0102] 도 6은 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

[0103] 단계 S110에서, 뉴럴 텐서 압축기(140)는 피처맵과 웨이트에 대한 사칙 연산이 반복된 결과인 텐서를 수신할 수 있다.

[0104] 단계 S120에서, 뉴럴 텐서 압축기(140)는 복수의 셀 중 제로 셀을 압축하기 위해 텐서를 반복 공간 분할한 쿼드 트리 압축 결과, 적어도 하나의 파라미터를 추출할 수 있다.

[0105] 단계 S130에서, 뉴럴 텐서 압축기(140)는 적어도 하나의 파라미터에 기초하여 압축 모드를 결정할 수 있다.

[0106] 단계 S140에서, 뉴럴 텐서 압축기(140)는 압축 모드에 기초하여 비트스트림을 출력할 수 있다.

[0107] 도 7은 본 개시의 예시적 실시예에 따른 압축기의 동작 방법을 나타내는 흐름도이다.

[0108] 단계 S110 후, 단계 S121에서, 뉴럴 텐서 압축기(140)는 쿼드 트리 압축 결과로서 생성되는 쿼드 트리 비트 수(QTB)를 추출할 수 있다.

[0109] 단계 S122에서, 뉴럴 텐서 압축기(140)는 복수의 셀들 중 셀 값이 가장 큰 논-제로 셀의 비트 폭에 상응하는 논제로 비트 폭(NZW) 추출

[0110] 단계 S123에서, 뉴럴 텐서 압축기(140)는 복수의 셀들 중 셀 값이 제로인 셀들의 수에 상응하는 제로 개수(ZC)를 추출할 수 있다. 그 후, 단계 S130로 이동한다.

[0111] 도 8은 뉴럴 네트워크 구조의 일 예로서 컨볼루션 뉴럴 네트워크 구조를 설명하기 위한 도면이고, 도 9는 본 개시의 예시적 실시예에 따른 뉴럴 네트워크의 컨볼루션 연산을 설명하기 위한 도면이다.

[0112] 도 8을 참조하면, 뉴럴 네트워크(NN)는 복수의 레이어들(L1 내지 Ln)을 포함할 수 있다. 복수의 레이어들(L1 내지 Ln) 각각은 선형 레이어 또는 비선형 레이어일 수 있으며, 일 실시예에 있어서, 적어도 하나의 선형 레이어 및 적어도 하나의 비선형 레이어가 결합되어 하나의 레이어로 지칭될 수도 있다. 예시적으로, 선형 레이어는 컨볼루션 레이어(convolution layer) 및 풀리 커넥티드 레이어(fully connected layer)를 포함할 수 있으며, 비선형 레이어는 풀링(pooling layer) 및 활성 레이어(activation layer)를 포함할 수 있다.

[0113] 예시적으로, 제1 레이어(L1)는 컨볼루션 레이어이고, 제2 레이어(L2)는 풀링 레이어이고, 제n 레이어(Ln)는 출력 레이어로서 풀리 커넥티드 레이어일 수 있다. 뉴럴 네트워크(NN)는 활성 레이어를 더 포함할 수 있으며, 다른 종류의 연산을 수행하는 레이어를 더 포함할 수 있다.

[0114] 복수의 레이어들(L1 내지 Ln) 각각은 입력되는 데이터(예컨대, 이미지 프레임) 또는 이전 레이어에서 생성된 피처맵을 입력 피처맵으로서 수신하고, 입력 피처맵을 연산함으로써 출력 피처맵 또는 인식 신호(REC)를 생성할 수 있다. 이 때, 피처맵은 입력 데이터의 다양한 특징이 표현된 데이터를 의미한다. 피처맵들(FM1, FM2, FMn)은 예컨대 2차원 매트릭스 또는 3차원 매트릭스(또는 텐서(tensor)) 형태를 가질 수 있다. 피처맵들(FM1, FM2, FMn)은 너비(W)(또는 칼럼), 높이(H)(또는 로우) 및 깊이(D)를 가지며, 이는 좌표상의 x축, y축 및 z축에 각각 대응될 수 있다. 이 때, 깊이(D)는 채널 수로 지칭될 수 있다.

[0115] 제1 레이어(L1)는 제1 피처맵(FM1)을 웨이트 맵(WM)과 컨볼루션함으로써 제2 피처맵(FM2)을 생성할 수 있다. 웨

이트 맵(WM)은 제1 피처맵(FM1)을 필터링할 수 있으며, 필터 또는 커널로도 지칭될 수 있다. 웨이트 맵(WM)의 깊이, 즉 채널 개수는 제1 피처맵(FM1)의 깊이, 즉 채널 개수와 동일하며, 웨이트 맵(WM)과 제1 피처맵(FM1)의 동일한 채널끼리 컨볼루션 될 수 있다. 웨이트 맵(WM)이 제1 피처맵(FM1)을 슬라이딩 윈도우로 하여 횡단하는 방식으로 시프트 될 수 있다. 시프트되는 양은 "스트라이드(stride) 길이" 또는 "스트라이드"로 지칭될 수 있다. 각 시프트 동안, 웨이트 맵(WM)에 포함되는 웨이트 값들 각각이 제1 피처맵(FM1)과 중첩되는 영역에서의 모든 셀 데이터들과 곱해지고 더해질 수 있다. 웨이트 맵(WM)에 포함되는 웨이트 값들 각각이 제1 피처맵(FM1)과 중첩되는 영역에서의 제1 피처맵(FM1)의 데이터들을 추출 데이터라 칭할 수 있다. 제1 피처맵(FM1)과 웨이트 맵(WM)이 컨볼루션 됨에 따라, 제2 피처맵(FM2)의 하나의 채널이 생성될 수 있다. 도 3에는 하나의 웨이트 맵(WM)이 표시되었으나, 실질적으로는 복수의 웨이트 맵들이 제1 피처맵(FM1)과 컨볼루션 되어, 제2 피처맵(FM2)의 복수의 채널들이 생성될 수 있다. 다시 말해, 제2 피처맵(FM2)의 채널의 수는 웨이트 맵의 개수에 대응될 수 있다.

[0116] 제2 레이어(L2)는 풀링을 통해 제2 피처맵(FM2)의 공간적 크기(spatial size)를 변경함으로써, 제3 피처맵(FM3)을 생성할 수 있다. 풀링은 샘플링 또는 다운-샘플링으로 지칭될 수 있다. 2차원의 풀링 윈도우(PW)가 풀링 윈도우(PW)의 사이즈 단위로 제2 피처맵(FM2) 상에서 시프트 되고, 풀링 윈도우(PW)와 중첩되는 영역의 셀 데이터들 중 최대값(또는 셀 데이터들의 평균값)이 선택될 수 있다. 이에 따라, 제2 피처맵(FM2)으로부터 공간적 사이즈가 변경된 제3 피처맵(FM3)이 생성될 수 있다. 제3 피처맵(FM3)의 채널과 제2 피처맵(FM2)의 채널 개수는 동일하다.

[0117] 제n 레이어(Ln)는 제n 피처맵(FMn)의 피처들을 조합함으로써 입력 데이터의 클래스(class)(CL)를 분류할 수 있다. 또한, 제n 레이어(Ln)는 클래스에 대응되는 인식 신호(REC)를 생성할 수 있다. 실시예에 있어서, 입력 데이터는 비디오 스트림(video stream)에 포함되는 프레임 데이터에 대응될 수 있으며, 제n 레이어(Ln)는 이전 레이어로부터 제공되는 제n 피처맵(FMn)을 기초로 프레임 데이터가 나타내는 이미지에 포함되는 사물에 해당하는 클래스를 추출함으로써, 사물을 인식하고, 인식된 사물에 상응하는 인식 신호(REC)를 생성할 수 있다.

[0118] 도 9를 참조하면, 입력 피처맵들(201)은 D개의 채널들을 포함하고, 각 채널의 입력 피처맵은 H행 W열의 크기를 가질 수 있다(D, H, W는 자연수). 커널들(202) 각각은 R행 S열의 크기를 갖고, 커널들(202)은 입력 피처맵들(201)의 채널 수(또는 깊이)(D) 에 대응되는 개수의 채널들을 포함할 수 있다(R, S는 자연수). 출력 피처맵들(203)은 입력 피처맵들(201)과 커널들(202) 간의 3차원 컨볼루션 연산을 통해 생성될 수 있고, 컨볼루션 연산에 따라 Y개의 채널들을 포함할 수 있다.

[0119] 하나의 입력 피처맵과 하나의 커널 간의 컨볼루션 연산을 통해 출력 피처맵이 생성되는 과정은 도 4b를 참조해 설명될 수 있으며, 도 4b에서 설명되는 2차원 컨볼루션 연산이 전체 채널들의 입력 피처맵들(201)과 전체 채널들의 커널들(202) 간에 수행됨으로써, 전체 채널들의 출력 피처맵들(203)이 생성될 수 있다.

[0120] 도 10은 본 개시의 예시적 실시예에 따른 양자화기(247)를 더 포함하는 뉴럴 텐서 압축기(240)의 블록도이다.

[0121] 뉴럴 텐서 압축기(240)는 쿼드 트리 생성기(241), 모드 선택기(243), 비트스트림 생성기(245) 및 양자화기(247)를 포함할 수 있다. 양자화기(247)는 제공받은 텐서(TENSOR)를 양자화하고, 양자화된 텐서(TENSOR_Q)를 쿼드 트리 생성기(241)에 제공할 수 있다. 쿼드 트리 생성기(241)는 양자화된 텐서(TENSOR_Q)에 대해 쿼드 트리 방법을 적용해 쿼드 트리를 생성할 수 있다. 도 10의 쿼드 트리 생성기(241), 모드 선택기(243), 및 비트스트림 생성기(245)는 도 2의 쿼드 트리 생성기(141), 모드 선택기(143), 및 비트스트림 생성기(145)에 서로 대응되는 바, 중복되는 설명은 생략된다.

[0122] 뉴럴 네트워크의 컨볼루션 연산에 ReLU(Rectified Linear Unit) 연산이 적용되는 경우가 많기 때문에, 피처맵 및 피처맵을 포함하는 텐서에 제로 셀이 많이 분포할 수 있다. 제로 셀이 많이 존재한다면, 상대적으로 텐서에 포함된 셀 값의 대부분은 0 근처에 모여있다고 추측될 수 있다.

[0123] 본 개시의 기술적 사상에 따르면, 양자화기(247)는 비균일 양자화를 수행할 수 있다. 예시적인 실시예에서, 양자화기(247)는 텐서에 제로 셀이 많은 특성을 이용하여 비균일적 양자화를 수행할 수 있다.

[0124] 비균일적 양자화에는, 다음과 같은 수학식 4가 적용된다.

수학식 4

[0125] $$Q_{out} = \frac{[input + Qstep/2]}{Qstep} + offset$$

[0126] 수학식 4를 참조하면, 비균일적 양자화된 값은, 입력값(input)에 양자화 단계(Qstep)에 2를 나눈 값을 더한 후, 소숫점 버림한 값에 대해 양자화 단계(Qstep)를 나누고, 오프셋을 더한 값과 상응할 수 있다. 그러나 이에 제한되지 않고, 텐서의 특성을 고려한 다양한 비균일 양자화 방법이 적용될 수 있다.

[0127] 본 개시의 기술적 사상에 따른 양자화기(247)는 선택적으로 양자화를 수행할 수 있다. 예시적인 실시예에서, 뉴럴 네트워크 프로세서의 정확도가 고도로 요구될 때, 양자화기(247)는 양자화를 수행하지 않고 텐서(TENSOR)를 그대로 쿼드 트리 생성기(241)에 전달할 수 있다. 예를 들어, 양자화가 수행되지 않을 때, 양자화된 텐서(TENSOR_Q)는 텐서(TENSOR)와 동일할 수 있다. 예시적인 실시예에서, 저전력 모드로 작동하는 뉴럴 네트워크 프로세서거나, 상대적으로 저렴한 전자 기기에 실장되는 뉴럴 네트워크 프로세서의 경우, 양자화기(247)는 비균일 양자화를 수행할 수 있다.

[0128] 균일적 양자화를 수행하면, 상대적으로 정확도에 민감한 영향을 미칠 수 있는 낮은 셀 값에 데이터 손실이 발생할 수 있다. 본 개시의 기술적 사상에 따른 뉴럴 텐서 압축기(240)는 비균일적 양자화를 수행할 때, 데이터 처리 정확도에 관련있는 낮은 셀 값을 가지는 셀들에 대해서는 양자화를 수행하지 않고, 상대적으로 정확도와 관련이 적은 높은 셀값을 가지는 셀들에 대해 양자화를 수행함으로써, 데이터 손실을 최소화하며 데이터 압축률을 극대화할 수 있다.

[0129] 도 11은 본 개시의 예시적 실시예에 따른 셀 값에 따른 셀 분포를 도시하는 그래프이다. 그래프의 가로축은 셀 값을, 세로축은 셀 개수를 의미할 수 있다.

[0130] 도 11을 참조하면, 텐서에 포함된 복수의 셀들은 각각 최대 255의 셀 값을 가질 수 있다고 가정한다. 양자화 단계(Qstep)는 셀 최대값인 255을 4로 나누어 범주화할 수 있다. 예를 들어, 양자화 범위는 제1 범위(셀값 0~63)에 대한 양자화 단계 1(Qstep=1) 및 오프셋 0, 제2 범위(셀값 64~127)에 대한 양자화 단계 2(Qstep=2) 및 오프셋 32 및 제3 범위(셀값 128~255)에 대한 양자화 단계 4(Qstep=4) 및 오프셋 64로 구분될 수 있다. 그러나 이에 제한되지는 않는다.

[0131] 예시적인 실시예에 따르면, 제1 범위에 대해, 양자화 단계는 1(Qstep=1)이고, 오프셋은 0이므로, 비균일적 양자화된 값은 0~63의 값을 가질 수 있다.

[0132] 예시적인 실시예에 따르면, 제2 범위에 대해, 양자화 단계는 2(Qstep=2)이고, 오프셋은 32이므로, 비균일적 양자화된 값은 64~95의 값을 가질 수 있다.

[0133] 예시적인 실시예에 따르면, 제3 범위에 대해, 양자화 단계는 4(Qstep=4)이고, 오프셋은 64이므로, 비균일적 양자화된 값은 96~127의 값을 가질 수 있다.

[0134] 본 개시의 기술적 사상에 따른 양자화기(도 10, 247)는 0부터 255 사의의 값을 가지는 8비트 셀 값을, 0부터 127 사이의 값을 가지는 7비트 셀 값으로 양자화할 수 있다. 비균일적 양자화는 다수의 셀이 낮은 셀 값을 가지는 범위에 대해 양자화를 생략하였으므로 균일적 양자화에 비해 데이터 손실이 최소화될 수 있다.

[0135] 도 12는 본 개시의 예시적 실시예에 따른 양자화기의 동작 방법을 나타내는 흐름도이다. 도 10이 함께 참조된다.

[0136] 단계 S21에서, 양자화기(247)는 텐서(TENSOR)를 수신할 수 있다. 텐서(TENSOR)는 8비트 크기의 셀을 64개 포함할 수 있다.

[0137] 단계 S22에서, 양자화기(247)는 8비트 크기의 셀 값의 범위를 나눌 수 있다. 양자화기(247)는 셀 값이 64 미만인지 확인할 수 있다(S22). 셀 값이 64 미만이면, 셀은 양자화되지 않고 쿼드 트리 생성기(241)로 제공될 수 있다. 셀 값이 64 이상이면 단계 S23으로 이동한다.

[0138] 단계 S23에서, 양자화기는 셀 값이 128 이상인지 확인할 수 있다(S23).

[0139] 단계 S25에서, 셀 값이 128 미만이라면, 양자화기(247)는 셀 값을 2로 나누고(S24), 오프셋 32를 셀 값에 합산

할 수 있다(S25).

[0140] 셀 값이 128 이상이라면, 양자화기(247)는 셀 값을 4로 나누고(S26), 오프셋 64를 셀 값에 합산할 수 있다(S27).

[0141] 도 13은 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다. 도 10이 함께 참조된다.

[0142] 단계 S210에서, 뉴럴 텐서 압축기는 뉴럴 네트워크를 이용한 피처맵과 웨이트에 대한 연산 결과, 복수 개의 셀들을 포함하는 텐서를 수신할 수 있다.

[0143] 단계 S220에서, 뉴럴 텐서 압축기는 복수의 셀들 중 최대 셀 값에 기초하여 텐서의 양자화 범위를 설정할 수 있다.

[0144] 단계 S230에서, 뉴럴 텐서 압축기는 상기 양자화 범위에 기초하여 상기 텐서를 선택적으로 양자화할 수 있다. 예시적인 실시예에서, 뉴럴 텐서 압축기는 제1 범위에 포함된 셀을 양자화하지 않고, 제2 범위에 포함된 셀을 양자화할 수 있다. 예를 들어, 제1 범위는 낮은 셀 값 영역일 수 있고, 제2 범위는 상대적으로 높은 셀 값 영역일 수 있으나, 이에 제한되지 않는다.

[0145] 단계 S240에서, 뉴럴 텐서 압축기는 양자화된 텐서(TENSOR_Q)에 대해 쿼드 트리 데이터 구조 적용함으로써 복수의 파라미터들 추출할 수 있다.

[0146] 단계 S250에서, 뉴럴 텐서 압축기는 복수의 파라미터들(NZW, ZC, QTB)에 기초하여 쿼트 트리 기반 비트스트림의 생성 여부를 결정할 수 있다.

[0147] 도 14는 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

[0148] 단계 S210 후, 단계 S221에서, 양자화기(247)는 최대 셀 값을 4로 나눈 값을 소수점 내림한 제1 값을 계산할 수 있다.

[0149] 단계 S222에서, 양자화기(247)는 최대 셀 값을 2로 나눈 값을 소수점 내림한 제2 값을 계산할 수 있다.

[0150] 단계 S223에서, 양자화기(247)는 0 ≤ 제1 양자화 범위 ≤ 제1 값에 해당하는 제1 양자화 범위를 설정할 수 있다.

[0151] 단계 S224에서, 양자화기(247)는 제1 값 < 제2 양자화 범위 ≤ 제2 값에 해당하는 제2양자화 범위를 설정할 수 있다.

[0152] 단계 S225에서, 양자화기(247)는 제2 값 < 제3 양자화 범위 ≤ 최대 셀 값에 해당하는 제3 양자화 범위를 설정할 수 있다.

[0153] 단계 S226에서, 양자화기(247)는 각각의 셀 값에 상응하는 양자화 범위로 복수 개의 셀들을 분류할 수 있다.

[0154] 그 후, 단계 S230으로 이동한다.

[0155] 도 15는 본 개시의 예시적 실시예에 따른 뉴럴 텐서 압축기의 동작 방법을 나타내는 흐름도이다.

[0156] 단계 S220 후, 단계 S231에서, 양자화기(247)는 제2 양자화 범위에 포함되는 셀의 셀 값을 2로 나누고, 상기 제1 양자화 범위와의 오버랩을 방지하기 위한 제1 오프셋을 합산할 수 있다.

[0157] 단계 S232에서, 양자화기(247)는 제3 양자화 범위에 포함되는 셀의 셀 값을 4로 나누고, 오버랩을 방지하기 위한 제2 오프셋을 합산 할 수 있다.

[0158] 그 후, 단계 S240으로 이동한다.

[0159] 도 16은 본 개시의 예시적 실시예에 따른 전자 시스템을 나타내는 블록도이다.

[0160] 도 16을 참조하면, 전자 시스템(1000)은 뉴럴 네트워크를 기초로 입력 데이터를 실시간으로 분석하여 유효한 정보를 추출하고, 추출된 정보를 기초로 상황을 판단하거나 전자 시스템(1000)이 탑재되는 전자 장치의 구성들을 제어할 수 있다. 예를 들어, 전자 시스템(1000)은 드론(drone), 첨단 운전자 보조 시스템(Advanced Drivers Assistance System; ADAS), 로봇 장치, 스마트 TV, 스마트 폰, 의료 장치, 모바일 장치, 영상 표시 장치, 계측 장치, IoT(Internet of Things) 장치 등에 적용될 수 있으며, 이외에도 다양한 종류의 전자 장치 중 하나에 탑재될 수 있다.

[0161] 전자 시스템(1000)은 도시된 기능부 외에도 다양한 종류의 IP 블록들을 포함할 수 있다. 예를 들어, IP 블록들은 프로세싱 유닛(processing unit), 프로세싱 유닛에 포함된 복수의 코어들(cores), MFC(Multi-Format Codec), 비디오 모듈(예컨대, 카메라 인터페이스, JPEG(Joint Photographic Experts Group) 프로세서, 비디오 프로세서, 또는 믹서 등), 3D 그래픽 코어, 오디오 시스템, 드라이버, 디스플레이 드라이버, 휘발성 메모리, 비휘발성 메모리(non-volatile memory), 메모리 컨트롤러(memory controller), 입출력 인터페이스 블록(input and output interface block), 또는 캐시 메모리(cache memory) 등을 포함할 수 있다.

[0162] IP들을 연결하기 위한 기술에는 시스템 버스(System Bus)를 기반으로 한 연결 방식이 있다. 예를 들어, 표준 버스 규격으로서, ARM(Advanced RISC Machine) 사의 AMBA(Advanced Microcontroller Bus Architecture) 프로토콜이 적용될 수 있다. AMBA 프로토콜의 버스 타입에는 AHB(Advanced High-Performance Bus), APB(Advanced Peripheral Bus), AXI(Advanced eXtensible Interface), AXI4, ACE(AXI Coherency Extensions) 등이 포함될 수 있다. 전술된 버스 타입들 중 AXI는 IP들 사이의 인터페이스 프로토콜로서, 다중 아웃스탠딩 어드레스(multiple outstanding address) 기능과 데이터 인터리빙(data interleaving) 기능 등을 제공할 수 있다. 이외에도, 소닉사(SONICs Inc.)의 uNetwork 나 IBM사의 CoreConnect, OCP-IP의 오픈 코어 프로토콜(Open Core Protocol) 등 다른 타입의 프로토콜이 시스템 버스에 적용되어도 무방할 것이다.

[0163] 뉴럴 네트워크 프로세서(또는, NPU)(1100)는 시스템 버스를 통해 다양한 종류의 입력 데이터를 수신할 수 있고, 입력 데이터를 기초로 정보 신호를 생성할 수 있다. 예를 들어, NPU(1100)는 입력 데이터에 뉴럴 네트워크 연산을 수행함으로써 정보 신호를 생성해낼 수 있으며, 뉴럴 네트워크 연산은 컨볼루션 연산을 포함할 수 있다.

[0164] 메모리(1300)는 데이터를 저장하기 위한 저장 장소로서, 예를 들어, OS(Operating System), 각종 프로그램들 및 각종 데이터를 저장할 수 있다. 메모리(1300)는 DRAM일 수 있으나, 이에 한정되는 것은 아니다. 메모리(1300)는 휘발성 메모리(volatile memory)를 포함할 수 있다. 휘발성 메모리는 DRAM(Dynamic RAM), SRAM(Static RAM), SDRAM(Synchronous DRAM), PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM), FeRAM(Ferroelectric RAM) 등을 포함할 수 있다.

[0165] CPU(1500)는 전자 시스템(1000)의 전반적인 동작을 제어할 수 있으며, 일 예로서 CPU(1500)는 중앙 프로세싱 유닛(Central Processing Unit; CPU)일 수 있다. CPU(1500)는 하나의 프로세서 코어(Single Core)를 포함하거나, 복수의 프로세서 코어들(Multi-Core)을 포함할 수 있다. CPU(1500)는 메모리(1300)에 저장된 프로그램들 및/또는 데이터를 처리 또는 실행할 수 있다. 예를 들어, CPU(1500)는 메모리(1300)에 저장된 프로그램들을 실행함으로써 전자 시스템(1000)의 기능들을 제어할 수 있다.

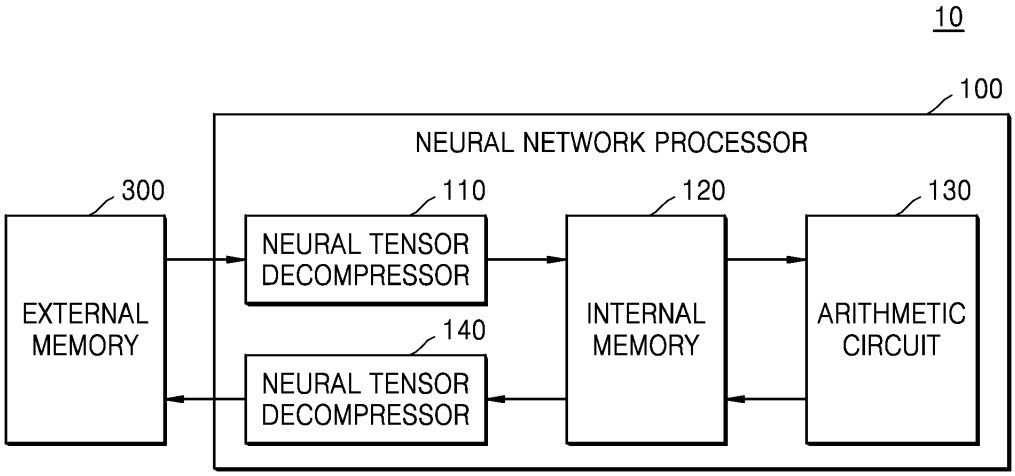
[0166] 스토리지(1700)는 데이터를 저장하기 위한 저장 장소로서, 각종 프로그램들 및 각종 데이터를 저장할 수 있다. 스토리지(1700)는 비휘발성 메모리(non-volatile memory)를 포함할 수 있다. 비휘발성 메모리는 ROM(Read Only Memory), PROM(Programmable ROM), EPROM(Electrically Programmable ROM), EEPROM(Electrically Erasable and Programmable ROM), 플래시 메모리, PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM), FRAM(Ferroelectric RAM) 등을 포함할 수 있다. 또한 일 실시예에 있어서, 스토리지(1700)는 HDD(Hard Disk Drive), SSD(Solid State Drive), CF(Compact Flash), SD(Secure Digital), Micro-SD(Micro Secure Digital), Mini-SD(Mini Secure Digital), xD(extreme digital) 또는 Memory Stick 중 적어도 하나를 포함할 수도 있다.

[0167] 센서(1900)는 전자 시스템(1000) 주변의 정보를 수집할 수 있다. 센서(1900)는 전자 시스템(1000) 외부로부터 이미지 신호를 센싱 또는 수신할 수 있고, 센싱 또는 수신된 이미지 신호를 이미지 데이터, 즉 이미지 프레임으로 변환할 수 있다. 이를 위해, 센서(1900)는 센싱 장치, 예컨대 촬상 장치, 이미지 센서, 라이더(LIDAR; light detection and ranging) 센서, 초음파 센서, 적외선 센서 등 다양한 종류의 센싱 장치들 중 적어도 하나를 포함하거나, 또는 상기 장치로부터 센싱 신호를 수신할 수 있다. 일 실시예에서, 센서(1900)는 이미지 프레임을 뉴럴 네트워크 프로세서(1100)에 제공할 수 있다. 예를 들어, 센서(1900)는 이미지 센서를 포함할 수 있으며, 전자 시스템(1000)의 외부 환경을 촬영함으로써 비디오 스트림을 생성하고, 비디오 스트림의 연속되는 이미지 프레임들을 뉴럴 네트워크 프로세서(1100)에 순차적으로 제공할 수 있다.

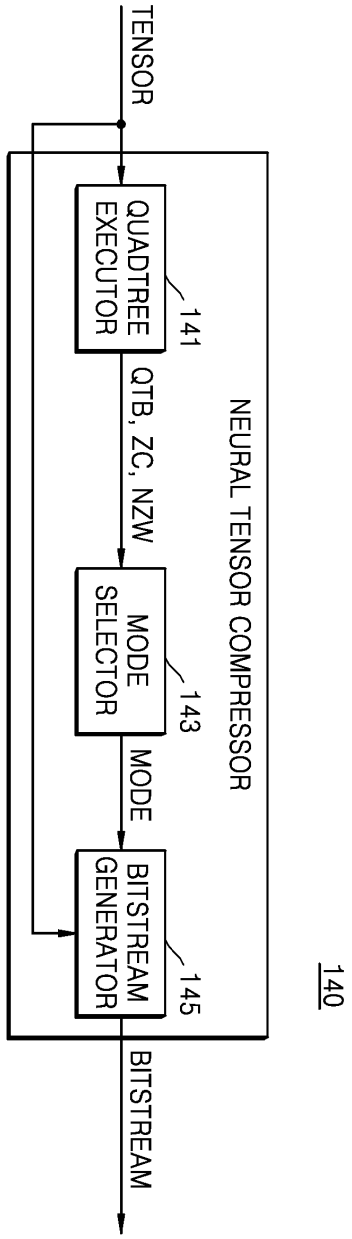
[0168] 이상에서와 같이 도면과 명세서에서 예시적인 실시예들이 개시되었다. 본 명세서에서 특정한 용어를 사용하여 실시예들을 설명되었으나, 이는 단지 본 개시의 기술적 사상을 설명하기 위한 목적에서 사용된 것이지 의미 한정이나 특허청구범위에 기재된 본 개시의 범위를 제한하기 위하여 사용된 것은 아니다. 그러므로 본 기술분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 따라서, 본 개시의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

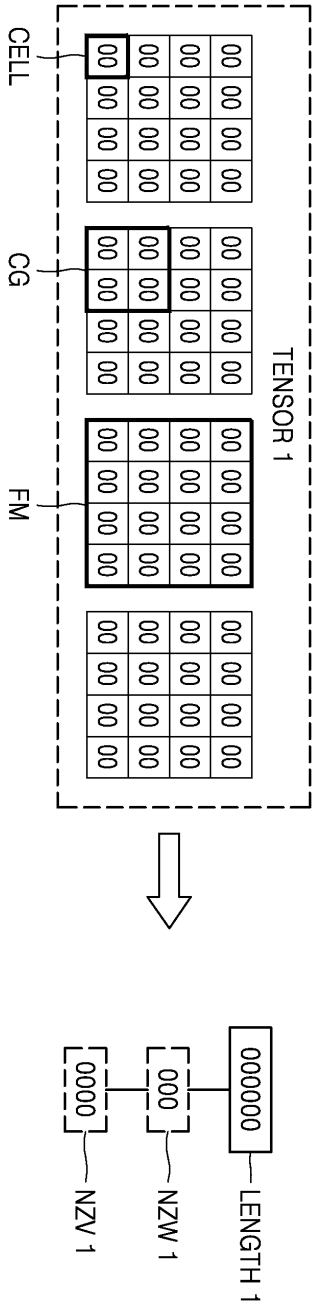
도면1


10
100
NEURAL NETWORK PROCESSOR
300
EXTERNAL MEMORY
110
NEURAL TENSOR DECOMPRESSOR
140
NEURAL TENSOR DECOMPRESSOR
120
INTERNAL MEMORY
130
ARITHMETIC CIRCUIT

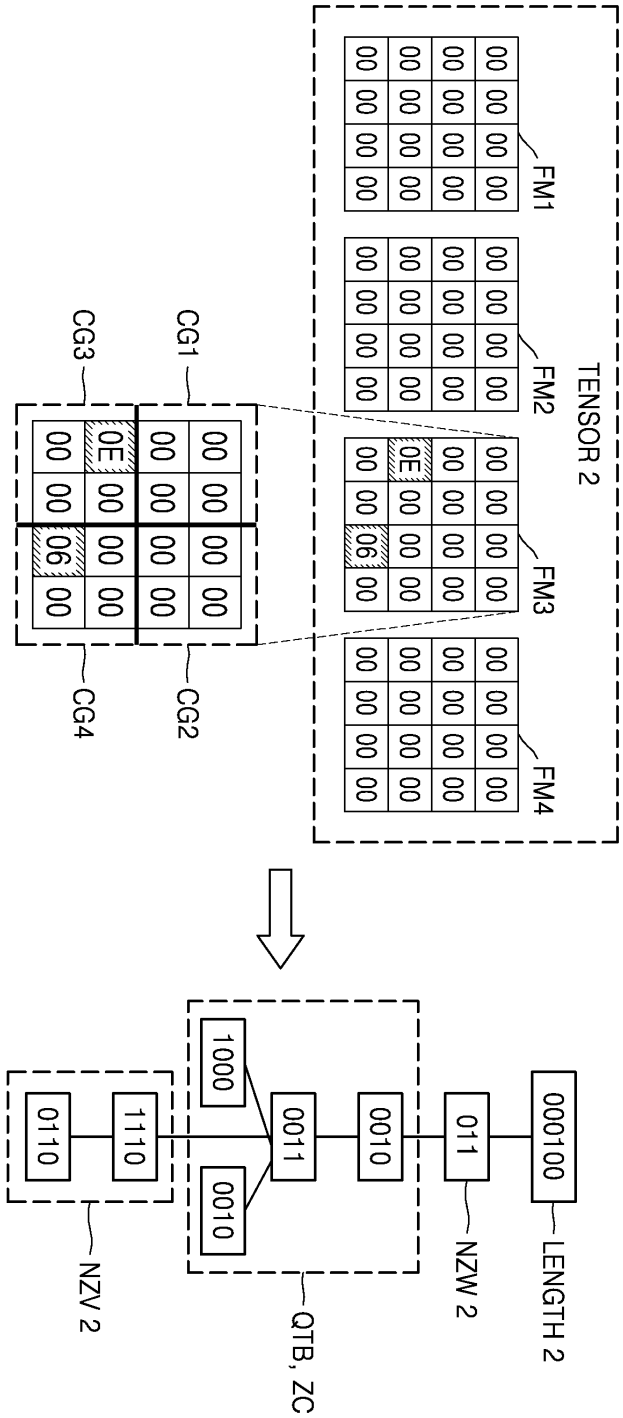
도면2


TENSOR
NEURAL TENSOR COMPRESSOR
QUADTREE EXECUTOR
141
QTB, ZC, NZW
MODE SELECTOR
143
MODE
BITSTREAM GENERATOR
145
BITSTREAM
140

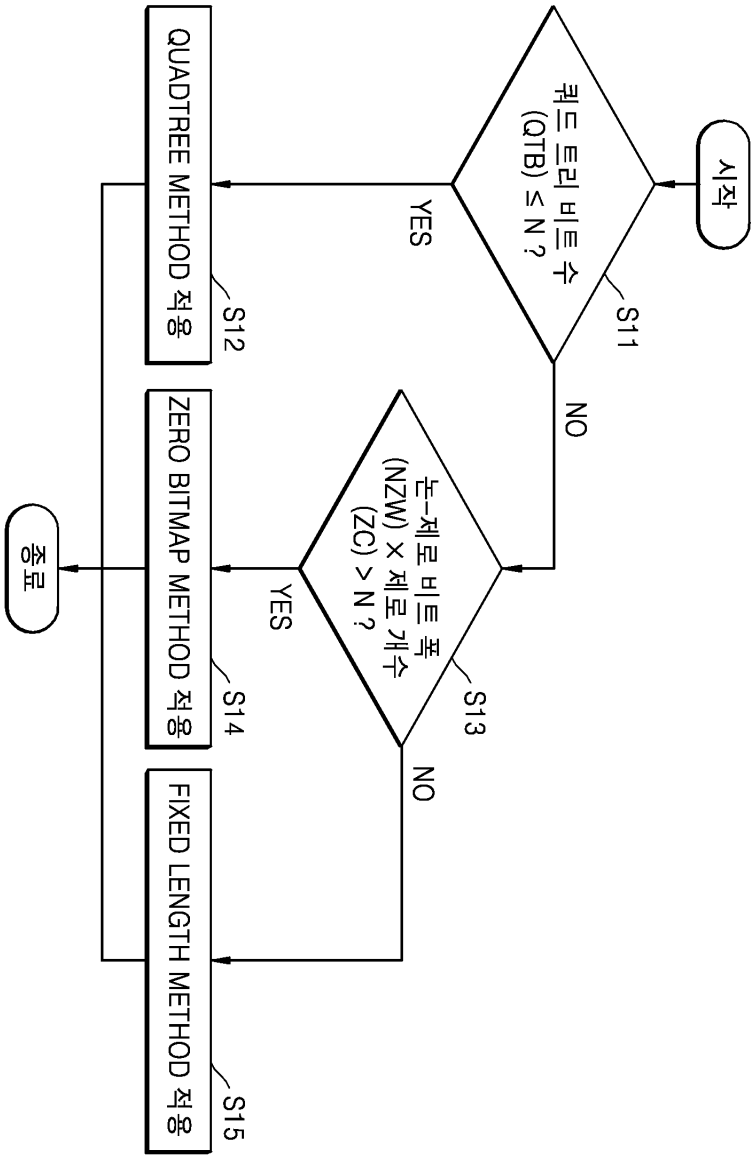
도면3a



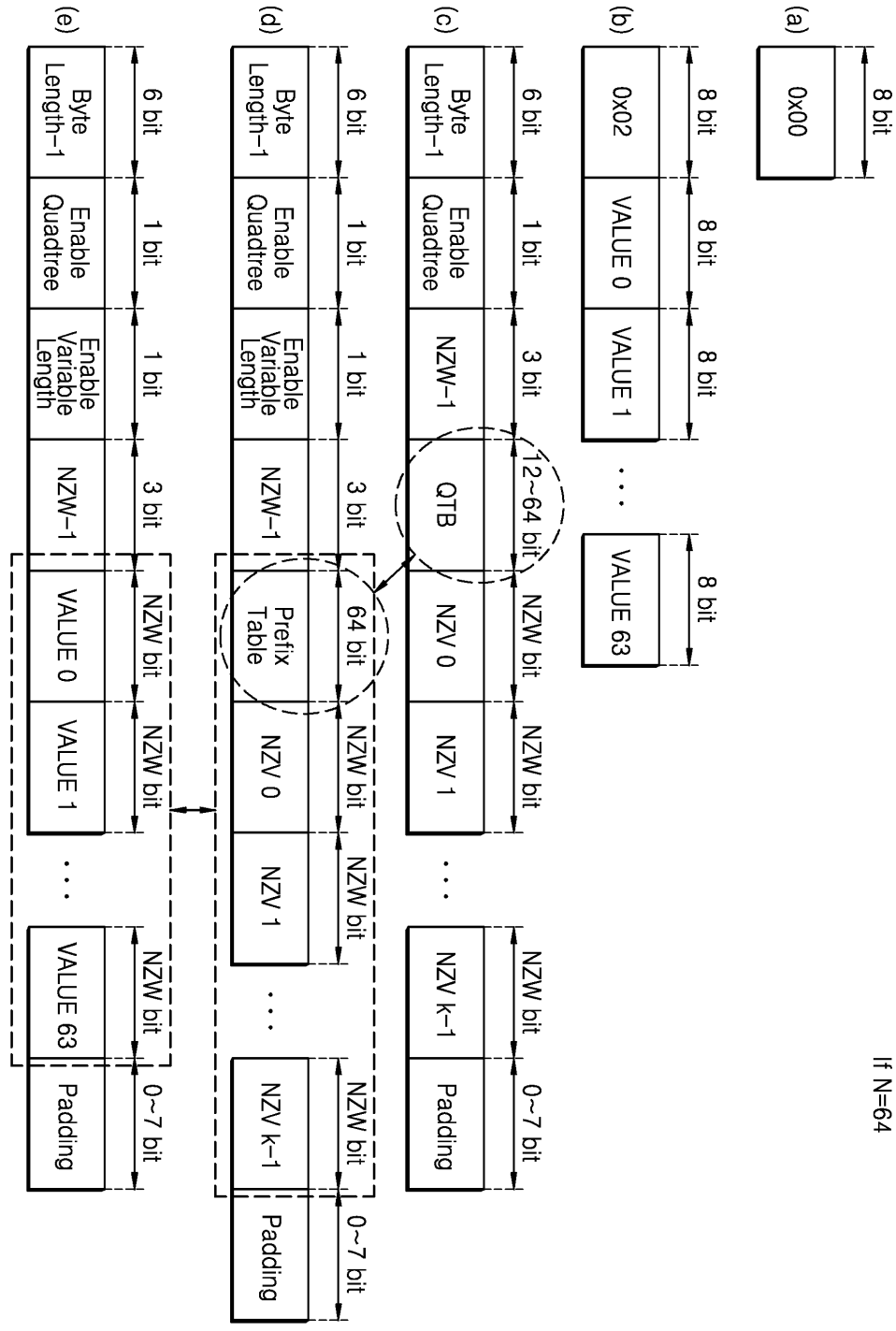
도면3b


TENSOR 2
FM1
FM2
FM3
FM4
CG1
CG2
CG3
CG4
0E
06
1000
0011
0010
0110
1110
011
000100
NZV 2
QTB, ZC
NZW 2
LENGTH 2

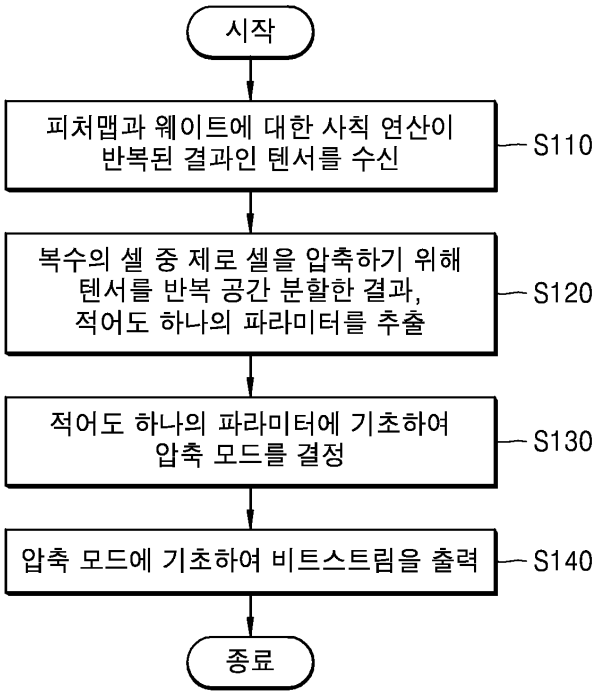
도면4


시작
쿼드 트리 비트 수 (QTB) ≤ N ?
S11
YES
NO
QUADTREE METHOD 적용
S12
논-제로 비트 폭 (NZW) × 제로 개수 (ZC) > N ?
S13
YES
NO
ZERO BITMAP METHOD 적용
S14
FIXED LENGTH METHOD 적용
S15
종료

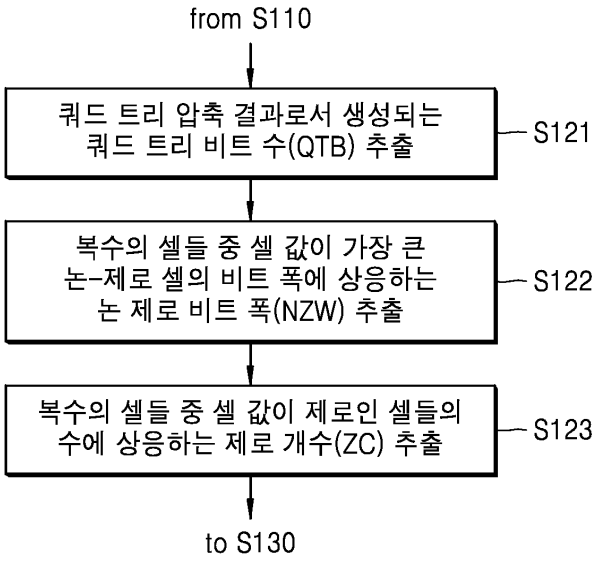
도면5


(a)
0x00
8 bit
(b)
0x02
VALUE 0
VALUE 1
VALUE 63
8 bit
(c)
Byte Length-1
Enable Quadtree
NZW-1
QTB
12~64 bit
NZV 0
NZV 1
NZV k-1
Padding
6 bit
1 bit
3 bit
NZW bit
0~7 bit
(d)
Byte Length-1
Enable Quadtree
Enable Variable Length
NZW-1
Prefix Table
64 bit
NZV 0
NZV 1
NZV k-1
Padding
(e)
Byte Length-1
Enable Quadtree
Enable Variable Length
NZW-1
VALUE 0
VALUE 1
VALUE 63
Padding
If N=64

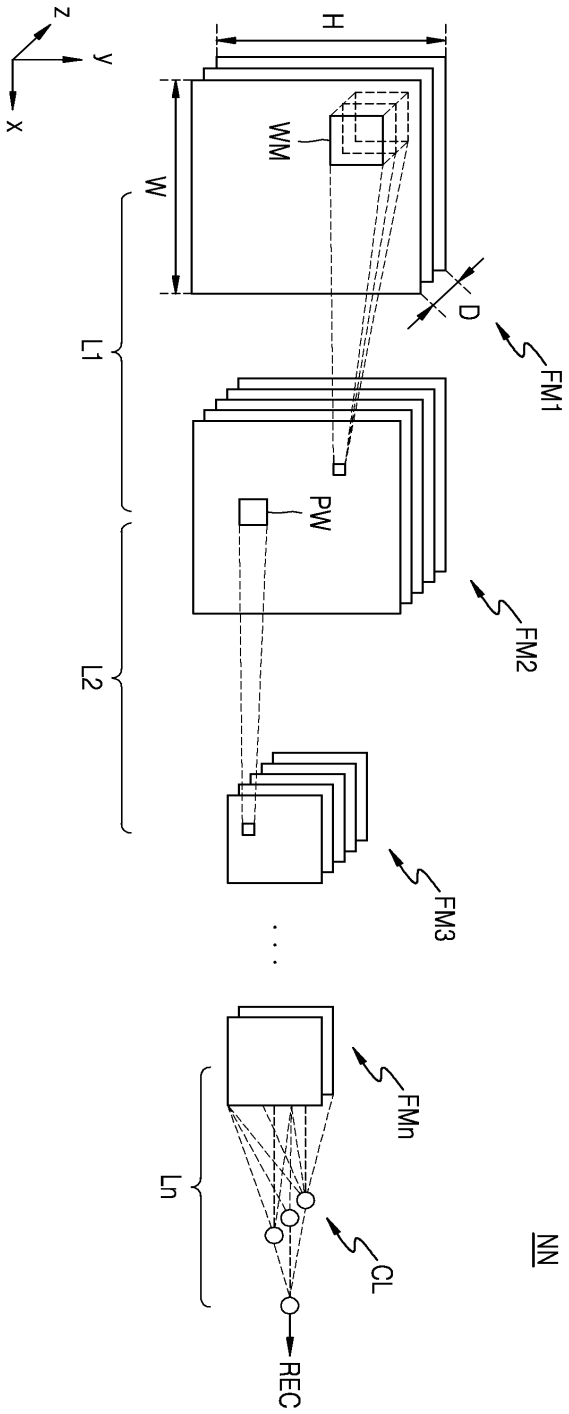
도면6


시작
피처맵과 웨이트에 대한 사칙 연산이 반복된 결과인 텐서를 수신
S110
복수의 셀 중 제로 셀을 압축하기 위해 텐서를 반복 공간 분할한 결과, 적어도 하나의 파라미터를 추출
S120
적어도 하나의 파라미터에 기초하여 압축 모드를 결정
S130
압축 모드에 기초하여 비트스트림을 출력
S140
종료

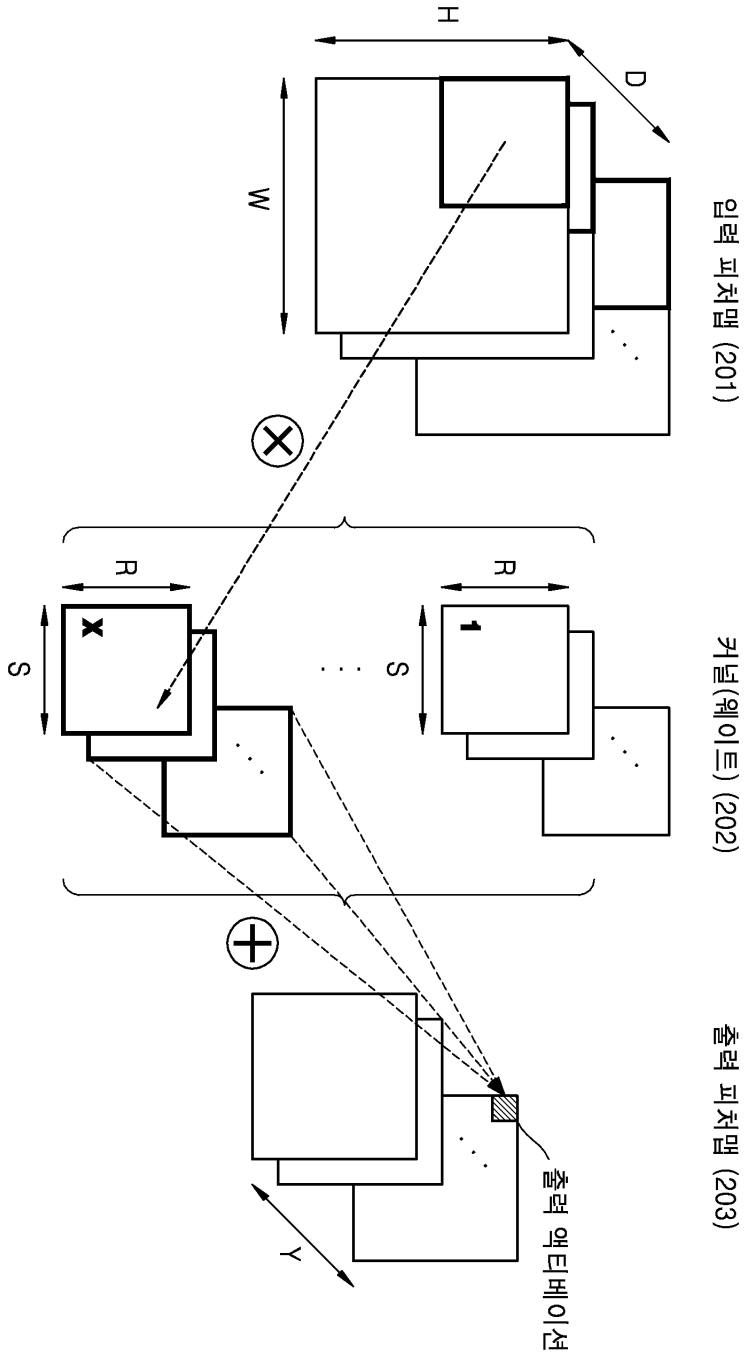
도면7


from S110
쿼드 트리 압축 결과로서 생성되는 쿼드 트리 비트 수(QTB) 추출
S121
복수의 셀들 중 셀 값이 가장 큰 논-제로 셀의 비트 폭에 상응하는 논 제로 비트 폭(NZW) 추출
S122
복수의 셀들 중 셀 값이 제로인 셀들의 수에 상응하는 제로 개수(ZC) 추출
S123
to S130

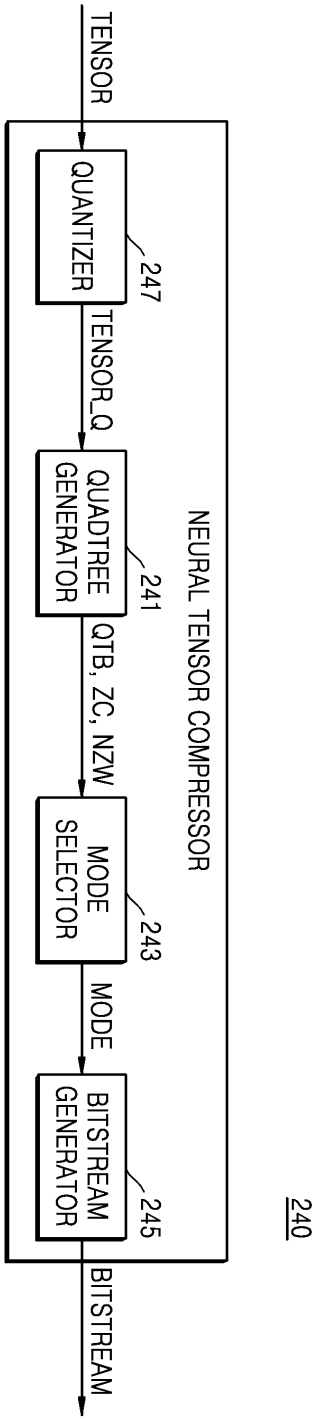
도면8



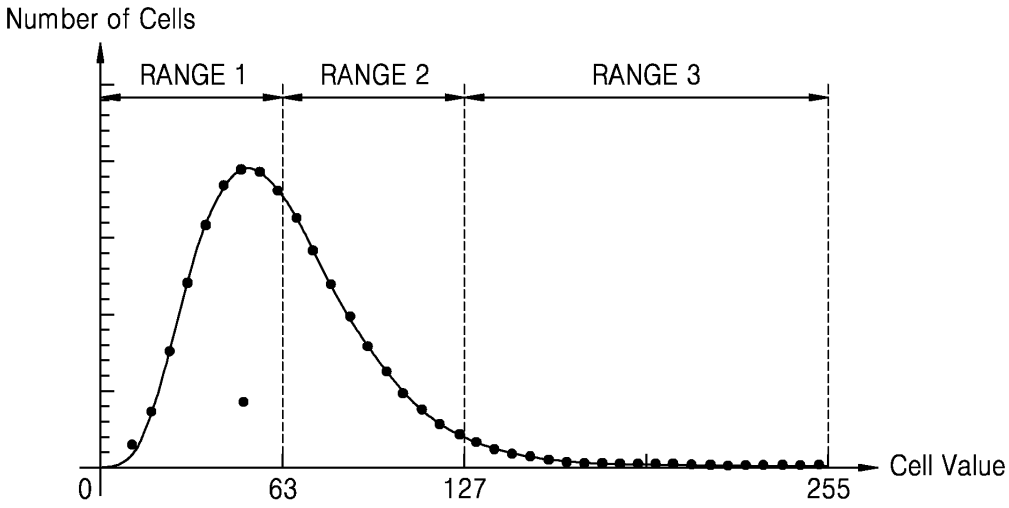
도면9



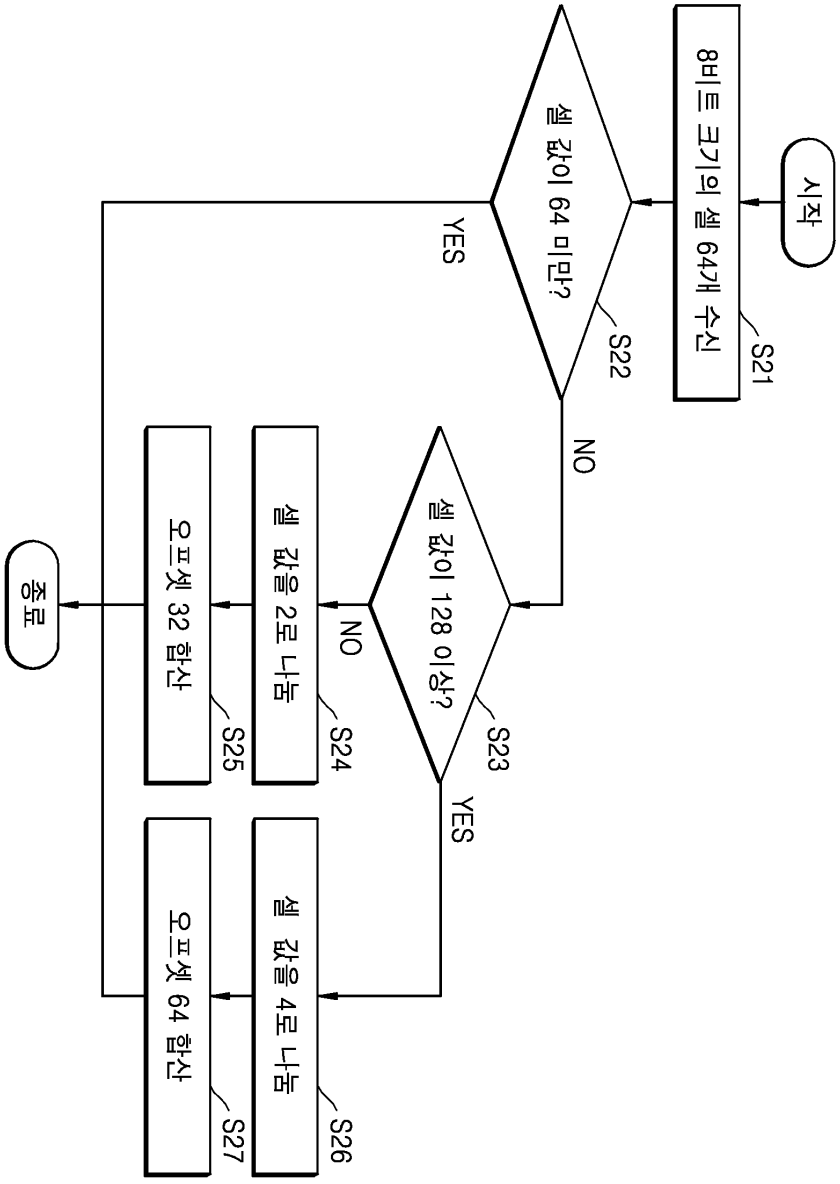
도면10


TENSOR
QUANTIZER
247
TENSOR_Q
QUADTREE GENERATOR
241
QTB, ZC, NZW
MODE SELECTOR
243
MODE
BITSTREAM GENERATOR
245
NEURAL TENSOR COMPRESSOR
240
BITSTREAM

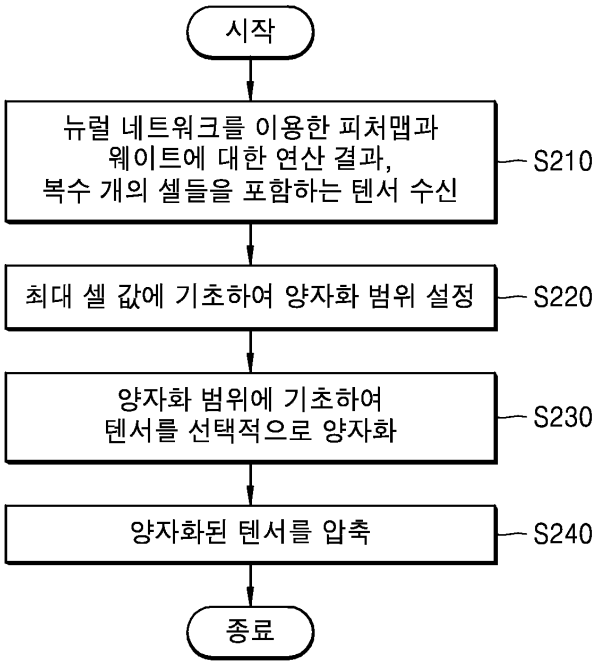
도면11


Number of Cells
RANGE 1
RANGE 2
RANGE 3
0
63
127
255
Cell Value

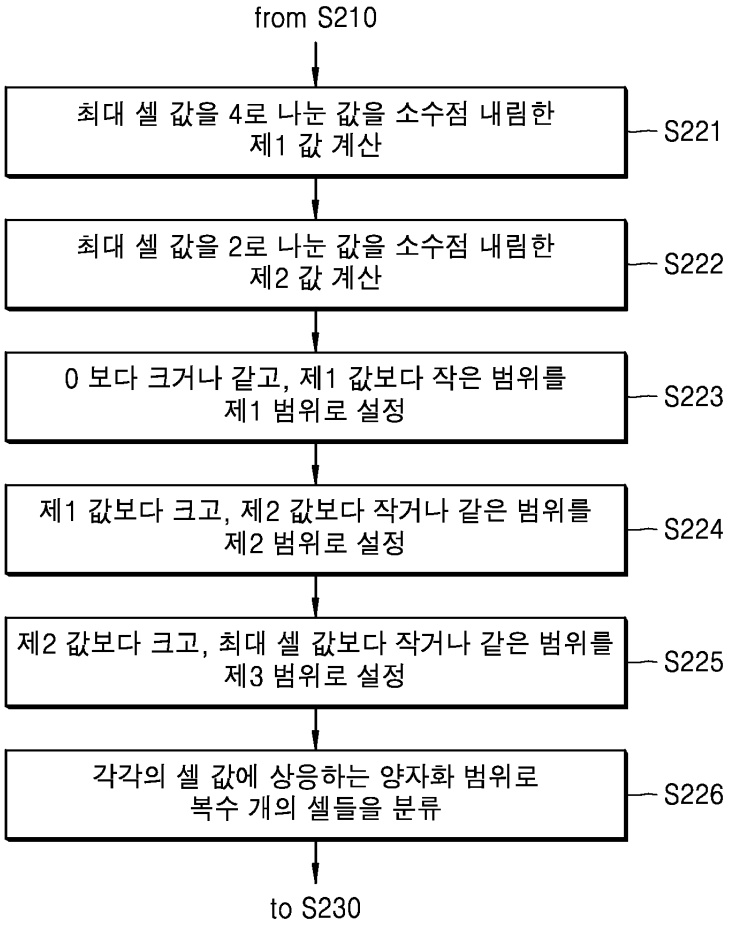
도면12


시작
8비트 크기의 셀 64개 수신
S21
셀 값이 64 미만?
S22
YES
NO
셀 값이 128 이상?
S23
NO
YES
셀 값을 2로 나눔
S24
셀 값을 4로 나눔
S26
오프셋 32 합산
S25
오프셋 64 합산
S27
종료

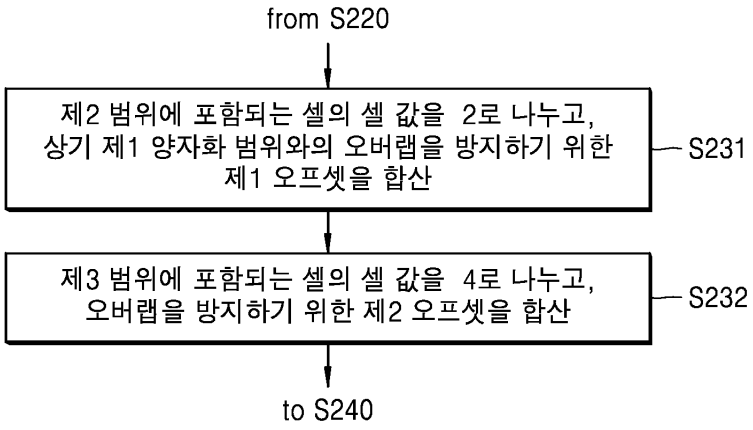
도면13


시작
뉴럴 네트워크를 이용한 피처맵과 웨이트에 대한 연산 결과, 복수 개의 셀들을 포함하는 텐서 수신
S210
최대 셀 값에 기초하여 양자화 범위 설정
S220
양자화 범위에 기초하여 텐서를 선택적으로 양자화
S230
양자화된 텐서를 압축
S240
종료

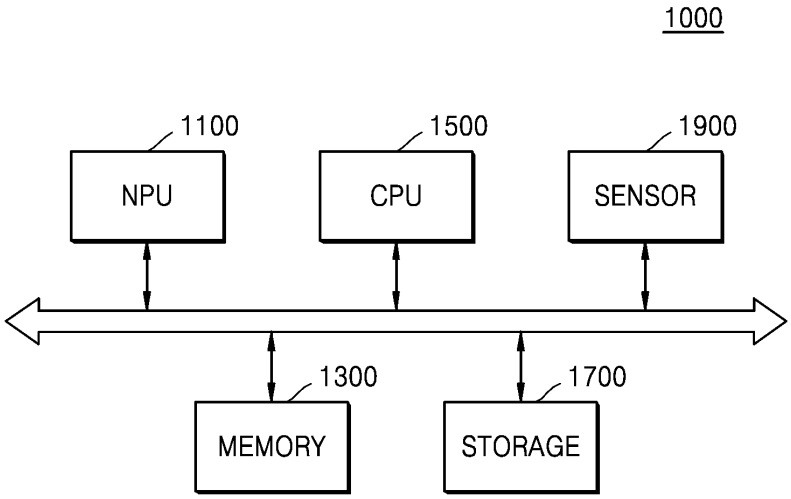
도면14


from S210
최대 셀 값을 4로 나눈 값을 소수점 내림한 제1 값 계산
S221
최대 셀 값을 2로 나눈 값을 소수점 내림한 제2 값 계산
S222
0 보다 크거나 같고, 제1 값보다 작은 범위를 제1 범위로 설정
S223
제1 값보다 크고, 제2 값보다 작거나 같은 범위를 제2 범위로 설정
S224
제2 값보다 크고, 최대 셀 값보다 작거나 같은 범위를 제3 범위로 설정
S225
각각의 셀 값에 상응하는 양자화 범위로 복수 개의 셀들을 분류
S226
to S230

도면15


from S220
제2 범위에 포함되는 셀의 셀 값을 2로 나누고, 상기 제1 양자화 범위와의 오버랩을 방지하기 위한 제1 오프셋을 합산
S231
제3 범위에 포함되는 셀의 셀 값을 4로 나누고, 오버랩을 방지하기 위한 제2 오프셋을 합산
S232
to S240

도면16


1000
1100
NPU
1500
CPU
1900
SENSOR
1300
MEMORY
1700
STORAGE