(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년02월21일
(11) 등록번호 10-2502162
(24) 등록일자 2023년02월16일

(51) 국제특허분류(Int. Cl.)
G06T 3/40 (2006.01) *G06N 3/04* (2023.01)
G06T 1/20 (2018.01) *G06T 1/60* (2006.01)
(52) CPC특허분류
G06T 3/40 (2013.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2021-0056723
(22) 출원일자 2021년04월30일
심사청구일자 2021년04월30일
(65) 공개번호 10-2022-0149281
(43) 공개일자 2022년11월08일
(56) 선행기술조사문헌
(뒷면에 계속)
전체 청구항 수 : 총 20 항

(73) 특허권자
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
강석형
경상북도 포항시 남구 지곡로 155, 5동 203호 (지곡동, 교수아파트)
김성훈
경기도 화성시 경기대로1021번길 6-4, 401호(병점동)
(뒷면에 계속)
(74) 대리인
특허법인 무한

심사관 : 이정은

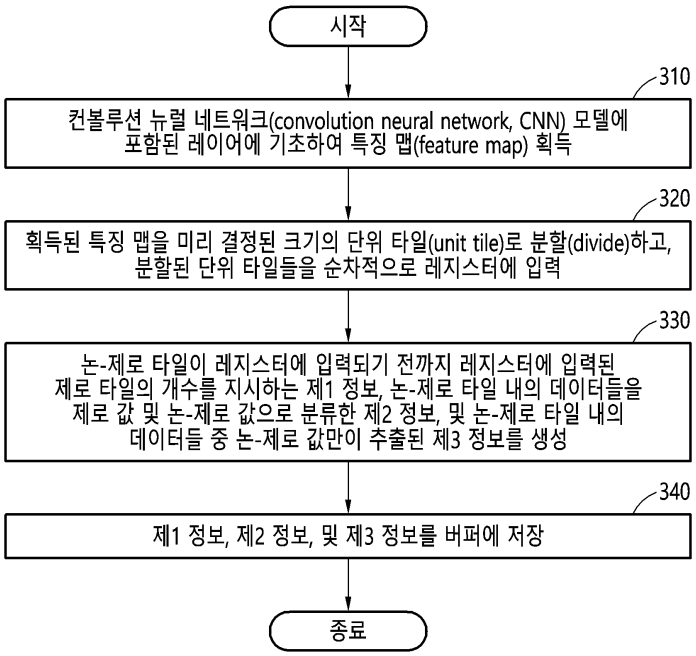
(54) 발명의 명칭 **특징 맵을 컴프레싱하는 장치 및 방법**

(57) 요 약

일 실시예에 따른 컴프레서에 의해 수행되는 특징 맵(feature map)을 컴프레싱하는 방법은, 입력 이미지로부터 컨볼루션 뉴럴 네트워크(convolution neural network, CNN) 모델에 포함된 하나 이상의 레이어에 기초하여 특징 맵(feature map)을 획득하는 단계, 획득된 특징 맵을 미리 결정된 크기의 단위 타일(unit tile)로 분할(divid

(뒷면에 계속)

대 표 도 - 도3



e)하고, 분할된 단위 타일들을 순차적으로 레지스터에 입력시키는 단계, 논-제로 타일이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하는 단계, 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하는 단계, 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하는 단계. 및 제1 정보, 제2 정보, 및 제3 정보를 버퍼(buffer)에 저장하는 단계를 포함하고, 논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며, 제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타낼 수 있다.

(52) CPC특허분류
G06T 1/20 (2013.01)
G06T 1/60 (2013.01)

(72) 발명자
권은지
부산광역시 금정구 수림로19번길 26, 101동 401호 (부곡동, 동원로얄듀크)
박윤호
서울특별시 송파구 송이로 88, 3동 1209호(가락동, 가락대림아파트)
강예성
경기도 용인시 기흥구 공세로 76, 101동 509호(고매동, 세원아파트)

(56) 선행기술조사문헌
Jorge Albericio, et al., "Cnvlutin: ineffectual-neuron-free deep neural network computing", ACM SIGARCH Computer Architecture NewsVolume 44 Issue 3, (2016.06.18.)
Li-Na Wang, et al., "Compressing Deep Networks by Neuron Agglomerative Clustering", Smart Sensing and Advanced Machine Learning Based Emerging Intelligent Systems, (2020.10.23.)
Rutishauser, et al., "An On-the-Fly Feature Map Compression Engine for Background Memory Access Cost Reduction in DNN Inference", (2020.01.09.)
Lin, et al., "Supporting Compressed-Sparse Activations and Weights on SIMD-like Accelerator for Sparse Convolutional Neural Networks", 2018 23rd Asia and South Pacific Design Automation Conference

이 발명을 지원한 국가연구개발사업
과제고유번호 1711121286
과제번호 2020M3F3A2A02082435
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 차세대지능형반도체기술개발(R&D)
연구과제명 신소자 개발지원을 위한 CMOS 통합 집적 센서 인터페이스 SoC 개발
기 여 율 1/2
과제수행기관명 포항공과대학교
연구기간 2020.07.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711152830
과제번호 2021-0-00754-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 SW컴퓨팅산업원천기술개발
연구과제명 인공지능 반도체 설계 SW(Software) 개발
기 여 율 1/2
과제수행기관명 한국과학기술원
연구기간 2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

컴프레서에 의해 수행되는 특징 맵(feature map)을 컴프레싱하는 방법에 있어서,

입력 이미지로부터 컨볼루션 뉴럴 네트워크(convolution neural network, CNN) 모델에 포함된 하나 이상의 레이어에 기초하여 특징 맵(feature map)을 획득하는 단계;

상기 획득된 특징 맵을 미리 결정된 크기의 단위 타일(unit tile)로 분할(divide)하고, 상기 분할된 단위 타일들을 순차적으로 레지스터에 입력시키는 단계;

논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하는 단계;

상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하는 단계;

상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하는 단계; 및

상기 제1 정보, 상기 제2 정보, 및 상기 제3 정보를 버퍼(buffer)에 저장하는 단계

를 포함하고,

논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며,

제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타내는,

특징 맵 컴프레싱 방법.

청구항 2

제1항에 있어서,

상기 제1 정보를 생성하는 단계는,

상기 레지스터에 입력된 단위 타일 내의 데이터들이 모두 제로 값인 경우, 카운트를 증가시키는 단계

를 포함하는 특징 맵 컴프레싱 방법.

청구항 3

제2항에 있어서,

상기 제1 정보를 생성하는 단계는,

상기 카운트가 임계값에 도달하는 경우, 상기 임계값으로 상기 제1 정보를 생성하는 단계

를 포함하고,

상기 제2 정보를 생성하는 단계는,

상기 카운트가 임계값에 도달하는 경우, 제로 값으로 상기 제2 정보를 생성하는 단계

를 포함하는 특징 맵 컴프레싱 방법.

청구항 4

제1항에 있어서,

상기 제3 정보를 생성하는 단계는,

상기 논-제로 타일 내의 논-제로 값의 데이터를 제1 비트 폭(bit width)의 데이터로 나타내어 상기 제3 정보를 생성하는 단계

를 포함하는 특징 맵 컴프레싱 방법.

청구항 5

제1항에 있어서,

상기 제2 정보를 생성하는 단계는,

상기 논-제로 타일 내의 데이터를 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하는 단계

를 포함하고,

상기 제3 정보를 생성하는 단계는,

아웃라이어로 분류된 데이터를 제1 비트 폭의 데이터로 나타내고, 논-아웃라이어로 분류된 데이터를 상기 제1 비트 폭 보다 작은 제2 비트 폭의 데이터로 나타냄으로써 상기 제3 정보를 생성하는 단계

를 포함하며,

아웃라이어는, N 비트 고정 소수점 시스템(fixed-point system)에서 N/2 비트를 사용하여 표현할 수 없는 데이터를 나타내고,

논-아웃라이어는, 상기 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현 가능한 데이터를 나타내는,

특징 맵 컴프레싱 방법.

청구항 6

제5항에 있어서,

상기 분할된 단위 타일들을 레지스터에 입력시키는 단계는,

단위 타일 내의 데이터들을 상위 비트 데이터 및 하위 비트 데이터로 나누어 상기 레지스터에 입력시키는 단계

를 포함하고,

상기 제2 정보를 생성하는 단계는,

대상 데이터에 대응하는 상위 비트 데이터 및 하위 비트 데이터가 제로 값인지 여부를 판단하여, 상기 대상 데이터를 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하는 단계

를 포함하는 특징 맵 컴프레싱 방법.

청구항 7

제6항에 있어서,

상기 분류하는 단계는,

상위 비트 데이터 및 하위 비트 데이터가 모두 논-제로 값인 경우, 상기 대상 데이터를 아웃라이어로 분류하는 단계;

상위 비트 데이터가 제로 값이고, 하위 비트 데이터가 논-제로 값인 경우 상기 대상 데이터를 논-아웃라이어로 분류하는 단계; 및

상위 비트 데이터 및 하위 비트 데이터가 모두 제로 값인 경우, 상기 대상 데이터를 제로 값으로 분류하는 단계

를 포함하는 특징 맵 컴프레싱 방법.

청구항 8

제5항에 있어서,

상기 제2 정보를 생성하는 단계는,

상기 논-제로 타일 내에서 제로 값으로 분류된 데이터에 제1 값을 매핑하고, 논-아웃라이어로 분류된 데이터에 제2 값을 매핑하며, 아웃라이어로 분류된 데이터에 제3 값을 매핑하여 상기 제2 정보를 생성하는 단계

를 포함하고,

상기 제1 값, 상기 제2 값, 및 상기 제3 값은 서로 상이한 데이터인,

특징 맵 컴프레싱 방법.

청구항 9

제1항 내지 제8항 중 어느 한 항의 방법을 수행하기 위한 명령어를 포함하는 하나 이상의 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능 저장 매체.

청구항 10

특징 맵을 컴프레싱하는 컴프레서에 있어서,

입력 이미지로부터 컨볼루션 뉴럴 네트워크 모델에 포함된 하나 이상의 레이어에 기초하여 획득된 특징 맵(feature map)에 대하여, 상기 특징 맵이 미리 결정된 크기의 단위 타일들로 분할되고, 상기 분할된 단위 타일들이 순차적으로 입력되는 레지스터;

레지스터에 저장된 데이터들을 비교하는 비교기;

상기 비교의 결과에 기초하여 논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하며, 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하는 제어기; 및

상기 제1 정보, 상기 제2 정보, 및 상기 제3 정보를 저장하고, 저장된 데이터를 동적 랜덤 액세스 메모리(DRAM)으로 출력하는 버퍼

를 포함하고,

논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며,

제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타내는,

컴프레서.

청구항 11

제10항에 있어서,

상기 제어기는,

상기 레지스터에 입력된 단위 타일 내의 데이터들이 모두 제로 값인 경우, 카운트를 증가시키는,

컴프레서.

청구항 12

제11항에 있어서,

상기 제어기는,

상기 카운트가 임계값에 도달하는 경우, 상기 임계값으로 상기 제1 정보를 생성하고, 제로 값으로 상기 제2 정보를 생성하는,

컴프레서.

청구항 13

제10항에 있어서,

상기 제어기는,

상기 논-제로 타일 내의 논-제로 값의 데이터를 제1 비트 폭(bit width)의 데이터로 나타내어 상기 제3 정보를 생성하는,

컴프레서.

청구항 14

제10항에 있어서,

상기 제어기는,

상기 논-제로 타일 내의 데이터들 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하고,

아웃라이어로 분류된 데이터를 제1 비트 폭의 데이터로 나타내며, 논-아웃라이어로 분류된 데이터를 상기 제1 비트 폭 보다 작은 제2 비트 폭의 데이터로 나타냄으로써 상기 제3 정보를 생성하고,

아웃라이어는, N 비트 고정 소수점 시스템(fixed-point system)에서 N/2 비트를 사용하여 표현할 수 없는 데이터를 나타내며,

논-아웃라이어는, 상기 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현 가능한 데이터를 나타내는,

컴프레서.

청구항 15

제14항에 있어서,

상기 레지스터는,

단위 타일 내의 데이터들이 상위 비트 데이터 및 하위 비트 데이터로 나누어진 입력을 수신하고,

상기 비교기는,

대상 데이터에 대응하는 상위 비트 데이터 및 하위 비트 데이터가 제로 값인지 여부를 판단하고,

상기 제어기는,

상기 판단 결과에 기초하여 상기 대상 데이터를 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하는,

컴프레서.

청구항 16

제15항에 있어서,

상기 제어기는,

상기 판단 결과에 기초하여, 상위 비트 데이터 및 하위 비트 데이터가 모두 논-제로 값인 경우, 상기 대상 데이터를 아웃라이어로 분류하고, 상위 비트 데이터가 제로 값이고, 하위 비트 데이터가 논-제로 값인 경우 상기 대상 데이터를 논-아웃라이어로 분류하며, 상위 비트 데이터 및 하위 비트 데이터가 모두 제로 값인 경우, 상기 대상 데이터를 제로 값으로 분류하는,

컴프레서.

청구항 17

제14항에 있어서,

상기 제어기는,

상기 논-제로 타일 내에서 제로 값으로 분류된 데이터에 제1 값을 매핑하고, 논-아웃라이어로 분류된 데이터에 제2 값을 매핑하며, 아웃라이어로 분류된 데이터에 제3 값을 매핑하여 상기 제2 정보를 생성하고,

상기 제1 값, 상기 제2 값, 및 상기 제3 값을 서로 상이한 데이터인,

컴프레서.

청구항 18

뉴럴 네트워크 가속기 장치에 있어서,

컨볼루션 뉴럴 네트워크 연산을 수행하는 프로세싱 엘리먼트 어레이(PE array)를 이용하여 컨볼루션 뉴런 네트워크 모델에 포함된 하나 이상의 레이어에 기초하여 입력 이미지로부터 추출된 특징 맵을 획득하는 버퍼;

상기 획득된 특징 맵을 미리 결정된 크기의 단위 타일들로 분할하고, 상기 분할된 단위 타일들을 순차적으로 컴프레서에 입력시키는 제어기(controller); 및

논-제로 타일이 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하며, 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하고, 상기 제1 정보, 상기 제2 정보, 및 상기 제3 정보를 저장하며, 저장된 데이터를 동적 랜덤 액세스 메모리(DRAM)으로 출력하는 컴프레서(compressor)

를 포함하고,

논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내고,

제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타내는,

뉴럴 네트워크 가속기 장치.

청구항 19

디컴프레서에 의해 수행되는 특징 맵을 디컴프레싱하는 방법에 있어서,

컨볼루션 네트워크 모델의 레이어에 기초하여 획득된 특징 맵이 컴프레싱된 데이터를 레지스터에 입력시키는 단

계; 및

논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보에 대응하는 개수의 제로 타일을 생성하여 버퍼로 전송하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보 및 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보에 기초하여 논-제로 타일을 생성하여 버퍼로 전송하는 단계

를 포함하고,

상기 제2 정보 및 상기 제3 정보에 기초하여 논-제로 타일을 생성하여 상기 버퍼로 전송하는 단계는,

상기 제2 정보를 이용하여 상기 논-제로 타일 내에서 제로 값의 데이터와 논-제로 값의 데이터를 구분하고, 논-제로 값의 데이터를 상기 제3 정보와 순차적으로 대응시켜 상기 논-제로 타일을 생성하는 단계

를 포함하고,

논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며,

제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타내는,

특징 맵 디컴프레싱 방법.

청구항 20

특징 맵을 디컴프레싱하는 디컴프레서에 있어서,

컨볼루션 네트워크 모델의 레이어에 기초하여 획득된 특징 맵이 압축된 데이터가 입력되는 레지스터;

논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보에 대응하는 개수의 제로 타일을 생성하여 버퍼로 전송하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보 및 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보에 기초하여 논-제로 타일을 생성하여 버퍼로 전송하는 제어기

를 포함하고,

상기 제어기는,

상기 제2 정보를 이용하여 상기 논-제로 타일 내에서 제로 값의 데이터와 논-제로 값의 데이터를 구분하고, 논-제로 값의 데이터를 상기 제3 정보와 순차적으로 대응시켜 상기 논-제로 타일을 생성하고,

논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며,

제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타내는,

디컴프레서.

발명의 설명

기 술 분 야

[0001] 이하, 컨볼루션 뉴럴 네트워크 모델의 레이어에 기초하여 획득된 특징 맵을 컴프레싱하는 장치 및 방법에 관한 기술이 제공된다.

배 경 기 술

[0002] 컨볼루션 뉴럴 네트워크 구조는 많은 양의 오프 칩 메모리(off-chip memory)(예를 들어, 동적 랜덤 액세스 메모리(dynamic random access memory, DRAM))에 대한 액세스(access)를 필요로 하며, 오프 칩 메모리에 대한 액세스를 통해 매개변수를 내부 버퍼에 저장한다. 이러한 외부 DRAM에 대한 액세스는 네트워크 에너지 소비의 대부분을 차지한다. 일반적으로 컨볼루션 뉴럴 네트워크 가속기 장치는 CNN 모델에 포함된 레이어의 출력 특징 맵(feature map)을 컴프레싱하여 전력 소모를 줄이는 방법을 사용한다. 그러나, 종래의 특징 맵 컴프레싱 방법은

특징 맵 내에서 제로('0') 값이 많지 않게 되면 압축률(compression ratio)이 떨어진다는 단점이 존재한다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0003] 일 실시예에 따른 컴프레서에 의해 수행되는 특징 맵(feature map)을 컴프레싱하는 방법은, 입력 이미지로부터 컨볼루션 뉴럴 네트워크(convolution neural network, CNN) 모델에 포함된 하나 이상의 레이어에 기초하여 특징 맵(feature map)을 획득하는 단계, 상기 획득된 특징 맵을 미리 결정된 크기의 단위 타일(unit tile)로 분할(divide)하고, 상기 분할된 단위 타일들을 순차적으로 레지스터에 입력시키는 단계, 논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하는 단계, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하는 단계, 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하는 단계, 및 상기 제1 정보, 상기 제2 정보, 및 상기 제3 정보를 버퍼(buffer)에 저장하는 단계를 포함하고, 논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며, 제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타낼 수 있다.

상기 제1 정보를 생성하는 단계는, 상기 레지스터에 입력된 단위 타일 내의 데이터들이 모두 제로 값인 경우, 카운트를 증가시키는 단계를 포함할 수 있다.

상기 제1 정보를 생성하는 단계는, 상기 카운트가 임계값에 도달하는 경우, 상기 임계값으로 상기 제1 정보를 생성하는 단계를 포함하고, 상기 제2 정보를 생성하는 단계는, 상기 카운트가 임계값에 도달하는 경우, 제로 값으로 상기 제2 정보를 생성하는 단계를 포함할 수 있다.

상기 제3 정보를 생성하는 단계는, 상기 논-제로 타일 내의 논-제로 값의 데이터를 제1 비트 폭(bit width)의 데이터로 나타내어 상기 제3 정보를 생성하는 단계를 포함할 수 있다.

상기 제2 정보를 생성하는 단계는, 상기 논-제로 타일 내의 데이터를 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하는 단계를 포함하고, 상기 제3 정보를 생성하는 단계는, 아웃라이어로 분류된 데이터를 제1 비트 폭의 데이터로 나타내고, 논-아웃라이어로 분류된 데이터를 상기 제1 비트 폭 보다 작은 제2 비트 폭의 데이터로 나타냄으로써 상기 제3 정보를 생성하는 단계를 포함하며, 아웃라이어는, N 비트 고정 소수점 시스템(fixed-point system)에서 N/2 비트를 사용하여 표현할 수 없는 데이터를 나타내고, 논-아웃라이어는, 상기 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현 가능한 데이터를 나타낼 수 있다.

상기 분할된 단위 타일들을 레지스터에 입력시키는 단계는, 단위 타일 내의 데이터들을 상위 비트 데이터 및 하위 비트 데이터로 나누어 상기 레지스터에 입력시키는 단계를 포함하고, 상기 제2 정보를 생성하는 단계는, 대상 데이터에 대응하는 상위 비트 데이터 및 하위 비트 데이터가 제로 값인지 여부를 판단하여, 상기 대상 데이터를 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하는 단계를 포함할 수 있다.

상기 분류하는 단계는, 상위 비트 데이터 및 하위 비트 데이터가 모두 논-제로 값인 경우, 상기 대상 데이터를 아웃라이어로 분류하는 단계, 상위 비트 데이터가 제로 값이고, 하위 비트 데이터가 논-제로 값인 경우 상기 대상 데이터를 논-아웃라이어로 분류하는 단계, 및 상위 비트 데이터 및 하위 비트 데이터가 모두 제로 값인 경우, 상기 대상 데이터를 제로 값으로 분류하는 단계를 포함할 수 있다.

상기 제2 정보를 생성하는 단계는, 상기 논-제로 타일 내에서 제로 값으로 분류된 데이터에 제1 값을 매핑하고, 논-아웃라이어로 분류된 데이터에 제2 값을 매핑하며, 아웃라이어로 분류된 데이터에 제3 값을 매핑하여 상기 제2 정보를 생성하는 단계를 포함하고, 상기 제1 값, 상기 제2 값, 및 상기 제3 값은 서로 상이한 데이터일 수 있다.

일 실시예에 따른 특징 맵을 컴프레싱하는 컴프레서는, 입력 이미지로부터 컨볼루션 뉴럴 네트워크 모델에 포함된 하나 이상의 레이어에 기초하여 획득된 특징 맵(feature map)에 대하여, 상기 특징 맵이 미리 결정된 크기의 단위 타일들로 분할되고, 상기 분할된 단위 타일들이 순차적으로 입력되는 레지스터, 레지스터에 저장된 데이터들을 비교하는 비교기, 상기 비교의 결과에 기초하여 논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레

지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하며, 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하는 제어기, 및 상기 제1 정보, 상기 제2 정보, 및 상기 제3 정보를 저장하고, 저장된 데이터를 동적 랜덤 액세스 메모리(DRAM)으로 출력하는 버퍼를 포함하고, 논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며, 제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타낼 수 있다.

상기 제어기는, 상기 레지스터에 입력된 단위 타일 내의 데이터들이 모두 제로 값인 경우, 카운트를 증가시킬 수 있다.

상기 제어기는, 상기 카운트가 임계값에 도달하는 경우, 상기 임계값으로 상기 제1 정보를 생성하고, 제로 값으로 상기 제2 정보를 생성할 수 있다.

상기 제어기는, 상기 논-제로 타일 내의 논-제로 값의 데이터를 제1 비트 폭(bit width)의 데이터로 나타내어 상기 제3 정보를 생성할 수 있다.

상기 제어기는, 상기 논-제로 타일 내의 데이터들 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류하고, 아웃라이어로 분류된 데이터를 제1 비트 폭의 데이터로 나타내며, 논-아웃라이어로 분류된 데이터를 상기 제1 비트 폭 보다 작은 제2 비트 폭의 데이터로 나타냄으로써 상기 제3 정보를 생성하고, 아웃라이어는, N 비트 고정 소수점 시스템(fixed-point system)에서 N/2 비트를 사용하여 표현할 수 없는 데이터를 나타내며, 논-아웃라이어는, 상기 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현 가능한 데이터를 나타낼 수 있다.

상기 레지스터는, 단위 타일 내의 데이터들이 상위 비트 데이터 및 하위 비트 데이터로 나누어진 입력을 수신하고, 상기 비교기는, 대상 데이터에 대응하는 상위 비트 데이터 및 하위 비트 데이터가 제로 값인지 여부를 판단하고, 상기 제어기는, 상기 판단 결과에 기초하여 상기 대상 데이터를 아웃라이어, 논-아웃라이어, 및 제로 값 중 하나로 분류할 수 있다.

상기 제어기는, 상기 판단 결과에 기초하여, 상위 비트 데이터 및 하위 비트 데이터가 모두 논-제로 값인 경우, 상기 대상 데이터를 아웃라이어로 분류하고, 상위 비트 데이터가 제로 값이고, 하위 비트 데이터가 논-제로 값인 경우 상기 대상 데이터를 논-아웃라이어로 분류하며, 상위 비트 데이터 및 하위 비트 데이터가 모두 제로 값인 경우, 상기 대상 데이터를 제로 값으로 분류할 수 있다.

상기 제어기는, 상기 논-제로 타일 내에서 제로 값으로 분류된 데이터에 제1 값을 매핑하고, 논-아웃라이어로 분류된 데이터에 제2 값을 매핑하며, 아웃라이어로 분류된 데이터에 제3 값을 매핑하여 상기 제2 정보를 생성하고, 상기 제1 값, 상기 제2 값, 및 상기 제3 값을 서로 상이한 데이터일 수 있다.

일 실시예에 따른 뉴럴 네트워크 가속기 장치는, 컨볼루션 뉴럴 네트워크 연산을 수행하는 프로세싱 엘리먼트 어레이(PE array)를 이용하여 컨볼루션 뉴런 네트워크 모델에 포함된 하나 이상의 레이어에 기초하여 입력 이미지로부터 추출된 특징 맵을 획득하는 버퍼, 상기 획득된 특징 맵을 미리 결정된 크기의 단위 타일들로 분할하고, 상기 분할된 단위 타일들을 순차적으로 컴프레서에 입력시키는 제어기(controller), 및 논-제로 타일이 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보를 생성하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성하며, 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성하고, 상기 제1 정보, 상기 제2 정보, 및 상기 제3 정보를 저장하며, 저장된 데이터를 동적 랜덤 액세스 메모리(DRAM)으로 출력하는 컴프레서(compressor)를 포함하고, 논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내고, 제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타낼 수 있다.

일 실시예에 따른 디컴프레서에 의해 수행되는 특징 맵을 디컴프레싱하는 방법은, 컨볼루션 네트워크 모델의 레이어에 기초하여 획득된 특징 맵이 컴프레싱된 데이터를 레지스터에 입력시키는 단계, 및 논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보에 대응하는 개수의 제로 타일을 생성하여 버퍼로 전송하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보 및 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보에 기초하여 논-제로 타일을 생성하여 버퍼로 전송하는 단계를 포함하고, 상기 제2 정보 및 상기 제3 정보에 기초하여 논-제로 타일을 생성하여 상기 버퍼로 전송하는 단계는, 상기 제2 정보를 이용하여 상기 논-제로 타일 내에서 제로 값의 데이터와 논-제로 값의 데이터를 구분하고, 논-제로 값의 데이터를 상기 제3 정보와 순차적으로 대응시켜 상기 논-제로 타일을 생성하는 단계를 포함하고, 논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을

나타내며, 제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타낼 수 있다.

일 실시예에 따른 특징 맵을 디컴프레싱하는 디컴프레서는, 컨볼루션 네트워크 모델의 레이어에 기초하여 획득된 특징 맵이 압축된 데이터가 입력되는 레지스터, 논-제로 타일이 상기 레지스터에 입력되기 전까지 상기 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보에 대응하는 개수의 제로 타일을 생성하여 버퍼로 전송하고, 상기 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보 및 상기 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보에 기초하여 논-제로 타일을 생성하여 버퍼로 전송하는 제어기를 포함하고, 상기 제어기는, 상기 제2 정보를 이용하여 상기 논-제로 타일 내에서 제로 값의 데이터와 논-제로 값의 데이터를 구분하고, 논-제로 값의 데이터를 상기 제3 정보와 순차적으로 대응시켜 상기 논-제로 타일을 생성하고, 논-제로 타일은, 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타내며, 제로 타일은, 타일 내의 데이터들이 모두 제로 값인 타일을 나타낼 수 있다.

[0004] 삭제

[0005] 삭제

[0006] 삭제

[0007] 삭제

[0008] 삭제

[0009] 삭제

[0010] 삭제

[0011] 삭제

[0012] 삭제

[0013] 삭제

[0014] 삭제

[0015] 삭제

[0016] 삭제

[0017] 삭제

[0018] 삭제

[0019] 삭제

[0020] 삭제

[0021] 삭제

도면의 간단한 설명

[0022] 도 1은 CNN 모델에서의 특징 맵 및 특징 맵 내에서 데이터의 분포에 관하여 설명한다.

도 2는 특징 맵 내에서 데이터들 사이의 공간적 상호 관계성에 관한 그래프이다.

도 3은 일 실시예에 따른 뉴럴 네트워크 가속기 장치의 동작을 설명하는 흐름도이다.

도 4는 일 실시예에 따른 가속기 장치가 특징 맵을 컴프레싱하는 방법에 관하여 설명한다.

도 5는 일 실시예에 따른 가속기 장치의 구성에 관하여 설명한다.

도 6은 일 실시예에 따른 컴프레서의 구조에 관하여 설명한다.

도 7은 일 실시예에 따른 특징 맵을 디컴프레싱하는 디컴프레서의 구조에 관하여 설명한다.

도 8은 단위 타일의 크기에 따른 정규화된 DRAM 액세스의 역수에 대한 그래프를 도시한다.

도 9는 다양한 컴프레싱 방법에 따른 특징 맵의 압축률 및 희소성(sparsity)에 관한 그래프를 나타낸다.

도 10은 양자화된 데이터에 대하여 컴프레싱된 특징 맵의 정규화된 DRAM 액세스의 값을 나타낸다.

도 11은 특징 맵을 저장 및 로드(load)하는데 소비되는 에너지에 관한 그래프를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0023] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0024] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0025] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0026] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0027] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0028] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면

부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.

도 1은 CNN 모델에서의 특징 맵 및 특징 맵 내에서 데이터의 분포에 관하여 설명한다.

[0029] 삭제

[0030] 최근 컨볼루션 뉴럴 네트워크(convolution neural network, CNN)(이하, 'CNN')의 크기가 커짐에 따라, CNN의 정확도가 향상되었다. 최신 CNN 모델에서는 수백만번의 연산(computation)과 수백 메가바이트(megabyte)의 매개변수(parameter)가 요구된다. CNN 가속기 장치는 DRAM에 대한 액세스를 통하여 매개변수를 내부 버퍼에 저장한다. DRAM에 대한 액세스는 CNN 가속기 장치의 에너지 소비의 대부분을 차지한다. CNN 가속기 장치는 에너지 소비를 줄이기 위하여, 특징 맵(feature map)을 컴프레싱 함으로써 DRAM 액세스 수를 최소화할 수 있다. 특징 맵을 컴프레싱하는 데에 추가적인 전력이 소모되지만, 간단한 산술 연산 또는 내부 버퍼 액세스를 위한 에너지 소비는 DRAM 액세스의 에너지 소비 보다 훨씬 적게 나타난다.

[0031] 종래에는 특징 맵에 대한 효과적인 컴프레싱을 위하여 렐루(rectified linear unit, ReLU) 및 양자화(quantization)로 인한 희소성(sparsity)에 초점을 두었다. 계산 복잡성의 감소와 베니싱-그라디언트(vanishing gradient) 문제를 해결하기 위해 렐루(ReLU)를 활성화 함수가 사용되었고, 특징 맵의 양자화를 통한 정수형(fixed-point) 데이터가 사용되었다. 렐루의 활성화 함수는 음수값 또는 제로값이 입력되면 제로값(0)을 반환하고, 양수값이 입력되면 해당 양수값을 그대로 반환하는 부분 선형 함수이다. 따라서, 특징 맵 내에서 모든 음수값들이 제로값이 되기 때문에 제로값이 많은 특징맵, 즉 희소성(sparsity)이 높은 특징 맵이 도출된다. 양자화(quantization)는 32 비트 실수형(floating point)로 표시되어 있는 값을 32 비트 또는 32 비트 보다 적은 비트를 사용하여 정수형 값으로 변환하는 것을 나타낸다. 양자화를 통하여 제로값에 가까운 값들을 제로값으로 양자화 시킬 수 있고, 희소성이 높은 특징 맵이 도출될 수 있다. 종래에는 렐루 또는 양자화를 통한 희소성이 높은 특징맵의 특성에만 의존하여 특징 맵을 컴프레싱하였다.

도 1을 참조하면, CNN 모델 중 하나의 예시적 모델인 VGG-16 네트워크(110)에 관하여 설명한다. 특징 맵(111, 112)은 VGG-16 네트워크(110)에 포함된 12번째 레이어에 기초하여 획득된 특징 맵을 나타낸다. 특징 맵(111, 112)에서는 제로값의 데이터들은 흰색으로 표시되고, 논-제로 값의 데이터들은 회색 또는 검은색으로 표시된다. 특징 맵(111, 112)에서 논-제로 값의 데이터들은 서로 군집되어 있다. 이는 CNN의 데이터들이 공간적 상호 관계성(spatial correlation)을 갖는다는 것을 나타낸다. 따라서 특정 데이터가 제로값이 아닌 경우, 근처 데이터들 역시 제로값이 아닐 확률이 높다. 일반적으로, 특징맵(111, 112) 내에서 논-제로 값의 데이터들(예를 들어, 검은색으로 표시된 데이터들)은 적은 양으로 나타나고, 특징 맵(111, 112) 내에서 제로값의 데이터들(예를 들어, 흰색으로 표시된 데이터들)은 많은 양으로 나타난다.

데이터 분포(120)는 VGG-16 네트워크(110)의 특징 맵에서 최대 비트 폭(maximum bit-width)에 따른 데이터의 분포를 나타낸다. 최대 비트 폭이란 데이터를 표현하기 위해 최소로 필요로 하는 비트의 수를 나타낼 수 있다. 예를 들어, 7의 데이터는 이진수로 $00000111_{(2)}$ 로 표현될 수 있으며, 최대 비트 폭은 3이다.

특징맵 내의 데이터들은 각각 최대 비트 폭에 따라 제로값(121), 논-아웃라이어(122), 또는 아웃라이어(123) 중 하나로 분류될 수 있다. 특징 맵 내에서 논-제로 값의 데이터는 대부분 N 비트 고정 소수점 시스템(fixed-point system)에서 N/2 비트를 사용하여 표현될 수 있다. 논-아웃라이어(122)은 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현 가능한 데이터를 나타낼 수 있다. 아웃라이어(123)는 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현할 수 없는 데이터를 나타낼 수 있다. 여기서 N은 1 이상의 자연수를 나타낼 수 있다. 제로값(121)은 최대 비트 폭이 '0'이다. 논-아웃라이어(122)는 N/2 이하의 최대 비트 폭을 가진다. 아웃라이어(123)는 N/2를 초과하는 최대 비트 폭을 가진다. 논-아웃라이어 및 아웃라이어는 논-제로 값이라고 나타낼 수 있다.

특징 맵 내에서 아웃라이어로 분류되는 데이터는 매우 적은 부분을 차지하지만, CNN의 정확도에 상당한 영향을 미친다. 아웃라이어는 상대적으로 논-아웃라이어에 비하여 많은 비트 수를 사용하여 표현되어야 하므로, 양자화에 한계점이 존재하고 이로 인하여 데이터의 크기가 클 수밖에 없는 단점이 있다. 반대로, 논-아웃라이어는 N 비트 고정 소수점 시스템에서 N/2 비트만을 사용하여 표현이 가능하다. CNN 특징 맵 내에서 데이터는 대부분 논-아웃라이어 또는 제로값 이므로, 고정된 N 비트를 사용하여 데이터를 표현하는 것은 메모리 사용 측면에서 많은 낭비이다. 그러나, 상대적으로 긴 비트 폭을 유지해야 하는 아웃라이어로 인하여, 데이터 표현의 비트 폭

을 줄일 수 없는 문제점이 존재한다. 반면, 일 실시예에 따른 특징 맵을 컴프레싱하는 컴프레서는 데이터들의 공간적 상호 관계성을 이용하고, 데이터들의 데이터 표현 비트 폭을 줄임으로써 특징 맵의 압축률을 향상시킬 수 있다.

[0032] 삭제

[0033] 삭제

[0034] 삭제

[0035] 삭제

[0037] 도 2는 특징 맵 내에서 데이터들 사이의 공간적 상호 관계성에 관한 그래프이다.

Moran의 I를 사용하여 특징 맵 내에서 데이터들 사이의 공간적 상호 관계성을 정량적으로 평가될 수 있다. ImageNet 2012 학습 데이터 기반으로 트레이닝이 수행된 8 비트 양자화된 VGG-16 네트워크에 관하여 일 예시로 설명한다. VGG-16 네트워크에서 인퍼런스(inference)를 수행하여 VGG-16 네트워크의 각 레이어 마다 특징 맵이 획득될 수 있다. 압축률은 데이터가 제로값인 지 여부에 따라 일정하게 평가될 수 있으나, 공간적 상호 관계성은 데이터의 크기에 따라 달리 평가될 수 있다. 데이터의 크기에 따라 공간적 상호 관계성이 다르게 평가되는 차이를 없애기 위하여, 특징 맵의 데이터들 중 논-제로 값은 모두 '1'의 값으로 변경되어 공간적 상호 관계성이 평가될 수 있다. 예를 들어, 특징 맵 내에서 (x_1, y_1)의 위치를 갖는 데이터와 (x_2, y_2)의 위치를 갖는 데이터 사이의 거리는 $|x_2-x_1|+|y_2-y_1|$으로 표현될 수 있다. 특정 거리(d) 내 데이터들 사이의 공간적 상호 관계성을 조사하기 위하여, 공간 계수(w_{ij})는 아래 수학식 1과 같이 정의될 수 있다. 여기서, i 및 j는 특징 맵 내에서 데이터의 위치 인덱스를 나타낸다.

[0038] 삭제

수학식 1

[0039]
$$w_{ij} = \begin{cases} 1, & \text{if } |i-j| \text{ is } d \\ 0, & \text{otherwise} \end{cases}$$

[0040] 논-제로 값을 '1'의 값으로 변환한 특징 맵의 공간적 자동 상호 관계성(spatial auto correlation)을 평가하기 위하여 Moran의 I가 계산될 수 있다. Moran의 I는 아래 수학식 2와 같이 정의 될 수 있다.

수학식 2

[0041]
$$I = \frac{L}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} \cdot (x_i - \bar{x}) \cdot (x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

수학식 2에서 L은 특징 맵 내의 데이터 총 개수를 나타낸다. x 및 $\bar{x}$는 각각 데이터 및 데이터의 평균을 나타낸다. Moran의 I는 -1 내지 1 사이의 값을 가진다. Moran의 I 값이 '-1'인 경우, 논-제로 값의 데이터들이 완벽하게 분산되어 있음을 나타내고, Moran의 I 값이 '1'인 경우, 논-제로 값의 데이터들이 완벽하게 군집되어 있음

을 나타낸다. 일반적으로 특정 데이터에 대한 Moran의 I가 $\frac{-1}{L-1}$ 보다 크다면, 특정 데이터가 양의 공간적 상호 관계성을 가지는 것으로 판단된다.

도 2는 서로 다른 컨볼루션 레이어 마다 계산된 특정 거리 내 데이터들 사이의 공간적 상호 관계성(Moran의 I)을 나타낸다. 그래프(211)는 거리가 '1'인 데이터들 사이의 공간적 상호 관계성에 관한 그래프이고, 그래프(212)는 거리가 '2'인 데이터들 사이의 공간적 상호 관계성에 관한 그래프이며, 그래프(213)은 거리가 '3'인 데이터들 사이의 공간적 상호 관계성에 관한 그래프이고, 그래프(214)는 거리가 '4'인 데이터들 사이의 공간적 상호 관계성에 관한 그래프이다. 도 2를 참조하면, 가까운 거리 내 데이터들 사이의 공간적 상호 관계성이 더 높은 것으로 나타난다. 거리가 '1'인 데이터들 사이의 평균 Moran의 I는 0.54로 나타나고, 거리가 증가함에 따라 데이터들 사이의 평균 Moran의 I는 0.35, 0.29, 및 0.26으로 감소하는 것으로 나타난다. 공간적 상호 관계성은 모든 거리에서 항상 $\frac{-1}{L-1}$ 보다 큰 것으로 나타나기 때문에, 특징 맵의 데이터들 양의 공간적 상호 관계성을 갖는 것으로 판단된다. 이하에서는, 특징 맵의 높은 공간적 상호 관계성을 이용하기 위하여, 타일(tile)을 기반으로 특징 맵을 컴프레싱하는 방법에 관하여 설명한다.

[0042] 삭제

[0043] 삭제

[0045] 도 3은 일 실시예에 따른 뉴럴 네트워크 가속기 장치의 동작을 설명하는 흐름도이다.

[0046] 동작(310)에서 일 실시예에 따른 뉴럴 네트워크 가속기 장치(이하, '가속기 장치')는 CNN 모델에 포함된 하나 이상의 레이어에 기초하여 특징 맵(feature map)을 획득할 수 있다.

동작(320)에서 일 실시예에 따른 가속기 장치는 획득된 특징 맵을 미리 결정된 크기의 단위 타일(unit tile)로 분할(divide)하고, 분할된 단위 타일들을 순차적으로 컴프레서의 레지스터에 입력할 수 있다. 가속기 장치는 특징 맵의 높은 공간적 상호 관계성을 활용하기 위하여 특징 맵을 복수의 단위 타일들로 분할할 수 있다. 예를 들어, 특징 맵을 m×n 단위 타일로 분할하고, 단위 타일 내의 데이터를 컴프레싱할 수 있다. 단위 타일의 폭(width) 및 높이(height)는 미리 결정될 수 있다. m×n 타일은 일 축으로 m개의 데이터가 배치되고, 일 축과 수직한 다른 축으로 n개의 데이터가 배치된 타일을 나타낼 수 있다.

동작(330)에서 일 실시예에 따른 가속기 장치는 논-제로 타일이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보(런(run)이라고도 함), 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보(마스크(mask)라고도 함), 및 논-제로 타일 내의 데이터들을 중 논-제로 값만이 추출된 제3 정보를 생성할 수 있다.

일 실시예예 따른 가속기 장치는 컴프레서의 레지스터에 입력된 단위 타일을 제로 타일 또는 논-제로 타일로 분류할 수 있다. 제로 타일은 타일 내의 데이터들이 모두 제로값인 타일을 나타낼 수 있다. 논-제로 타일은 타일 내의 데이터들 중 적어도 하나가 논-제로 값인 타일을 나타낼 수 있다. 가속기 장치는 논-제로 타일이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일의 개수를 카운팅하고, 카운팅 결과를 지시하는 제1 정보를 생성할 수 있다. 제1 정보는 연속적인 제로 타일의 개수를 나타낼 수 있다. 예를 들어, 레지스터에 두개의 연속적인 제로 타일이 입력되는 경우, 제1 정보는 $10_{(2)}$이다. 제1 정보의 비트 폭(bit width)은 미리 결정될 수 있다.

일 실시예에 따른 가속기 장치는 레지스터에 입력된 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보를 생성할 수 있다. 제2 정보의 비트 폭은 단위 타일의 크기($m \times n$)에 의하여 결정될 수 있다.

일 실시예에 따른 가속기 장치는 논-제로 타일 내의 데이터들 중 논-제로 값을 갖는 데이터들에 기초하여 제3 정보를 생성할 수 있다. 가속기 장치는 논-제로 타일 내의 데이터들 중 제로값의 데이터를 제외하고, 논-제로 값의 데이터들만을 추출하여 제3 정보를 생성할 수 있다.

동작(340)에서 일 실시예에 따른 가속기 장치는 생성된 제1 정보, 제2 정보, 및 제3 정보를 버퍼(buffer)에 저

장할 수 있다. 컴프레서는 저장된 데이터를 DRAM으로 출력할 수 있다.

[0047] 삭제

[0048] 삭제

[0049] 삭제

[0050] 삭제

[0051] 삭제

[0052] 삭제

[0054] 도 4는 일 실시예에 따른 가속기 장치가 특징 맵을 컴프레싱하는 방법에 관하여 설명한다.

이하에서는, 특징 맵을 컴프레싱하는 다양한 방법 중 일 실시예에 따른 제1 컴프레싱 방법(440)에 관하여 설명한다. 일 실시예에 따른 가속기 장치의 컴프레서(compressor)는 특징 맵(410)을 단위 타일로 분할(divide)할 수 있고, 분할된 단위 타일 내의 데이터들 각각을 제로값 또는 논-제로 값으로 분류할 수 있다. 일 실시예에 따른 컴프레서는 논-제로 타일(423)이 컴프레서의 레지스터에 입력되기 전까지 레지스터에 입력된 연속된 제로 타일(421, 422)의 개수를 지시하는 제1 정보(예를 들어, $10_{(2)}$)를 제1 정보(441)로 생성할 수 있다. 제1 정보의 비트 폭은 미리 결정될 수 있다.

컴프레서는 논-제로 타일(423) 내의 데이터들을 제로값의 데이터(431)와 논-제로 값의 데이터들(432, 433, 434)로 분류한 제2 정보(442)를 생성할 수 있다. 예를 들어, 컴프레서는 제로값의 데이터(431)가 '$0_{(2)}$'을 지시하는 것으로 결정할 수 있고, 논-제로 값의 데이터들(432, 433, 434)이 '1'을 지시하는 것으로 결정할 수 있다. 제2 정보(442)의 비트 폭은 단위 타일의 크기에 따라 결정될 수 있다. 예를 들어, 컴프레서는 단위 타일의 크기가 2×2로 결정된 경우, 제2 정보(442)를 4 비트 폭으로 결정할 수 있다. 이 경우, 컴프레서는 $0111_{(2)}$를 제2 정보로 생성할 수 있다.

일 실시예에 따른 컴프레서는 논-제로 타일(423) 내의 데이터들 중 논-제로 값의 데이터들(432, 433, 434)에 기초하여 제3 정보(443)를 생성할 수 있다. 일 실시예에 따르면, 컴프레서는 논-제로 값의 데이터들(432, 433, 434)을 고정된 제1 비트 폭의 데이터로 나타내어 생성할 수 있다. 예를 들어, N 비트 고정 소수점 시스템에서, 제1 비트 폭은 N 비트 폭을 나타낼 수 있다.

더 나아가, 일 실시예에 따른 컴프레서는 논-제로 타일이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일의 개수를 지시하는 카운트(count)가 임계값에 도달하는 경우, 임계값으로 제1 정보(441)로 생성하고, 제로 값으로 제2 정보(442)를 생성할 수 있다. 여기서, 임계값은 제1 정보의 미리 결정된 비트 폭에 따라 표현 가능한 최대값을 나타낼 수 있다. 예를 들어, 제1 정보에 대하여 미리 결정된 비트 폭이 2 비트인 경우, 임계값은 3($11_{(2)}$)으로 결정된다. 예를 들어, 컴프레서는 레지스터에 입력된 연속적인 제로 타일의 개수가 3개인 경우, 제1 정보를 $11_{(2)}$로 생성하고, 제2 정보는 $0000_{(2)}$로 생성할 수 있다. 컴프레서는 특징 맵의 단위 타일들의 데이터들을 모두 컴프레싱 할 때까지 데이터의 생성 및 저장을 반복하며, 마지막에는 종료 패킷(end of packet)(444)을 생성하고 저장한다. 종료 패킷(444)은 제로 데이터일 수 있다.

[0055] 삭제

[0056] 삭제

[0057] 삭제

[0058] 삭제

이하에서는, 특징 맵을 컴프레싱하는 다양한 방법 중 일 실시예에 따른 제2 컴프레싱 방법(450)에 관하여 설명한다. 가속기 장치의 컴프레서는 데이터들의 데이터 표현 비트 폭을 감소시킴으로써 보다 특징 맵의 압축률을 향상시킬 수 있다. 컴프레서는 특징 맵의 데이터를 제로값, 논-아웃라이어, 및 아웃라이어 중 하나로 분류하여 특징 맵을 컴프레싱 할 수 있다.

먼저, 실시예에 따른 컴프레서는 논-제로 타일(423)이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일(421, 422)의 개수를 지시하는 제1 정보(예를 들어, 10(2))를 제1 정보(451)로 생성할 수 있다.

컴프레서는 논-제로 타일(423) 내의 데이터를 제로값, 논-아웃라이어, 또는 아웃라이어로 분류하여, 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보(452)를 생성할 수 있다. 앞서 설명한 바와 같이, 논-아웃라이어는 N 비트 고정 소수점 시스템에서 N/2 비트를 사용하여 표현 가능한 데이터를 나타낼 수 있고, 아웃라이어는 N/2 비트를 사용하여 표현이 불가능한 데이터를 나타낼 수 있다.

컴프레서는 논-제로 타일(423) 내의 데이터들을 구분할 수 있다. 컴프레서는 논-제로 타일(423) 내에서 제로값의 데이터(431)에 제1 값을 매핑할 수 있고, 논-아웃라이어의 데이터들(432, 433)에 제2 값을 매핑할 수 있으며, 아웃라이어의 데이터(434)에 제3 값을 매핑할 수 있다. 예를 들어, 컴프레서는 논-제로 타일(423) 내에서 제로값의 데이터(431)가 '$00_{(2)}$'을 지시하는 값으로 결정할 수 있고, 논-아웃라이어의 데이터들(432, 433)이 '$01_{(2)}$'을 지시하는 값으로 결정할 수 있으며, 아웃라이어의 데이터(434)가 '$10_{(2)}$'을 지시하는 값으로 결정할 수 있다. 컴프레서는 각 데이터가 지시하는 값을 연결하여 제2 정보(452)를 생성할 수 있다. 제2 정보(452)는 데이터를 제로값, 논-아웃라이어, 또는 아웃라이어로 구분하기 위해, 데이터를 제로값 또는 논-제로값으로 구분하기 위한 경우와 비교하여 추가적인 비트가 필요하게 된다. 그러나, 대부분의 논-제로값의 데이터는 논-아웃라이어이기 때문에, 보다 적은 비트 폭을 사용하여 제3 정보(453)를 생성할 수 있고, 제1 컴프레싱 방법(440)과 비교하여 효과적인 압축률을 나타낼 수 있다.

일 실시예에 따른 컴프레서는 논-제로 타일(423) 내에서 논-제로 값의 데이터들(432, 433, 434)에 기초하여 제3 정보(453)를 생성할 수 있다. 일 실시예에 따른 컴프레서는 아웃라이어의 데이터(434)를 제1 비트 폭의 데이터로 나타내고, 논-아웃라이어의 데이터들(432, 433)을 제1 비트 폭 보다 작은 제2 비트 폭의 데이터로 나타냄으로써 제3 정보(453)를 생성할 수 있다. 예를 들어, N 비트 고정 소수점 시스템에서, 제1 비트 폭은 N 비트 폭을 나타낼 수 있고, 제2 비트 폭은 N/2 비트 폭을 나타낼 수 있다. 일 실시예에 따르면, 컴프레서는 제3 정보(453)를 생성하는 경우, 논-아웃라이어의 데이터를 N/2 비트 폭의 데이터로 나타낼 수 있다. 컴프레서는 논-아웃라이어의 데이터를 기존 고정된 N 비트 폭의 절반 만큼의 비트 폭을 사용하여 나타낼 수 있으므로 압축 효율을 높일 수 있다.

컴프레서는 특징 맵의 단위 타일들의 데이터들을 모두 컴프레싱 할 때까지 데이터의 생성 및 저장을 반복하며, 마지막에는 종료 패킷(end of packet)(454)을 생성하고 저장한다. 종료 패킷(454)은 제로 데이터일 수 있다.

[0060] 삭제

[0061] 삭제

[0062] 삭제

[0063] 삭제

[0064] 삭제

[0065] 삭제

[0067] 도 5는 일 실시예에 따른 가속기 장치의 구성에 관하여 설명한다.

[0068] 일 실시예에 따른 가속기 장치(510)는 프로세싱 엘리먼트 어레이(processing element array, PE array)(511), 온칩 글로벌 버퍼(On-chip global buffer)(512), 제어기(controller)(513), 컴프레서(514), 및 디컴프레서(515)를 포함할 수 있다.

[0069] 프로세싱 엘리먼트 어레이(511)는 컨볼루션 뉴럴 네트워크 연산을 수행할 수 있다. 프로세싱 엘리먼트 어레이(511)의 컨볼루션 뉴럴 네트워크 연산 수행을 위하여 오프 칩 메모리(520)(예를 들어, DRAM)에서 적절한 데이터가 디컴프레서(515)에 의하여 디컴프레싱되고, 디컴프레싱된 데이터가 온칩 글로벌 버퍼(512)에 로드될 수 있다. 프로세싱 엘리먼트 어레이(511)에서 컨볼루션 뉴럴 네트워크 연산을 통하여 계산된 데이터는 온칩 글로벌 버퍼(512)로 전송되어 저장되고, 컴프레서(514)를 통하여 컴프레싱된 데이터가 오프 칩 메모리(520)로 전송되어 저장될 수 있다. 이하, 도 6에서는 컴프레서의 구조에 관하여 보다 자세히 설명하고, 도 7에서는 디컴프레서의 구조에 관하여 보다 자세히 설명한다.

[0071] 도 6은 일 실시예에 따른 컴프레서의 구조에 관하여 설명한다.

[0072] 일 실시예에 따른 컴프레서는 특징 맵을 컴프레싱 할 수 있다. 일 실시예에 따른 컴프레서(600)는 레지스터(601), 비교기(602), 버퍼(603), 및 제어기(613)를 포함할 수 있다.

가속기 장치는 입력 이미지로부터 컨볼루션 뉴럴 네트워크 모델에 포함된 하나 이상의 레이어에 기초하여 획득된 특징 맵에 대하여, 미리 결정된 크기의 단위 타일들로 특징 맵을 분할할 수 있다. 일 실시예에 따른 컴프레서의 레지스터(601)는 특징 맵의 분할된 단위 타일들을 순차적으로 입력 받을 수 있다. 레지스터(601)에 단위 타일이 포함하는 데이터들이 입력될 수 있다. 비교기(602)는 레지스터에 입력되어 저장된 데이터들을 비교할 수 있다. 이하에서는, 일 실시예에 따른 제2 압축 방법(450)을 컴프레서(600)의 구조와 함께 설명한다.

첫 번째 주기(cycle)에서, 컴프레서의 병렬 레지스터(601)에 단위 타일의 데이터들(Din_0, Din_1, ..., Din_{k-1})이 입력되어 저장될 수 있다. 일 실시예에 따르면, 컴프레서의 병렬 레지스터(601)에 데이터가 상위 비트 데이터(641)와 하위 비트 데이터(642)로 나누어져 입력될 수 있다. 컴프레서의 제어기가 데이터를 상위 비트 데이터(641) 및 하위 비트 데이터(642)로 나누어 레지스터(601)에 입력할 수 있다. 상위 비트 데이터(641)는 N/2 비트 위치로부터 최상위 비트(most significant bit, MSB) 위치까지의 비트 위치에 대응하는 데이터를 나타낼 수 있다. 상위 비트 데이터(641)는 N 비트 고정 소수점 시스템에서 데이터를 N 비트로 표현하는 경우, 상위 N/2 비트에 대응하는 데이터를 나타낼 수 있다. 하위 비트 데이터(642)는 최하위 비트(least significant bit, LSB)로부터 N/2 비트 위치까지의 비트 위치에 대응하는 데이터를 나타낼 수 있다. 하위 비트 데이터(642)는 데이터를 N 비트로 표현하는 경우, 하위 N/2 비트에 대응하는 데이터를 나타낼 수 있다.

비교기(602)는 하나의 데이터의 상위 비트 데이터(641) 및 하위 비트 데이터(642)가 각각 제로값인지 여부를 판단할 수 있다. 제어기(613)는 판단 결과에 기초하여 데이터를 아웃라이어, 논-아웃라이어, 또는 제로값 중 하나로 분류할 수 있다. 컴프레서의 제어기는 비교 결과에 기초하여, 하나의 데이터의 상위 비트 데이터(641) 및 하위 비트 데이터(642)가 모두 논-제로 값인 경우, 하나의 데이터를 아웃라이어로 분류할 수 있다. 마찬가지로, 컴프레서의 제어기는 하나의 데이터의 상위 비트 데이터(641)가 제로값이고, 하위 비트 데이터(642)가 논-제로 값인 경우, 하나의 데이터를 논-아웃라이어로 분류하며, 하나의 데이터의 상위 비트 데이터(641) 및 하위 비트 데이터(642)가 모두 제로 값인 경우 하나의 데이터를 제로값으로 분류할 수 있다.

다음 주기에서, 일 실시예에 따른 컴프레서(600)는 입력된 데이터들 중 논-제로 값의 데이터에 기초하여 추가된

데이터(631)를 생성할 수 있다. 제어기는 데이터들 중 제로값을 제거하는 시프트 앤 어팬드(shift-and-append) 로직을 통하여 논-제로 값을 순차적으로 추출할 수 있다. 컴프레서는 먹스(mux)(604)를 사용하여 아웃라이어로 분류된 데이터는 제1 비트 폭 데이터로 나타내고, 논-아웃라이어로 분류된 데이터는 제1 비트 폭 보다 작은 제2 비트 폭의 데이터로 나타냄으로써 제3 정보(630)를 생성할 수 있다.

마지막 주기에서, 컴프레서는 제1 정보(610), 제2 정보(620), 제3 정보(630)를 버퍼(603)에 저장할 수 있다. 일 실시예에 따른 컴프레서의 제어기는 비교기의 비교 결과에 따라 제로값으로 분류된 데이터를 제1 값에 대응시키고, 논-아웃라이어로 분류된 데이터를 제2 값에 대응시키며, 아웃라이어로 분류된 데이터를 제3 값에 대응시킴으로써 제2 정보(620)를 생성할 수 있다. 제어기는 비교기의 비교 결과에 따라 제로값으로 분류된 데이터에 제1 값을 매핑하고, 논-아웃라이어로 분류된 데이터에 제2 값을 매핑하며, 아웃라이어로 분류된 데이터에 제3 값을 매핑하여 제2 정보(620)를 생성할 수 있다. 제1 값, 제2 값, 및 제3 값은 서로 상이한 데이터일 수 있다. 예를 들어, 제1 값은 '$00_{(2)}$'일 수 있고, 제2 값은 '$01_{(2)}$'일 수 있으며, 제3 값은 '$10_{(2)}$'일 수 있다. 각 데이터가 지시하는 값을 순차적으로 연결함으로써 제2 정보(620)를 생성할 수 있다. 여기서, 제2 정보의 비트 수는 단위 타일이 포함하는 데이터들의 수의 2배일 수 있다.

일 실시예에 따른 컴프레서의 제어기는 입력된 단위 타일을 제로 타일 또는 논-제로 타일로 분류할 수 있다. 컴프레서의 제어기는 단위 타일의 입력된 데이터들이 모두 제로값인 경우에 입력된 단위 타일이 제로 타일인 것으로 판단할 수 있다. 보다 구체적으로, 제어기는 제2 정보(620)를 사용하여 입력된 단위 타일이 제로 타일인지 판단할 수 있다. 컴프레서(600)는 입력된 단위 타일 내에서 각 데이터가 지시하는 값의 합(611)이 제로값인지 여부를 판단할 수 있다. 컴프레서(600)는 각 데이터가 지시하는 값의 합(611)이 제로값인 경우, 입력된 단위 타일이 제로 타일인 것으로 판단할 수 있다. 컴프레서는 각 데이터가 지시하는 값의 합이 논-제로 값인 경우, 입력된 단위 타일이 논-제로 타일인 것으로 판단할 수 있다. 일 실시예에 따른 컴프레서는 입력된 단위 타일이 제로 타일인 것으로 판단된 경우, 논-제로 타일이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일의 개수를 지시하는 카운트(612)를 '1'만큼 증가시킬 수 있다. 다시 말해, 컴프레서는 입력된 단위 타일의 각 데이터가 지시하는 값의 합(611)이 제로값인 경우, 카운트(612)를 '1'만큼 증가시킬 수 있다. 제어기(613)는 입력된 단위 타일의 각 데이터가 지시하는 값의 합(611)이 논-제로 인 경우, 입력된 단위 타일이 논-제로 타일인 것으로 판단할 수 있다. 컴프레서의 제어기는 논-제로 타일이 레지스터에 입력된 경우, 제1 정보의 카운트(612)에 저장된 값에 따라 제1 정보(610)를 생성하고, 생성된 제1 정보(610)를 버퍼(603)에 저장할 수 있다. 또한, 컴프레서는 논-제로 타일이 레지스터에 입력된 경우, 생성된 제2 정보(620)를 버퍼(603)에 저장할 수 있고, 제3 정보(630)를 버퍼(603)에 저장할 수 있다.

단위 타일을 컴프레싱하기 위한 지연 시간은 3 내지 4+k 사이클(cycle)로 다양하게 나타난다. 여기서, k는 단위 타일에 포함되는 데이터의 수를 나타낼 수 있다. 최소 주기는 제로 타일이 레지스터에 입력되는 경우에 나타나며, 최대 주기는 단위 타일의 모든 데이터들이 논-제로 값인 경우에 나타난다. 컴프레서가 n개 사용되는 경우에 특징 맵을 컴프레싱 하기 위한 시간은 n 배 감소한다. 여기서, n은 자연수를 나타낼 수 있다.

[0073] 삭제

[0074] 삭제

[0075] 삭제

[0076] 삭제

[0077] 삭제

[0078] 삭제

[0079] 삭제

[0081] 도 7은 일 실시예에 따른 특징 맵을 디컴프레싱하는 디컴프레서의 구조에 관하여 설명한다.

[0082] 일 실시예에 따른 특징 맵을 디컴프레싱하는 디컴프레서(700)(이하, '디컴프레서')는 CNN 모델의 레이어에 기초하여 획득된 특징 맵이 컴프레싱된 데이터를 디컴프레싱 할 수 있다. 디컴프레서의 레지스터에는 특징 맵이 컴프레싱된 데이터가 입력될 수 있다.

첫 번째 주기에서 제1 정보(710), 제2 정보(720)가 동시에 디컴프레서(700)의 레지스터로 입력되어 저장될 수 있다. 다음 주기에서 제3 정보(730)가 디컴프레서(700)의 레지스터로 입력되어 저장될 수 있다. 디컴프레서(700)는 논-제로 값의 데이터들을 나타내는 제3 정보(730)를 입력되는 순서대로 레지스터에 저장할 수 있다. 다음 주기에서 디컴프레서의 제어기(713)는 제1 정보(710)에 대응하는 개수의 제로 타일을 생성하여 버퍼로 전송할 수 있다. 마지막 주기에서 디컴프레서의 제어기는 제2 정보(720) 및 제3 정보(730)를 사용하여 논-제로 타일을 생성하고, 버퍼로 전송할 수 있다. 디컴프레서의 제어기(713)는 제2 정보(720)를 이용하여 논-제로 타일 내에서 제로값의 데이터와 논-제로 값의 데이터를 구분할 수 있다. 디컴프레서의 제어기(713)는 제로값의 데이터는 제로값과 대응시키고, 논-제로 값의 데이터는 제3 정보(730)와 순차적으로 대응시켜 논-제로 타일을 생성할 수 있다. 디컴프레서는 종료 패킷이 전송될 때까지 위의 순서를 반복한다.

특징 맵 컴프레싱 데이터를 디컴프레싱 하기 위한 지연 시간은 3 내지 3+l 사이클(cycle)로 다양하게 나타난다. 여기서, l은 제1 정보의 최대 크기(magnitude)를 나타낼 수 있다. 최소 주기는 종료 패킷을 디컴프레싱하는 경우에 나타나며, 최대 주기는 제1 정보가 최대 크기를 가지는 경우에 나타날 수 있다. 디컴프레서가 n개 사용되는 경우에 특징 맵 컴프레싱 데이터의 디컴프레싱을 위한 시간은 n 배 감소한다. 여기서, n은 자연수를 나타낼 수 있다.

[0083] 삭제

[0084] 삭제

[0086] 아래 표 1은 일 실시예에 따른 컴프레서와 디컴프레서에 대한 하드웨어의 크기 및 전력 소모에 대하여 나타낸다.

표 1

[0087]

장치	크기[μm^2]	전력 소모[mW]
도 6에 따른 컴프레서	11,124	0.488
도 7에 따른 디컴프레서	4,060	0.24
MAC(multiply-accumulate)	1,283	0.248
256 MACs	328,448	63.4888

[0088] 표 1에서는 일 실시예에 따른 컴프레서 및 디컴프레서에 대한 하드웨어의 크기 및 전력 소모와 함께, CNN 가속기 장치 내 MAC(multiply-accumulate) 유닛에 대한 크기 및 전력 소모, 일반적인 CNN 가속기 장치에서 사용되는 개수(예를 들어, 256개) 만큼의 MAC 유닛 전체의 크기 및 전력소모에 대하여 나타낸다. MAC 유닛은 8-비트 곱셈기와 28-비트 누산기로 이루어져 있다. 일 실시예에 따른 컴프레서 및 디컴프레서의 크기 및 전력 소모는 한 개의 MAC 유닛과 비슷한 수준을 보인다. MAC 유닛은 CNN 가속기 장치에서 아주 작은 부분을 차지 하고 있기 때문에, 일 실시예에 따른 컴프레서 또는 디컴프레서의 크기와 전력소모는 아주 작다고 할 수 있다. 일 실시예에 따른 컴프레서 또는 디컴프레서의 크기와 전력소모는 일반적인 CNN 가속기 장치와 비교하여 크기는 4.62%, 전력 소모는 1.15% 정도에 불과하다.

[0090] 도 8은 단위 타일의 크기에 따른 정규화된 DRAM 액세스의 역수에 대한 그래프를 도시한다.

[0091] 이하에서는, 일 실시예에 따른 가속기 장치가 컴프레서 또는 디컴프레서를 사용하여 특징 맵을 컴프레싱 또는 디컴프레싱한 결과에 관하여 설명한다. VGG16, ExtractionNet, 및 ResNet-18의 세가지 CNN 네트워크를 사용하

여 특징 맵이 추론(inference)될 수 있다. 3개의 CNN 특징 맵은 아래 수학식 3을 사용하여 k 비트로 양자화될 수 있다.

수학식 3

[0092] $$Quant(x) = round(\frac{x(2^{k-1}-1)}{max(|x|)}) \cdot \frac{max(|x|)}{2^{k-1}-1}$$

[0093] 특징 맵을 컴프레싱하는 방법에 대한 압축률(Compression ratio)은 아래 수학식 4와 같이 정의된다.

수학식 4

[0094] $$Compression\,ratio = \frac{data\ size\ before\ compression}{data\ size\ after\ compression}$$

[0095] 압축률(Compression ration)은 양의 유리수로 나타나되, 압축되지 않은 데이터의 압축률은 '1'로 나타난다.

[0096] 일 실시예에 따른 특징 맵을 컴프레싱하는 방법 또는 특징 맵을 디컴프레싱하는 방법에 대한 압축 효과(compression effect)는 정규화된 DRAM 액세스의 역수를 사용하여 평가될 수 있다. 도 8은 세가지의 CNN 네트워크에서 단위 타일의 크기에 따른 정규화된 DRAM 액세스의 역수에 대한 그래프를 도시한다. 그래프(810)은 VGG-16 네트워크에서 단위 타일의 크기($m \times n$)에 따른 정규화된 DRAM 액세스의 역수에 대한 그래프를 나타낸다. 그래프(820)은 ExtractionNet 네트워크에서 단위 타일의 크기에 따른 정규화된 DRAM 액세스의 역수에 대한 그래프를 나타낸다. 그래프(830)은 ResNet-18 네트워크에서 단위 타일의 크기에 따른 정규화된 DRAM 액세스의 역수에 대한 그래프를 나타낸다.

각 DRAM 액세스는 다양한 비트로 양자화된 특징 맵(예를 들어, 4 비트 내지 8 비트로 양자화된 특징 맵)의 평균 액세스의 수를 나타내고, 각 평균 액세스의 수는 압축 전의 데이터 크기로 정규화 되었다. VGG-16 네트워크에서는 m 및 n이 2인 경우에 DRAM 액세스의 수가 가장 작게 나타난다. VGG-16 네트워크에서 m 또는 n이 2보다 큰 값을 가지는 경우에는, DRAM 액세스의 수가 증가하는 것으로 나타난다. 큰 단위 타일을 사용하여 특징 맵을 분할하는 경우, 제로 타일로 많은 제로값을 한 번에 컴프레싱 할 수 있으나, 큰 단위 타일 내에서 데이터들이 모두 제로값인 경우는 거의 없기 때문에 제로 타일의 수가 급격히 감소하여 압축률이 좋지 않게 나타난다. 이러한 경향성의 원인은, 도 2를 참조하면, VGG-16 네트워크에서 평균적인 공간적 상호 관계성(Moran의 I)이 데이터들 사이의 거리가 '3'이상인 경우에 크게 감소하기 때문에 발생한다. 따라서, 데이터들 사이의 거리, 즉 단위 타일의 가로 길이 또는 세로 길이가 '3' 이상인 경우, 제로 타일의 개수가 줄어든다. 더 나아가, 크기가 큰 단위 타일일수록 많은 비트 폭(bit width)의 제2 정보를 필요로 하기 때문에, 단위 타일의 크기는 공간적 상호 관계성을 고려하여 결정되어야 한다. 결과적으로 DRAM 액세스의 수는 VGG-16 네트워크에서 0.36으로 m=2, n=2에서 최소이다. 또한, 그래프(820) 및 그래프(830)을 참조하면, ExtractionNet 네트워크 및 ResNet-18 네트워크에서도 DRAM 액세스의 수는 각각 0.37 및 0.34로 m=2, n=2에서 최소이다.

[0097] 삭제

[0099] 도 9는 다양한 컴프레싱 방법에 따른 특징 맵의 압축률 및 희소성(sparsity)에 관한 그래프를 나타낸다.

[0100] 그래프(910)는 VGG-16 네트워크에서 컴프레싱 방법에 따른 각 레이어에 대한 특징 맵의 압축률 및 희소성에 관한 그래프를 나타낸다. 그래프(920)는 ExtractionNet 네트워크에서 컴프레싱 방법에 따른 각 레이어에 대한 특징 맵의 압축률 및 희소성에 관한 그래프를 나타낸다. 그래프(930)는 ResNet-18 네트워크에서 컴프레싱 방법에 따른 각 레이어에 대한 특징 맵의 압축률 및 희소성에 관한 그래프를 나타낸다.

[0101] 그래프(911), 그래프(912), 그래프(913), 그래프(914), 및 그래프(915)는 VGG-16 네트워크에서 각각 RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440), 제2 컴프레싱 방법(450)에 따른 레이어에 대한 특징 맵의 압축률을 나타낸

다. 그래프(916)은 VGG-16 네트워크에서 레이어에 대한 특징 맵의 희소성에 관한 그래프를 나타낸다.

[0102] RLC(Run-length Compression, 런 렝스 부호화)는 특징 맵 내에서 연속적인 제로 타일들을 하나의 런(run) 데이터로 표현함으로써 DRAM 액세스를 낮추는 방법이다. RLC는 런 데이터 뒤에 제로값이 아닌 데이터를 저장한다. RLC4 및 RLC8은 각각 4 비트의 런 데이터 및 8 비트의 런 데이터를 사용하는 컴프레싱 방법을 나타낸다. ZVC(Zero-value Compression)은 특징 맵 내에서 제로값이 아닌 데이터를 저장하고, 해당 데이터의 위치값을 나타내는 마스크(mask) 데이터를 저장하는 방법이다.

[0103] 도 9를 참조하면, 압축률은 특징 맵의 희소성과 상당한 연관관계를 나타낸다. 뒤 쪽의 레이어에서는 특징 맵의 희소성이 상승하게 된다. 따라서 레이어가 더 깊을수록, 모든 알고리즘의 압축률은 증가하는 것으로 나타난다. 일 실시예에 따른 제2 컴프레싱 방법(450)은 2.77의 가장 높은 평균 압축률을 나타내며, RLC8, RLC4, ZVC의 평균 압축률은 각각 1.60, 1.86 및 2.21로 나타난다.

[0104] 특정 레이어에서 특징 맵의 낮은 희소성으로 인해 RLC8 및 RLC4의 압축률은 1.0 이하로 나타난다. 그러나, ZVC, 제1 컴프레싱 방법(440) 및 제2 컴프레싱 방법(450)에 따른 압축률은 항상 1.0 이상으로 나타난다. 8 번째 레이어부터는 제1 컴프레싱 방법 또는 제2 컴프레싱 방법의 압축률과 ZVC의 압축률 간의 차이가 점진적으로 증가하는 것으로 나타나는데, 이는 특징 맵의 희소성이 상승하였을 때, 제1 컴프레싱 방법 또는 제2 컴프레싱 방법은 논-제로 타일을 이용하여 효과적으로 제로값을 압축할 수 있기 때문이다. 비록 몇 개의 레이어(예를 들어, 2, 3, 4, 및 7번째 레이어)에서 ZVC가 제1 컴프레싱 방법(440)에 비하여 더 높은 압축률을 갖는 것으로 나타나나, 해당 레이어에서 ZVC와 제1 컴프레싱 방법(440) 간의 압축률의 차이는 0.04 이하로 아주 작게 나타난다.

[0105] 그래프(921), 그래프(922), 그래프(923), 그래프(924), 및 그래프(95)는 ExtractionNet 네트워크에서 각각 RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440), 제2 컴프레싱 방법(450)에 따른 레이어에 대한 특징 맵의 압축률을 나타낸다. 그래프(926)은 ExtractionNet 네트워크에서 레이어에 대한 특징 맵의 희소성에 관한 그래프를 나타낸다. ExtractionNet 네트워크에서도 제2 컴프레싱 방법(450)에 따른 압축률이 가장 좋은 압축률을 나타낸다. ExtractionNet에 대한 RLC8, RLC4, ZVC 및 제2 컴프레싱 방법(450)의 평균 압축 비율은 각각 1.73, 1.97, 2.32 및 2.84로 나타난다.

[0106] 그래프(931), 그래프(932), 그래프(933), 그래프(934), 및 그래프(935)는 ResNet-18 네트워크에서 각각 RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440), 제2 컴프레싱 방법(450)에 따른 레이어에 대한 특징 맵의 압축률을 나타낸다. 그래프(926)은 ResNet-18 네트워크에서 레이어에 대한 특징 맵의 희소성에 관한 그래프를 나타낸다. RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440) 및 제2 컴프레싱 방법(450)의 평균 압축률은 각각 2.21, 2.36, 2.66, 3.30 및 3.52로 나타난다. ResNet-18의 희소성이 출렁이는 이유는 숏컷-레이어(shortcut-layer)의 영향이다.

도 10은 양자화된 데이터에 대하여 컴프레싱된 특징 맵의 정규화된 DRAM 액세스의 값을 나타낸다.

그래프(1010), 그래프(1020), 그래프(1030)는 각각 VGG-16 네트워크, Extraction 네트워크, ResNet 네트워크에서 4비트부터 8비트까지 양자화된 데이터에 대하여 컴프레싱된 특징 맵의 정규화된 DRAM 액세스 값을 나타낸다.

[0108] 삭제

[0109] 삭제

[0110] 그래프(1011), 그래프(1012), 그래프(1013), 그래프(1014), 그래프(1015)는 VGG-16 네트워크에서 각각 RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440), 제2 컴프레싱 방법(450)에 따른 컴프레싱된 특징 맵의 정규화된 DRAM 액세스의 값에 대한 그래프를 나타낸다.

[0111] 4비트부터 8비트까지 양자화된 데이터에 대하여 컴프레싱된 특징 맵의 정규화된 DRAM 액세스 값이 측정된다. 4비트로 양자화된 특징 맵을 제외하고, 제2 컴프레싱 방법(450)은 가장 낮은 정규화된 DRAM 액세스 값을 나타낸다. 제1 컴프레싱 방법(440)이 4비트로 양자화된 특징 맵에서 가장 높은 DRAM 액세스 값의 감소를 나타낸다. 8비트로 양자화된 특징 맵에서 제2 컴프레싱 방법(450)은 50%의 DRAM 액세스 값의 감소를 나타내고, ZVC, RLC4,

및 RLC8은 각각 39%, 20%, 10%의 DRAM 액세스 값의 감소를 나타낸다. 8비트 이하로 특징맵이 양자화되는 경우 컴프레싱 방법에 상관 없이 DRAM 액세스 값이 감소한다. 이는, 특징 맵의 희소성이 증가함으로써 압축률이 상승하기 때문이다. 이러한 DRAM 액세스 값의 감소는, ZVC가 고정된 크기의 마스크(mask) 데이터를 사용하기 때문에, ZVC에서 가장 작게 나타난다. 제1 컴프레싱 방법(440) 및 제2 컴프레싱 방법(450)은 양자화된 특징 맵을 효과적으로 컴프레싱 할 수 있다. 특히, 4비트로 양자화된 특징맵에서 제1 컴프레싱 방법(440) 및 제2 컴프레싱 방법(450)은 DRAM 액세스 값을 82% 및 80%로 줄일 수 있다. 또한, 제1 컴프레싱 방법(440)이 4비트로 양자화된 특징맵에서 더 작은 DRAM 액세스 값을 나타내는데, 이는 특징 맵이 충분히 희소성이 높고 제2 컴프레싱 방법(450)은 데이터를 아웃라이어, 논-아웃라이어, 제로값으로 표현하기 위하여 비트 폭이 더 큰 제2 정보(마스크)를 사용하기 때문이다.

[0112] 그래프(1011), 그래프(1012), 그래프(1013), 그래프(1014), 그래프(1015)는 ExtractionNet 네트워크에서 각각 RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440), 제2 컴프레싱 방법(450)에 따른 컴프레싱된 특징 맵의 정규화된 DRAM 액세스의 값에 대한 그래프를 나타낸다.

[0113] 그래프(1011), 그래프(1012), 그래프(1013), 그래프(1014), 그래프(1015)는 ResNet-18 네트워크에서 각각 RLC8, RLC4, ZVC, 제1 컴프레싱 방법(440), 제2 컴프레싱 방법(450)에 따른 컴프레싱된 특징 맵의 정규화된 DRAM 액세스의 값에 대한 그래프를 나타낸다.

[0114] ExtractionNet 및 ResNet-18에서 DRAM 액세스 값은 VGG-16과 유사한 경향을 나타낸다. 8비트로 양자화된 특징맵에서 제2 컴프레싱 방법(450)은 DRAM 액세스 값을 각각 54 % 및 57 % 감소시키며, 제1 컴프레싱 방법은 DRAM 액세스 값을 각각 48 % 및 50 % 감소시킨다. 4 비트로 양자화된 특징맵에서 제2 컴프레싱 방법(450)은 DRAM 액세스 값을 각각 76 % 및 78 % 감소시키고, 제1 컴프레싱 방법(440)은 DRAM 액세스 값을 각각 78 % 및 80 % 감소시킨다.

[0116] 도 11은 특징 맵을 저장 및 로드(load)하는데 소비되는 에너지에 관한 그래프를 나타낸다.

[0117] 그래프(1111), 그래프(1121) 및 그래프(1131)은 VGG-16, ExtractionNet, ResNet-18 네트워크에서 RLC4를 사용하여 특징 맵을 컴프레싱하는 경우, 특징 맵을 저장 및 로드하는데 소비되는 에너지에 관한 그래프를 나타낸다. 그래프(1112), 그래프(1122) 및 그래프(1132)은 VGG-16, ExtractionNet, ResNet-18 네트워크에서 ZVC를 사용하여 특징 맵을 컴프레싱하는 경우, 특징 맵을 저장 및 로드하는데 소비되는 에너지에 관한 그래프를 나타낸다. 그래프(1113), 그래프(1123) 및 그래프(1133)은 VGG-16, ExtractionNet, ResNet-18 네트워크에서 일 실시예에 따른 컴프레싱 방법을 사용하여 특징 맵을 컴프레싱하는 경우, 특징 맵을 저장 및 로드하는데 소비되는 에너지에 관한 그래프를 나타낸다.

[0118] 에너지 소비는 컴프레서가 없는 CNN 가속기 장치에 대하여 정규화된다. 앞서 설명한 바와 같이, DRAM 액세스는 대부분의 에너지 소비를 담당한다. 따라서 일 실시예에 따른 컴프레싱 방법은 하드웨어의 에너지 소비량이 RLC4 및 ZVC보다 높더라도, 특징 맵을 저장하고 로드 하기 위한 에너지 소비량이 RLC4 및 ZVC 보다 감소한다. 일 실시예에 따른 컴프레싱 방법은 VGG-16, ExtractionNet 및 ResNet-18 네트워크에서 DRAM 액세스를 위한 에너지 소비를 각각 0.35, 0.38 및 0.40으로 감소시킨다. 결과적으로, 일 실시예에 따른 컴프레싱 방법은 평균 37.7 %의 에너지 소비를 감소시킬 수 있다.

[0120] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 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 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0121] 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 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0122] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

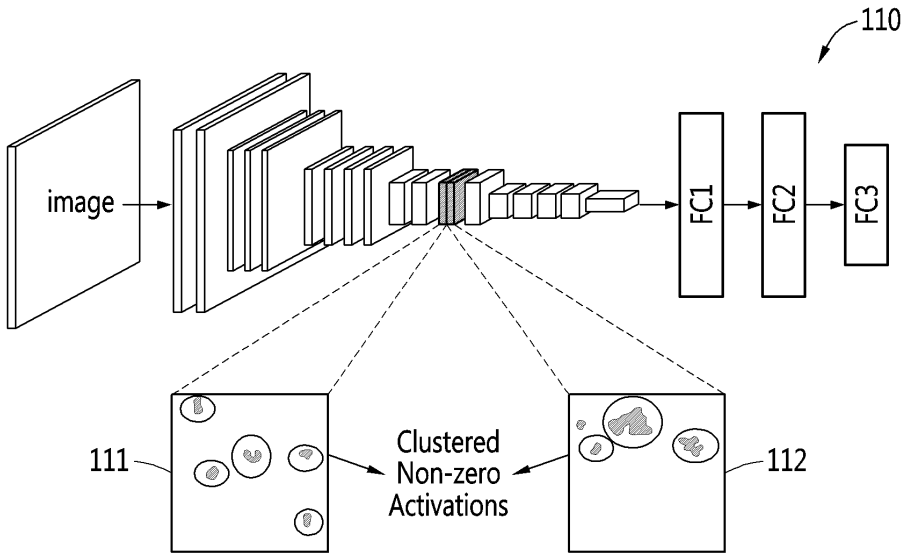
[0123] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

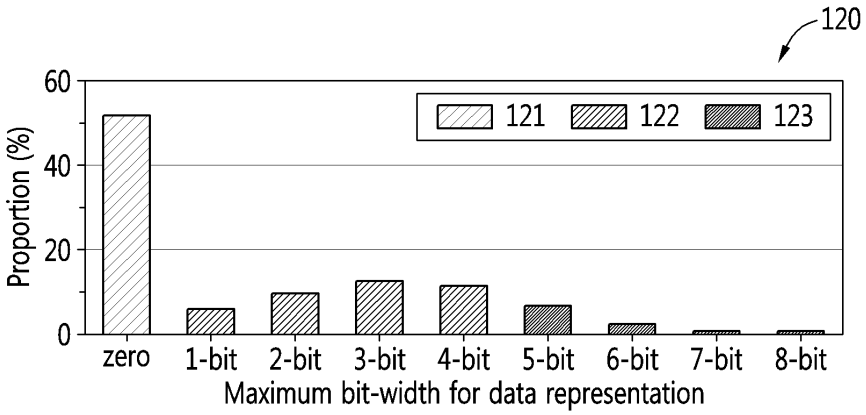
[0124] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0125] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

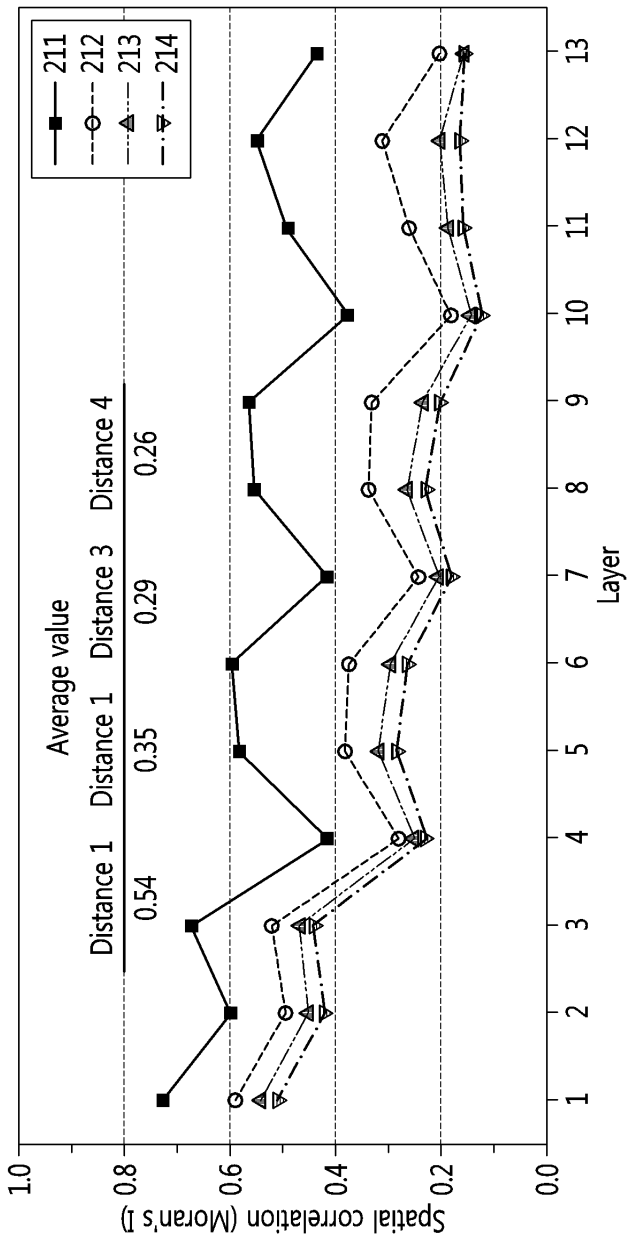
도면

도면1

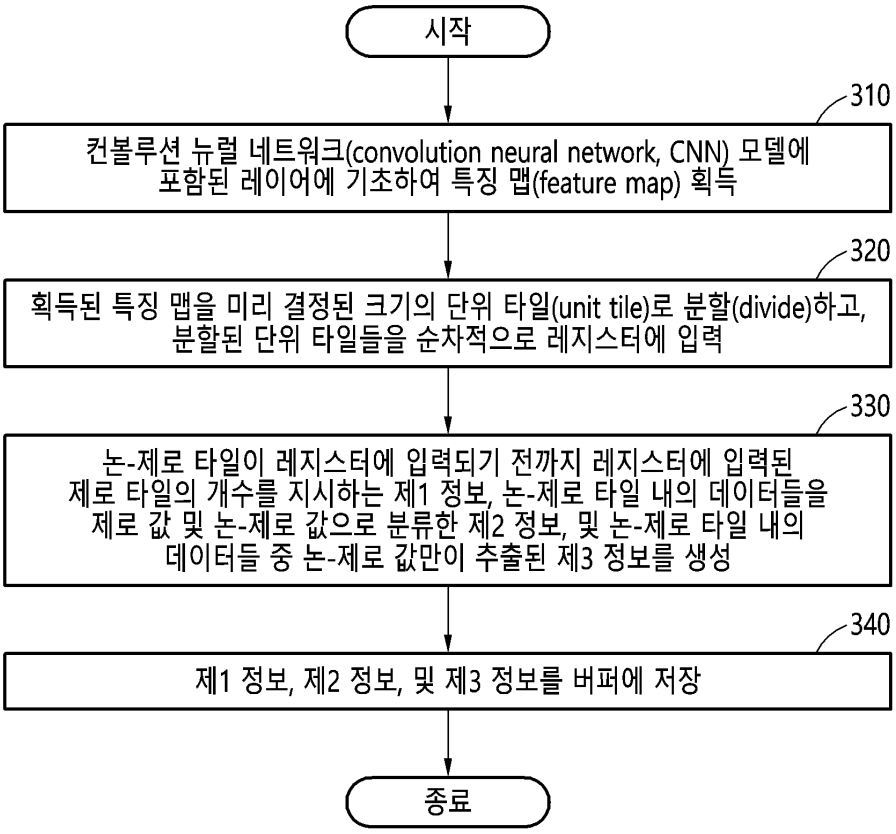

110
image
FC1
FC2
FC3
111
Clustered
Non-zero
Activations
112


120
60
40
20
0
Proportion (%)
121
122
123
zero
1-bit
2-bit
3-bit
4-bit
5-bit
6-bit
7-bit
8-bit
Maximum bit-width for data representation

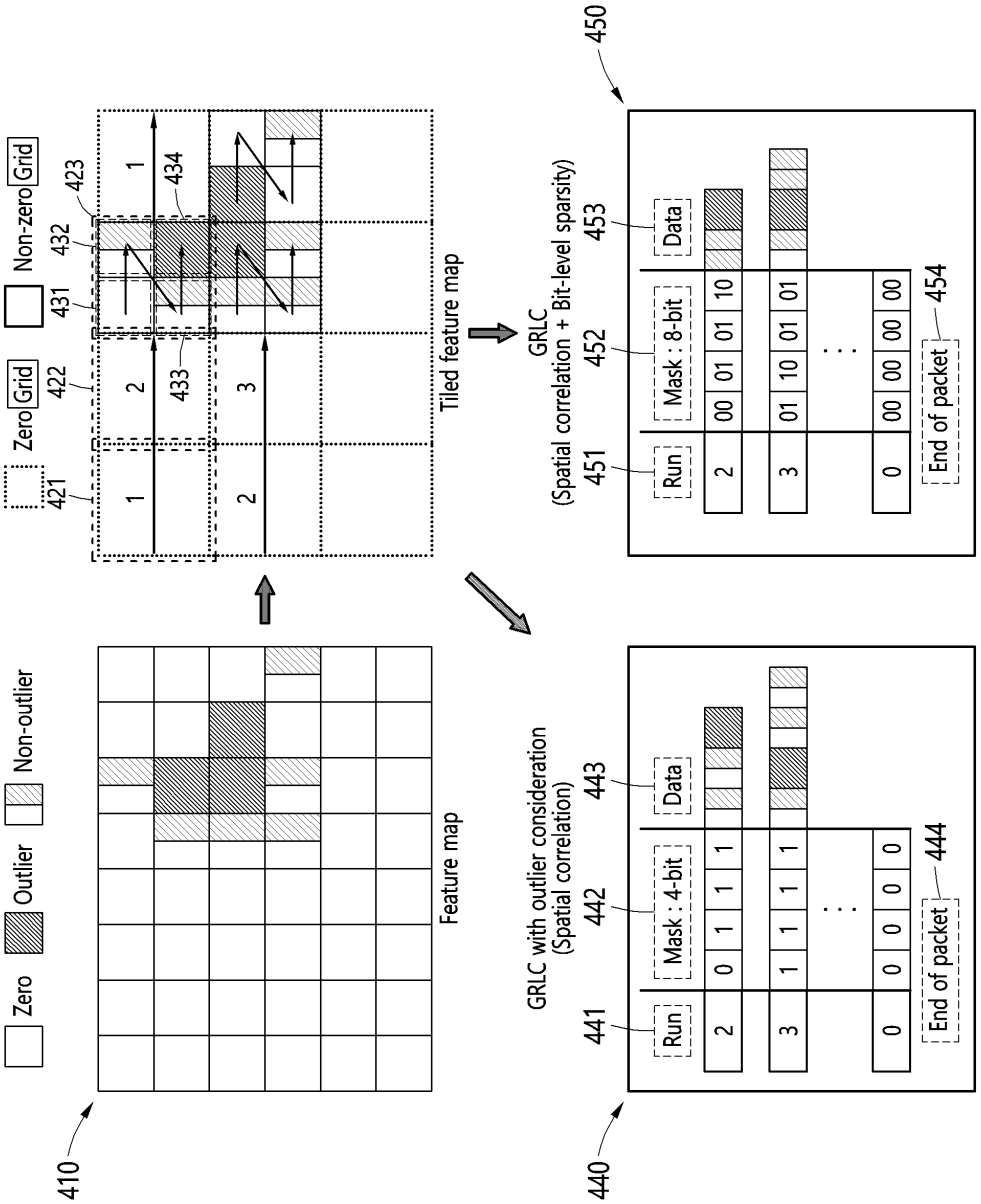
도면2


211
212
213
214
Average value
Distance 1 Distance 1 Distance 3 Distance 4
0.54 0.35 0.29 0.26
Spatial correlation (Moran's I)
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13
Layer

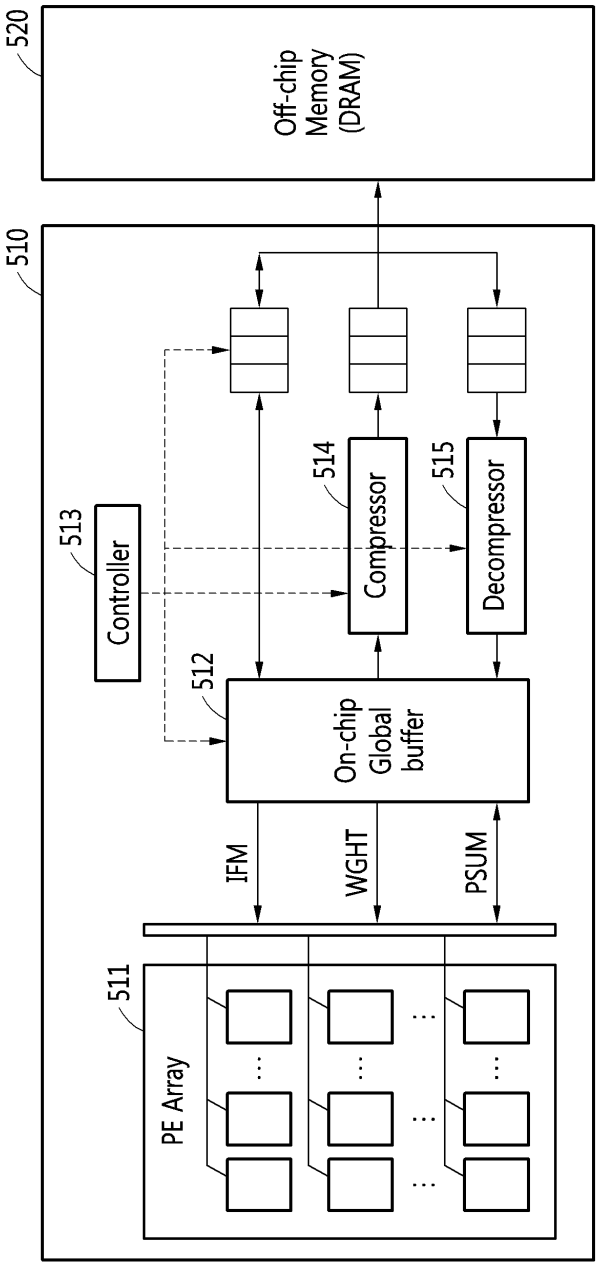
도면3


시작
310
컨볼루션 뉴럴 네트워크(convolution neural network, CNN) 모델에 포함된 레이어에 기초하여 특징 맵(feature map) 획득
320
획득된 특징 맵을 미리 결정된 크기의 단위 타일(unit tile)로 분할(divide)하고, 분할된 단위 타일들을 순차적으로 레지스터에 입력
330
논-제로 타일이 레지스터에 입력되기 전까지 레지스터에 입력된 제로 타일의 개수를 지시하는 제1 정보, 논-제로 타일 내의 데이터들을 제로 값 및 논-제로 값으로 분류한 제2 정보, 및 논-제로 타일 내의 데이터들 중 논-제로 값만이 추출된 제3 정보를 생성
340
제1 정보, 제2 정보, 및 제3 정보를 버퍼에 저장
종료

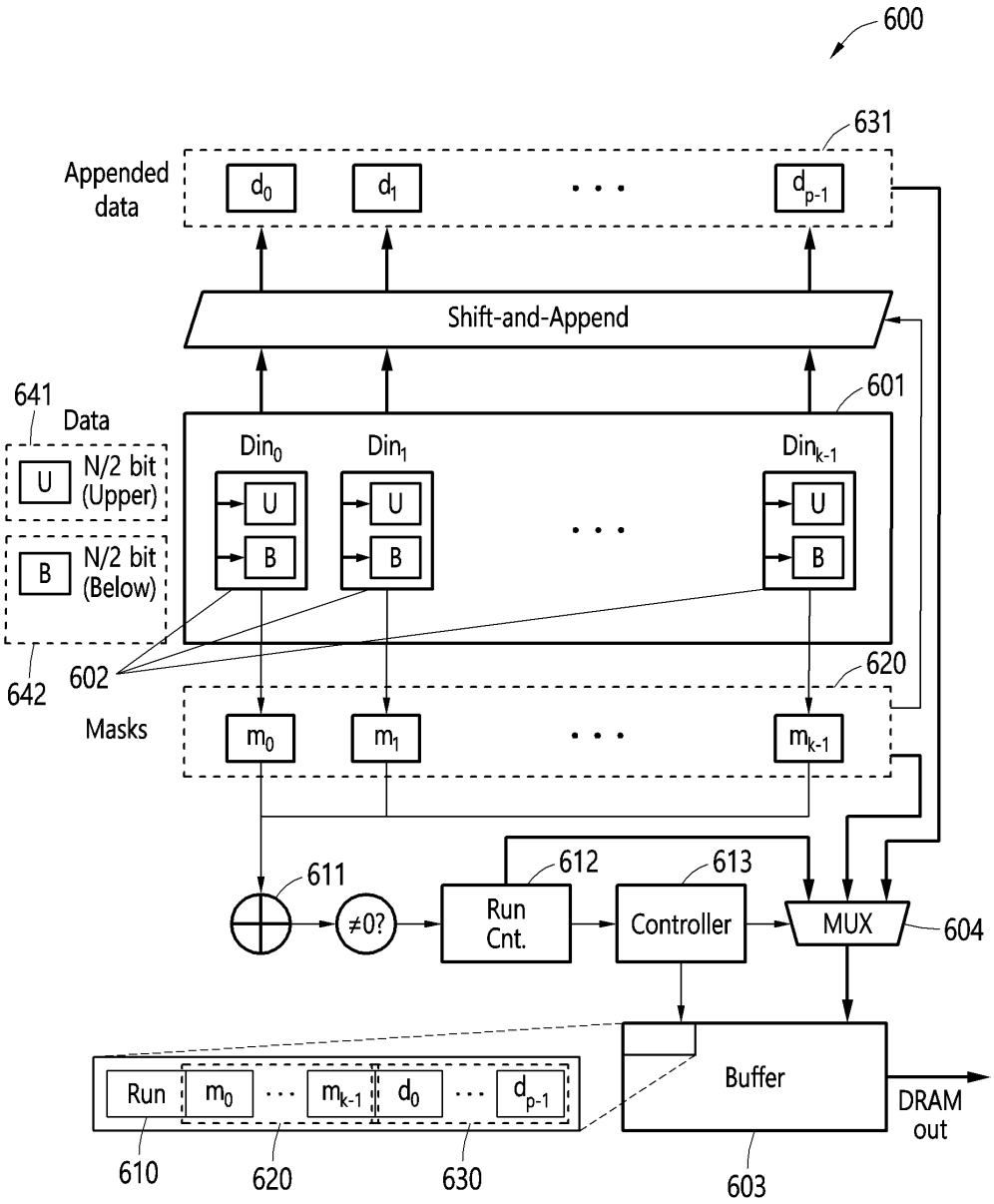
도면4


Zero
Outlier
Non-outlier
Feature map
410
Zero Grid
Non-zero Grid
421
422
423
431
432
433
434
Tiled feature map
GRLC with outlier consideration (Spatial correlation)
440
441
442
443
444
Run
Mask : 4-bit
Data
End of packet
GRLC (Spatial correlation + Bit-level sparsity)
450
451
452
453
454
Run
Mask : 8-bit
Data
End of packet

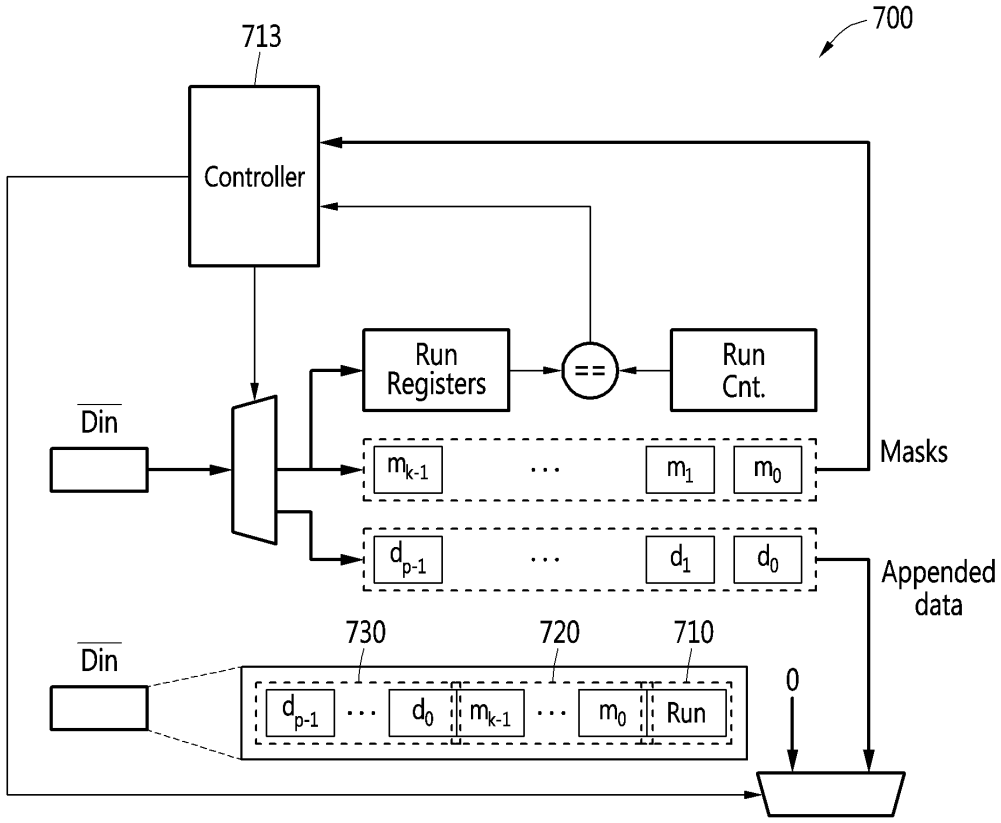
도면5


520
Off-chip Memory (DRAM)
510
513
Controller
514
Compressor
515
Decompressor
512
On-chip Global buffer
IFM
WGHT
PSUM
511
PE Array

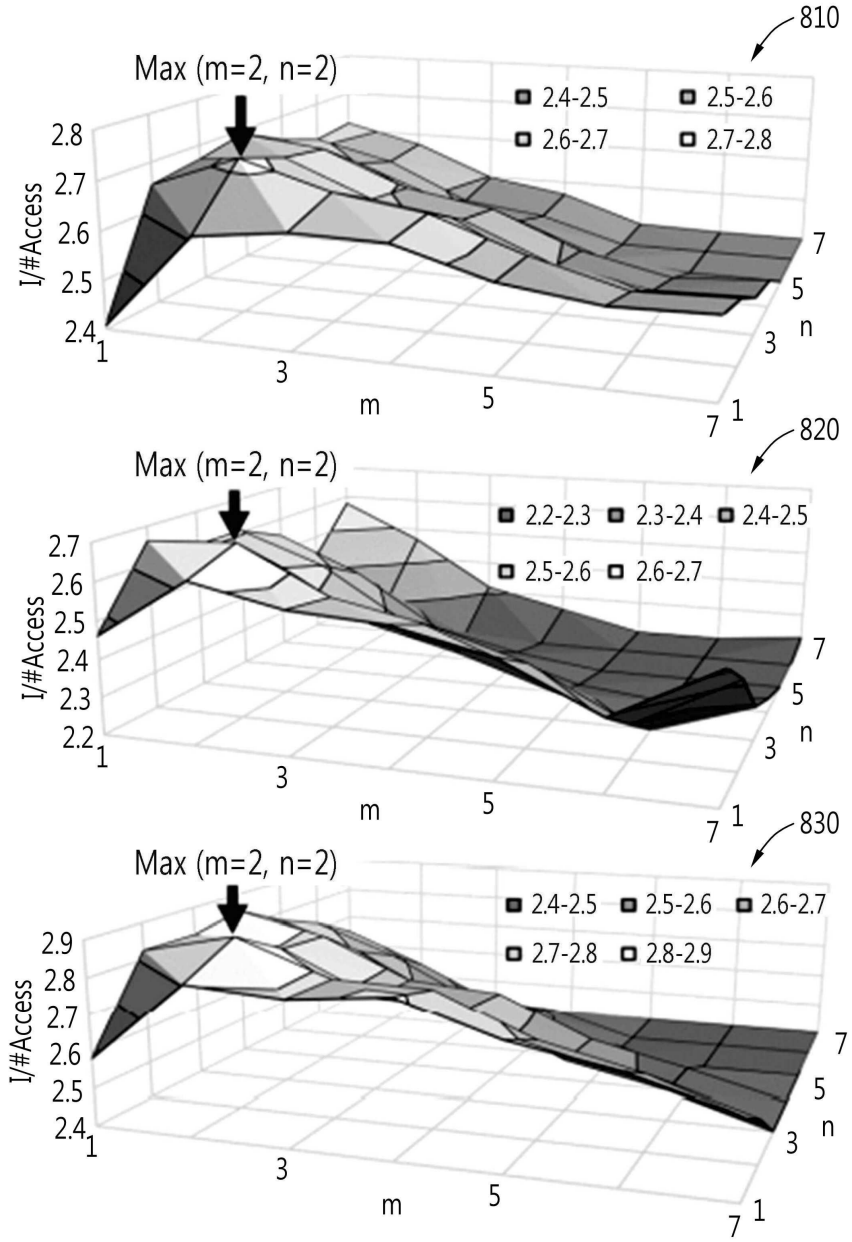
도면6


600
631
Appended data
d_0
d_1
d_{p-1}
Shift-and-Append
641
Data
U
N/2 bit (Upper)
B
N/2 bit (Below)
642
602
601
Din_0
Din_1
Din_{k-1}
620
Masks
m_0
m_1
m_{k-1}
611
≠0?
612
Run Cnt.
613
Controller
MUX
604
Buffer
DRAM out
Run
610
620
630
603

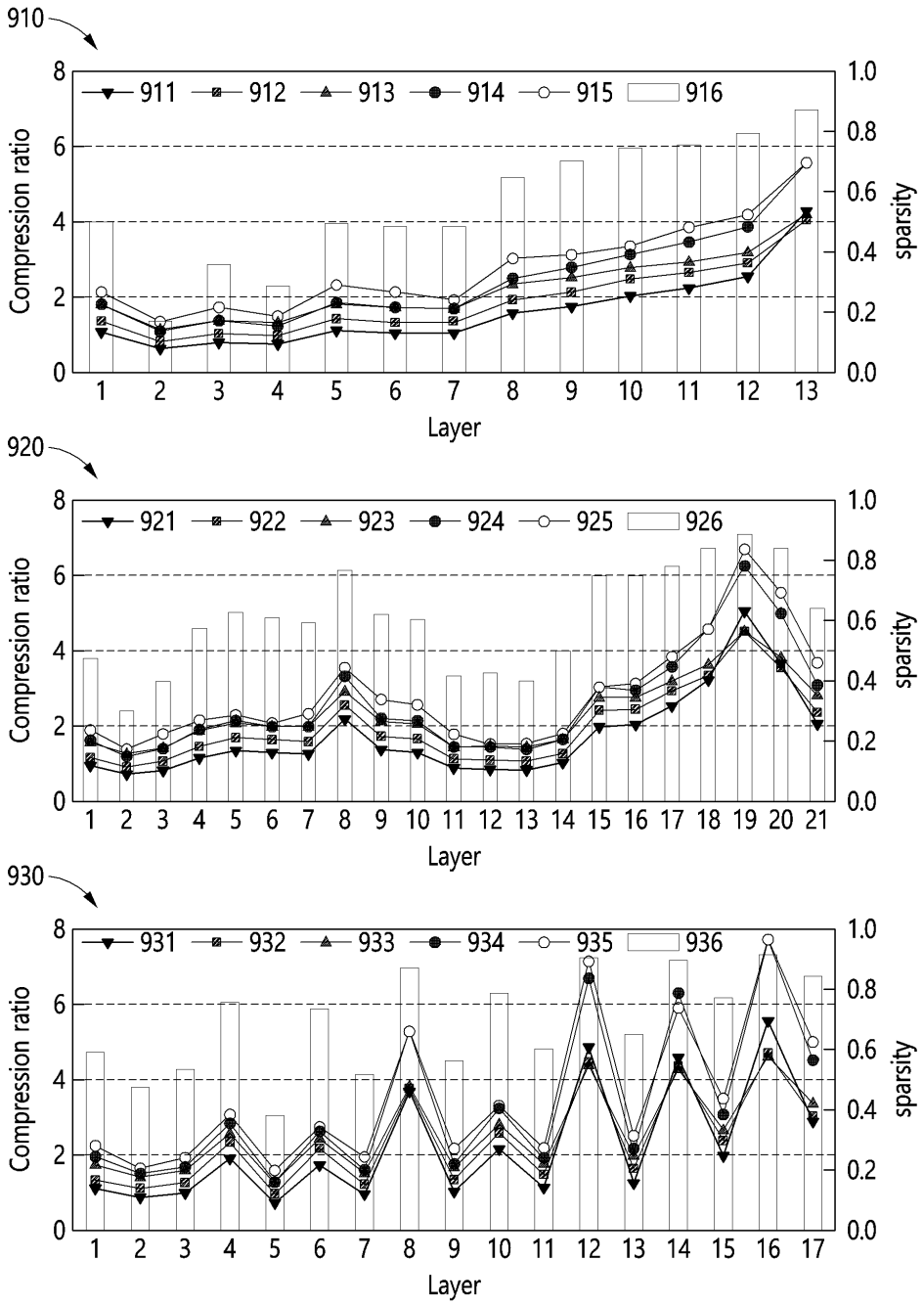
도면7


700
713
Controller
Run Registers
==
Run Cnt.
Din
m_{k-1} ... m_1 m_0
Masks
d_{p-1} ... d_1 d_0
Appended data
730
720
710
Din
d_{p-1} ... d_0 m_{k-1} ... m_0 Run
0

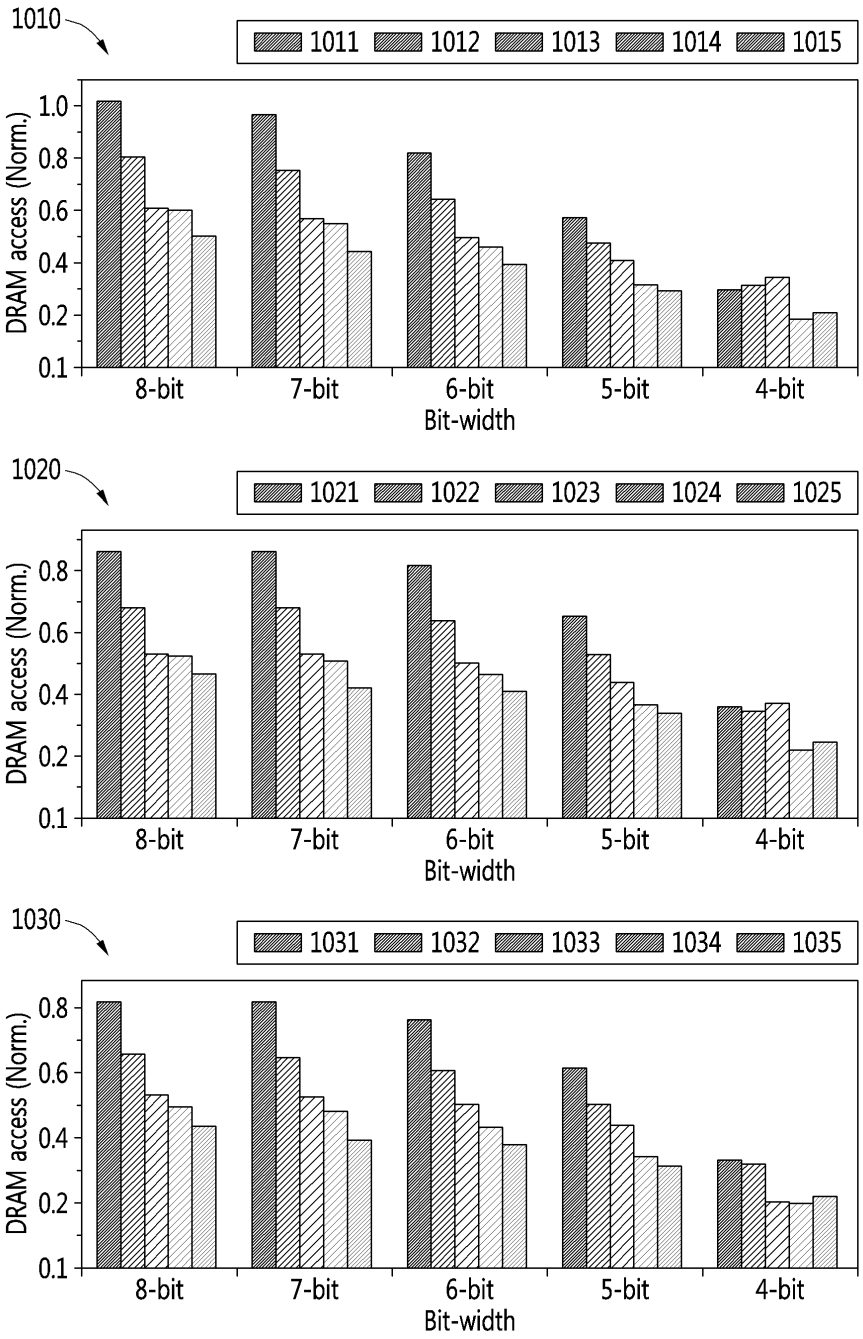
도면8


810
Max (m=2, n=2)
2.4-2.5
2.5-2.6
2.6-2.7
2.7-2.8
I/#Access
2.8
2.7
2.6
2.5
2.4
1
3
m
5
7
1
3
5
7
n
820
Max (m=2, n=2)
2.2-2.3
2.3-2.4
2.4-2.5
2.5-2.6
2.6-2.7
I/#Access
2.7
2.6
2.5
2.4
2.3
2.2
830
Max (m=2, n=2)
2.4-2.5
2.5-2.6
2.6-2.7
2.7-2.8
2.8-2.9
I/#Access
2.9
2.8
2.7
2.6
2.5
2.4

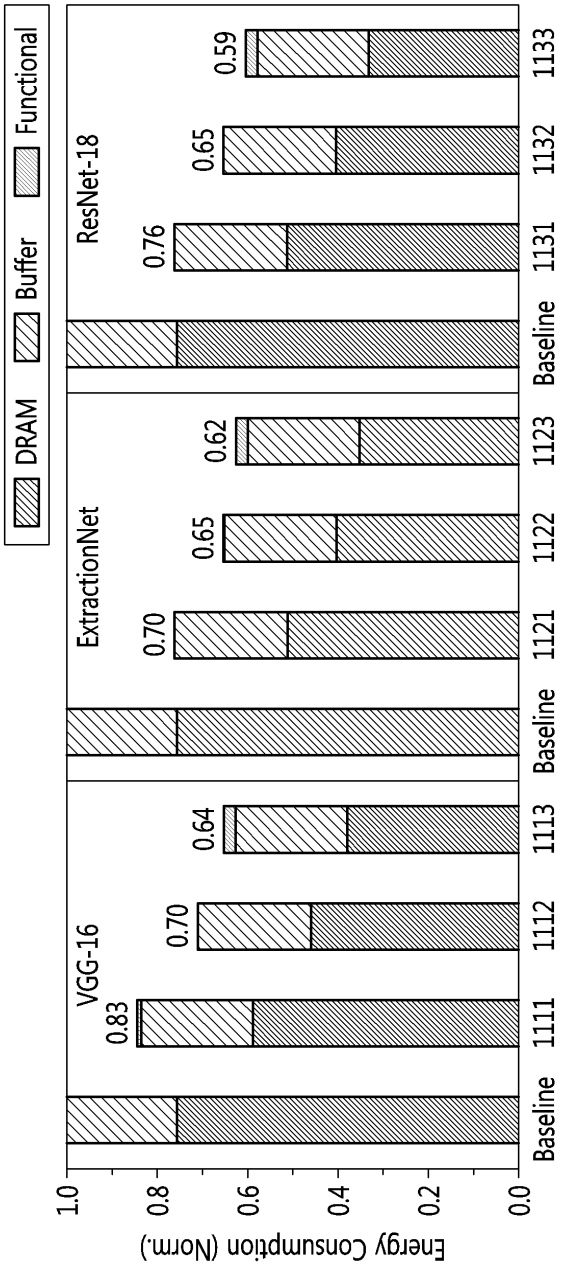
도면9


910
911
912
913
914
915
916
Compression ratio
sparsity
Layer
920
921
922
923
924
925
926
Compression ratio
sparsity
Layer
930
931
932
933
934
935
936
Compression ratio
sparsity
Layer

도면10


1010
1011
1012
1013
1014
1015
DRAM access (Norm.)
1.0
0.8
0.6
0.4
0.2
0.1
8-bit
7-bit
6-bit
5-bit
4-bit
Bit-width
1020
1021
1022
1023
1024
1025
DRAM access (Norm.)
0.8
0.6
0.4
0.2
0.1
8-bit
7-bit
6-bit
5-bit
4-bit
Bit-width
1030
1031
1032
1033
1034
1035
DRAM access (Norm.)
0.8
0.6
0.4
0.2
0.1
8-bit
7-bit
6-bit
5-bit
4-bit
Bit-width

도면11


DRAM
Buffer
Functional
Energy Consumption (Norm.)
1.0
0.8
0.6
0.4
0.2
0.0
VGG-16
ExtractionNet
ResNet-18
0.83
0.70
0.64
0.70
0.65
0.62
0.76
0.65
0.59
Baseline
1111
1112
1113
Baseline
1121
1122
1123
Baseline
1131
1132
1133