(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0100872
(43) 공개일자 2021년08월18일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
(21) 출원번호 10-2020-0014776
(22) 출원일자 2020년02월07일
심사청구일자 2020년02월07일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
이한호
인천광역시 연수구 컨벤시아대로252번길 30, 1505동 1701호(송도동, 송도 더샵 퍼스트파크 F15BL)
최동우
인천광역시 남동구 남동대로765번길 25, 508호(구월동, 지엔이츠)
(74) 대리인
양성보

전체 청구항 수 : 총 8 항

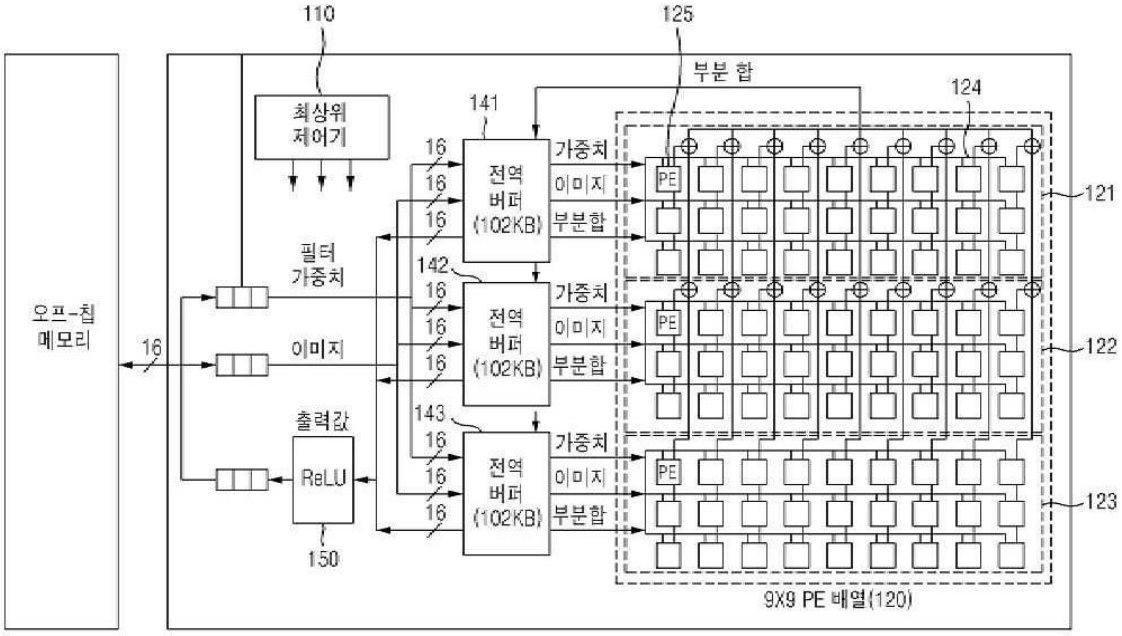
(54) 발명의 명칭 **필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속기 장치 및 방법**

(57) 요 약

필터 분해 기법을 이용한 높은 에너지 효율과 재구성형 구조를 가지는 CNN 가속기 장치 및 방법이 제시된다. 본 발명에서 제안하는 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속기 장치는 PE 사이를 연결하는 NoC, 전역 버퍼를 제어하는 최상위 제어 장치, 이미지와 필터 가중치에 대한 컨볼루션 연산을 수행하는 PE 배열,

(뒷면에 계속)

대 표 도



오프-칩 메모리로부터 읽어온 데이터를 저장하고, PE 배열의 입출력 데이터를 저장하는 전역 버퍼 레지스터, 컨볼루션 연산 결과에 대해 ReLU 활성함수를 적용하는 ReLU 장치를 포함한다. 또한, 다양한 크기의 필터를 고정된 3x3 크기의 필터로 분해하여 CNN 모델에 따라 재구성 가능하게 사용하는 방법과 이를 기반으로 기존 RS 데이터 흐름에 비해 데이터 처리율을 증가시킬 수 있는 SR 구조를 설계하는 방법을 포함한다. PE는 부분 합을 연산하는 3-병렬 SR 구조의 곱셈기와 가산기, 피연산자를 저장하는 내부 스크래치패드, 입출력 데이터를 일시적으로 저장하는 FIFO 버퍼 및 PE의 동작을 제어하는 하위 제어 장치를 포함한다.

이 발명을 지원한 국가연구개발사업

과제고유번호	1415159933
부처명	산업통상자원부
과제관리(전문)기관명	한국전자정보통신산업진흥회(KEA)
연구사업명	산업전문인력역량강화사업
연구과제명	미래형자동차R&D전문인력양성사업
기 여 율	1/2
과제수행기관명	인하대학교산학협력단
연구기간	2019.03.01 ~ 2020.02.29

이 발명을 지원한 국가연구개발사업

과제고유번호	2019R1F1A1061926
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	이공분야기초연구사업
연구과제명	고성능 저복잡도 동형암호 및 오류정정 아키텍처 연구
기 여 율	1/2
과제수행기관명	인하대학교 산학협력단
연구기간	2019.06.01 ~ 2020.02.29

명 세 서

청구범위

청구항 1

PE 사이를 연결하는 NoC, 전역 버퍼를 제어하는 최상위 제어기;

이미지와 필터 가중치에 대한 컨볼루션 연산을 수행하는 PE 배열;

입력된 데이터에 대해 부분 합 연산을 수행하는 부분 합 연산기;

오프-칩 메모리로부터 읽어온 데이터를 저장하고, PE 배열의 입출력 데이터를 저장하는 전역 버퍼; 및

컨볼루션 연산 결과에 대해 ReLU 활성함수를 적용하는 ReLU

를 포함하는 CNN 가속 장치.

청구항 2

제1항에 있어서,

최상위 제어기는,

구성 비트(Configuration bit)에 따라서 다양한 CNN 모델에 대해 CNN 가속기의 동작을 제어하는

CNN 가속 장치.

청구항 3

제1항에 있어서,

PE 배열은,

입력 이미지와 필터에 대해 컨볼루션 연산을 수행하고,

3 채널로 구성된 복수의 PE와 PE 사이의 네트워크인 NoC를 포함하고, NoC는 아비터와 데이터 경로로 구성되어 PE 사이에서 데이터를 주고받으며 제어하는

CNN 가속 장치.

청구항 4

제1항에 있어서,

전역 버퍼는,

오프-칩 메모리로부터 데이터를 읽어와 PE 배열에 입력할 데이터를 저장하는

CNN 가속 장치.

청구항 5

제1항에 있어서,

ReLU는,

다음 단계의 컨볼루션 연산을 줄이기 위해 전역 버퍼에 저장되어 있는 이전 단계의 컨볼루션 연산 결과에 활성함수를 적용하는

CNN 가속 장치.

청구항 6

제3항에 있어서,

PE는,

PE 내부의 메모리와 MUX를 제어하는 하위 제어기;

부분 합을 연산하는 복수의 3-병렬 SR 구조의 곱셈기와 가산기; 및

데이터를 저장하기 위한 내부 스크래치패드와 버퍼

를 포함하고,

필터와 입력 이미지의 부분 합을 연산하고, 3-병렬 SR 구조는 시간에 따라 순차적인 TR 구조와 대비되는 공간적으로 병렬 수행하는

CNN 가속 장치.

청구항 7

제6항에 있어서,

하위 제어기는,

최상위 제어기의 제어 신호와 PE 내부 메모리들의 상태 신호를 받아 PE의 동작을 제어하고,

PE 내부 메모리의 주소 값을 생성하는 주소 발생기; 및

PE 내부 메모리 및 MUX 제어 신호를 생성하는 제어신호 발생기

를 포함하는 CNN 가속 장치.

청구항 8

필터 분해 기법을 기반으로 입력 필터를 3x3 크기의 필터로 분해하는 단계;

구성 비트에 따라 해당하는 컨볼루션을 진행하기 위해 최상위 제어신호를 조절하는 단계;

최상위 제어신호에 따라 PE 배열을 이용하여 입력된 3x3 필터 가중치와 이미지에 대해 컨볼루션을 수행하는 단계; 및

수행된 컨볼루션 결과 데이터에 ReLU를 이용하여 활성 함수를 적용시키는 단계

를 포함하는 컨볼루션 연산 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 필터 분해 기법을 이용한 높은 에너지 효율과 재구성형 구조를 가지는 CNN(Convolutional Neural Network) 가속기 장치 및 방법에 관한 것으로, 보다 상세하게는, PE(Processing Element)와 PE 배열의 구조를 필터 분해 기법 기반의 효율적인 접근법을 적용하여 기존 RS(Row Stationary) 기반의 에너지 효율적 CNN 가속 장치보다 면적 증가율은 최소화하고 데이터 처리율은 높이는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] CNN은 딥러닝 기술 중에서 이미지를 대상으로 하는 인공 신경망이다. CNN 은 생체 인식, 자율 주행, 방범 시스템, 영상 화질 개선 등의 다양한 분야에서 응용 솔루션으로 사용되고 있다. CNN은 크게 컨볼루션(Convolution) 계층과 풀링(Pooling) 계층 그리고 완전 연결(Fully-connected) 계층으로 나뉜다. 연산 과정 중 컨볼루션 계층에서 연산의 80~90%가 이루어진다. 현재는 방대한 량의 연산을 수행하는 CNN을 가속하기 위해서 GPU(Graphics Processing Unit)이 사용되고 있지만, 에지 디바이스 (Edge-device)에 적용하기에는 오버헤드가 크기 때문에 CNN 연산에 특화된 가속기 하드웨어가 요구된다.

[0003] 에지 디바이스의 조건인 높은 에너지 효율을 충족하기 위해 매사추세츠 공과대학(MIT)에서는 에너지 효율적인 데이터 흐름인 RS(Row Stationary) 데이터 흐름을 제안하였다. RS 데이터 흐름은 오프-칩(Off-chip) 메모리, 전역 버퍼(Global buffer), PE 내부 스크래치패드(Scratchpad)로 구성된 계층적 메모리 구조를 사용한다. RS 데이

터 흐름은 PE 내부 스크래치패드에서 데이터 재사용을 최대화하여 오프-칩 메모리의 접근을 줄이고 높은 에너지 효율을 갖는다. MIT에서 제안한 RS 데이터 흐름 기반 CNN 가속기는 입력 데이터의 크기에 상관없이 연산이 가능한 재구성형 CNN 가속기이며, 최상위 제어 장치, 전역 버퍼(Global buffer), 12x14 PE 배열로 구성되어 있다. PE와 전역 버퍼는 데이터의 흐름에 최적화된 NoC(Network on Chip)로 연결되어 있다. PE는 필터(filter)와 입력 이미지의 부분 합(Partial sum)을 연산하는 장치이며, 곱셈기와 가산기, 내부 스크래치패드와 버퍼로 구성된다. MIT에서 제안한 RS 데이터 흐름 기반 CNN 가속기의 PE는 시간에 흐름에 따라 순차적으로 연산하는 TR(Temporal Reduction) 구조이므로 입력 필터의 크기가 클수록 소요되는 지연시간이 길어지고, 낮은 데이터 처리율을 갖는다는 단점이 있다.

발명의 내용

해결하려는 과제

[0004] 본 발명이 이루고자 하는 기술적 과제는 RS 데이터 흐름 기반 CNN 가속 기의 낮은 데이터 처리율을 극복하는 방법과 높은 에너지 효율을 가지는 CNN 가속기를 제공하는데 있다. 또한, 필터 분해 기법(Filter decomposition technique)을 사용함으로써, 낮은 복잡도를 갖고 재구성 가능(Reconfigurable)하게 동작하는 CNN 가속기 및 그 방법을 제공하고자 한다.

과제의 해결 수단

[0005] 일 측면에 있어서, 본 발명에서 제안하는 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속기 장치는 PE 배열과 NoC, 전역 버퍼를 제어하는 최상위 제어기, 이미지와 필터 가중치에 대한 컨볼루션 연산을 수행하는 PE 배열, 입력된 데이터에 대해 부분 합 연산을 수행하는 부분 합 연산기, 오프-칩 메모리로부터 읽어온 데이터를 저장하고, PE 배열의 입출력 데이터를 저장하는 전역 버퍼 및 컨볼루션 연산 결과에 대해 ReLU 활성함수를 적용하는 ReLU를 포함한다.

[0006] 최상위 제어기는 구성 비트(Configuration bit)에 따라서 다양한 CNN 모델에 대해 CNN 가속기의 동작을 제어한다.

[0007] PE 배열은 입력 이미지와 필터에 대해 컨볼루션 연산을 수행하고, 3 채널로 구성된 복수의 PE와 PE 사이의 네트워크인 NoC를 포함하고, NoC는 아비터와 데이터 경로로 구성되어 PE 사이에서 데이터를 주고받으며 제어한다.

[0008] 전역 버퍼는 오프-칩 메모리로부터 데이터를 읽어와 PE 배열에 입력할 데이터를 저장한다.

[0009] ReLU는 다음 단계의 컨볼루션 연산을 줄이기 위해 전역 버퍼에 저장되어 있는 이전 단계의 컨볼루션 연산 결과에 활성함수를 적용한다.

[0010] PE는 PE 내부의 메모리와 MUX를 제어하는 하위 제어기, 부분 합을 연산하는 복수의 3-병렬 SR 구조의 곱셈기와 가산기 및 데이터를 저장하기 위한 내부 스크래치패드와 버퍼를 포함하고, 필터와 입력 이미지의 부분 합을 연산하고, 3-병렬 SR 구조는 시간에 따라 순차적인 TR 구조와 대비되는 공간적으로 병렬 수행한다.

[0011] 하위 제어기는 최상위 제어기의 제어 신호와 PE 내부 메모리들의 상태 신호를 받아 PE의 동작을 제어하고, PE 내부 메모리의 주소 값을 생성하는 주소 발생기 및 PE 내부 메모리 및 MUX 제어 신호를 생성하는 제어신호 발생기를 포함한다.

[0012] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속기 장치의 컨볼루션 연산 방법은 필터 분해 기법을 기반으로 입력 필터를 3x3 크기의 필터로 분해하는 단계, 구성 비트에 따라 해당하는 컨볼루션을 진행하기 위해 최상위 제어신호를 조절하는 단계, 최상위 제어신호에 따라 PE 배열을 이용하여 입력된 3x3 필터 가중치와 이미지에 대해 컨볼루션을 수행하는 단계 및 수행된 컨볼루션 결과 데이터에 ReLU를 이용하여 활성 함수를 적용시키는 단계를 포함한다.

발명의 효과

[0013] 본 발명의 실시예들에 따르면 CNN 가속기 설계에 있어서 오프-칩 메모리의 접근을 줄이고 데이터의 재사용을 최대화하는 RS 데이터 흐름과 계층적 메모리 구조를 사용함으로써, 높은 에너지 효율을 가질 수 있다. 또한, 필터 분해 기법을 이용하여 3x3 컨볼루션만으로 다양한 크기의 필터에 대해 재구성형으로 동작하고 3x3 컨볼루션에 최적화된 SR 구조의 PE를 사용하여 컨볼루션 연산 과정을 병렬 처리함으로써, 기존 RS 데이터 흐름 기반 CNN 가

속기 장치에 비해 높은 데이터 처리량을 제공할 수 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시예에 따른 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속 장치의 구성을 나타내는 도면이다.

도 2는 본 발명의 일 실시예에 따른 필터 분해 기법을 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시예에 따른 필터 분해 기법과 SR 구조를 이용한 컨볼루션 연산 과정을 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 3-병렬 SR구조 기반 PE 아키텍처를 나타낸다.

도 5는 본 발명의 일 실시예에 따른 3-병렬 SR구조 기반 PE 내부 하위 제어 장치의 블록도를 나타낸다.

도 6은 본 발명의 일 실시예에 따른 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속 장치의 컨볼루션 연산 방법을 설명하기 위한 흐름도이다.

도 7은 본 발명의 일 실시예에 따른 3-병렬 SR구조 기반 PE의 동작 상태도를 나타낸다.

도 8은 본 발명의 일 실시예에 따른 1 채널 PE의 컨볼루션 연산 과정을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0015] 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0017] 도 1은 본 발명의 일 실시예에 따른 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속 장치의 구성을 나타내는 도면이다.

[0018] 제안하는 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속 장치는 최상위 제어기(110), PE 배열(120), 부분 합 연산기(다시 말해, PE(125)), 전역 버퍼(141, 142, 143), ReLU(150)를 포함한다.

[0019] 최상위 제어기(110)는 PE 배열(120)과 NoC(124), 전역 버퍼(143)를 제어한다. 구성 비트(Configuration bit)에 따라서 다양한 CNN 모델에 대해 CNN 가속기의 동작을 제어한다.

[0020] PE 배열(120)은 이미지와 필터 가중치에 대한 컨볼루션 연산을 수행한다. PE 배열(120)은 입력 이미지와 필터에 대해 컨볼루션 연산을 수행한다. 3 채널로 구성된 복수의 PE와 PE 사이의 네트워크인 NoC를 포함하고, NoC는 아비터와 데이터 경로로 구성되어 PE 사이에서 데이터를 주고받으며 제어한다.

[0021] 본 발명의 실시예에 따른 PE 배열은 9x9개의 PE와 NoC로 구성되어 있으며, 3개의 채널로 구분되어 3x3 크기의 2차원 컨볼루션 연산을 수행한다.

[0022] PE(125)는 PE 내부의 메모리와 MUX를 제어하는 하위 제어기, 부분 합을 연산하는 복수의 3-병렬 SR 구조의 곱셈기와 가산기 및 데이터를 저장하기 위한 내부 스크래치패드와 버퍼를 포함한다. 여기서, 3-병렬 SR 구조는 시간에 따라 순차적인 TR 구조와 대비되는 공간적으로 병렬 수행한다.

[0023] 본 발명의 실시예에 따른 PE는 이미지와 필터 가중치를 입력 받아 최상위 제어 장치의 제어 신호에 따라 부분 합(곱의 합)을 연산하는 장치이다.

[0024] 하위 제어기는 최상위 제어기의 제어 신호와 PE 내부 메모리들의 상태 신호를 받아 PE의 동작을 제어한다. 하위 제어기는 PE 내부 메모리의 주소 값을 생성하는 주소 발생기 및 PE 내부 메모리 및 MUX 제어 신호를 생성하는 제어신호 발생기를 포함한다.

[0025] 전역 버퍼(143)는 오프-칩 메모리로부터 읽어온 데이터를 저장하고, PE 배열의 입출력 데이터를 저장한다. 오프-칩 메모리로부터 데이터를 읽어와 PE 배열에 입력할 데이터를 저장한다.

[0026] ReLU(150)는 컨볼루션 연산 결과에 대해 ReLU 활성함수를 적용한다. ReLU(150)는 다음 단계의 컨볼루션 연산을 줄이기 위해 전역 버퍼에 저장되어 있는 이전 단계의 컨볼루션 연산 결과에 활성함수를 적용한다.

[0027] 더욱 상세하게는, 도 1를 참조하면, 이미지 데이터와 필터 가중치가 오프-칩 메모리로부터 3개의 전역 버퍼로

전송되고, 각 PE가 NoC로 연결된 3-채널 PE 배열은 전역 버퍼로부터 이미지와 필터 가중치를 받아 각각의 PE에 할당하여 2D 컨볼루션을 수행한다. 이 과정은 에너지 효율을 위해 데이터 재사용을 최대한으로 하는 RS 데이터 흐름으로 설계되었다. PE 내부의 스크래치패드에서 데이터의 재사용을 최대화함으로써 오프-칩 메모리의 접근을 줄이고 높은 에너지 효율을 갖는다. 제안된 CNN 가속기는 최상위 제어기의 제어 신호에 따라 다양한 CNN 알고리즘에 대해 재구성 가능하게 동작한다. 2D 컨볼루션을 수식으로 표현하면 수학식 1과 같다.

[0028] $$y[m,n] = ifm[m,n] * w[m,n] = \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} ifm[m-i, n-j] \cdot w[i,j]$$

[0029] 수학식 1

[0030] ifm은 이미지 픽셀 값이고 w는 필터 가중치 값이다. m,n은 2D 컨볼루션 결과인 출력특징 지도의 좌표 값이며, i,j는 필터의 좌표 값이다.

[0031] 제안된 CNN 가속기는 필터 분해 기법에 따라 3x3 컨볼루션만으로 다양한 크기의 필터 크기를 지원할 수 있도록 최상위 블록 및 PE의 구조가 설계되었다. 3x3 컨볼루션에 최적화된 구조로 설계되었으므로 면적당 데이터 처리율이 기존 RS데이터 흐름기반의 CNN 가속기 보다 향상되었다.

[0033] 도 2는 본 발명의 일 실시예에 따른 필터 분해 기법을 설명하기 위한 도면이다.

[0034] CNN의 컨볼루션 연산의 크기는 다양하게 존재한다. AlexNet만 하더라도 3x3, 5x5, 11x11 세 가지 크기의 컨볼루션 연산을 수행한다. 이를 모두 수행하는 CNN 가속기를 설계하기 위해서는 상당히 복잡한 컨트롤이 필요하다. 필터의 크기에 따라 사용하는 연산기를 지정하여 주고, 나머지 연산기는 절전(sleep) 모드로 동작 시키거나, 채널의 수를 나누어서 필터의 크기가 작을 경우에는 병렬적으로 컨볼루션을 수행하도록 컨트롤할 수 있다. 하지만 두 경우 모두 필터 크기에 따라 사용하지 않는 연산기가 발생하게 된다. 이러한 문제를 해결할 수 있는 방법이 바로 필터 분해(filter decomposition) 기법이다. 예를 들면 도 2과 같이 5x5의 커널을 3의 배수의 크기만큼 제로(zero) 패딩(padding)을 해준다. 즉, 5x5 커널을 6x6으로 만든 뒤에 4개의 3x3 필터로 나누어 줄 수 있다. 결국 필터의 가중치 값을 입력할 때 분해된 가중치 값들을 이용하면, 어떠한 크기의 컨볼루션이든 3x3 컨볼루션을 반복 수행하는 것만으로 모두 연산이 가능하다. 또한, 고정된 3x3 크기의 필터만을 사용하게 되므로 낮은 유연성이 문제였던 SR구조의 사용이 가능해진다.

[0036] 도 3은 본 발명의 일 실시예에 따른 필터 분해 기법과 SR 구조를 이용한 컨볼루션 연산 과정을 설명하기 위한 도면이다.

[0037] 도 3은 필터 분해 기법과 SR 구조를 사용한 1D 컨볼루션 과정과 기존 TR 구조의 1D 컨볼루션 과정의 비교를 나타낸다. 6개의 입력 픽셀과 5개의 필터 가중치에 일반적인 1D 컨볼루션을 수행하면 10 cycle의 수행시간이 요구된다. 하지만 제안된 필터 분해 기법과 SR 구조를 적용한 컨볼루션 연산은 데이터를 읽을 주소 값 결정에 대해서 복잡한 제어를 요구하지만, 4 cycle만에 연산을 수행할 수 있다. 이처럼 필터 분해 기법은 데이터 처리율을 높이기 위한 SR 구조의 중요한 기반 기술이 된다.

[0039] 도 4는 본 발명의 일 실시예에 따른 3-병렬 SR구조 기반 PE 아키텍처를 나타낸다.

[0040] 본 발명의 실시예에 따른 PE 배열은 9x9개의 PE와 NoC로 구성되어 있으며, 3개의 채널로 구분되어 3x3 크기의 2차원 컨볼루션 연산을 수행한다.

[0041] PE는 PE 내부의 메모리와 MUX를 제어하는 하위 제어기(410), 부분 합을 연산하는 복수의 3-병렬 SR 구조의 곱셈기(431, 432, 433)와 가산기(421, 422, 423) 및 데이터를 저장하기 위한 내부 스크래치패드(441, 442, 443)와 버퍼를 포함한다. 여기서, 3-병렬 SR 구조는 시간에 따라 순차적인 TR 구조와 대비되는 공간적으로 병렬 수행한다.

[0042] 본 발명의 실시예에 따른 PE는 이미지와 필터 가중치를 입력 받아 최상위 제어 장치의 제어 신호에 따라 부분합(곱의 합)을 연산하는 장치이다.

[0043] 하위 제어기(410)는 최상위 제어기의 제어 신호와 PE 내부 메모리들의 상태 신호를 받아 PE의 동작을 제어한다.

하위 제어기(410)는 PE 내부 메모리의 주소 값을 생성하는 주소 발생기(411) 및 PE 내부 메모리 및 MUX 제어 신호를 생성하는 제어신호 발생기(412)를 포함한다.

[0044] 본 발명의 실시예에 따른 PE는 4개의 입출력 First-In-First-Out(FIFO) 버퍼 레지스터, 파라미터를 저장하기 위한 3개의 스크래치패드, 3개의 곱셈기 및 가산기 그리고 하위 제어 장치(sub-controller)로 구성되어 있으며, 3x3 컨볼루션 연산에 최적화된 3-병렬 SR 구조로 설계되었다. 앞서 설명한 바와 같이 3x3 컨볼루션만으로 재구성형으로 동작할 수 있기 때문에 입력에 대해 유연성이 낮은 SR구조의 사용이 가능하다. 기존 RS 데이터 흐름 기반의 CNN 가속기는 시간에 흐름에 따라 순차적인 TR구조를 사용하기 때문에 데이터 처리율이 다른 CNN 가속기에 비해 매우 낮다는 단점이 있었다. 하지만 제안된 PE 구조는 병렬적인 SR구조를 사용함으로써 보다 높은 데이터 처리율을 갖는다. PE는 3개의 행이 하나의 채널을 구성한다. PE의 동작에는 먼저 입력 이미지와 필터 가중치가 FIFO 버퍼 레지스터에 쓰이고, 제어 신호에 따라 스크래치패드로 이동된다. 제어 신호에 의해 스크래치패드에서 데이터가 읽어지면 곱셈기와 가산기를 이용하여 부분 합을 구하고 마지막 행의 PE라면 누적 연산 없이 연산결과를 출력 FIFO 버퍼 레지스터에 저장하고 클럭 신호에 따라 순차적으로 출력하며, 첫 번째 혹은 두 번째 행의 PE는 이전 연산 결과를 누적하기 위해 입력 부분 합과 누적연산을 수행한다.

[0046] 도 5는 본 발명의 일 실시예에 따른 3-병렬 SR구조 기반 PE 내부 하위 제어 장치의 블록도를 나타낸다.

[0047] PE를 제어하는 하위 제어기의 블록도이다. 하위 제어기는 제어신호 생성기와 주소 발생기로 구성된다. 하위 제어기는 최상위 제어 신호에 따라서 PE 내부의 메모리 및 MUX를 제어하는 제어신호를 생성하고, 각 메모리의 주소 값을 생성한다.

[0049] 도 6은 본 발명의 일 실시예에 따른 필터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속 장치의 컨볼루션 연산 방법을 설명하기 위한 흐름도이다.

[0050] 제안하는 터 분해 기법을 이용한 에너지 효율적 재구성형 CNN 가속 장치의 컨볼루션 연산 방법은 필터 분해 기법을 기반으로 입력 필터를 3x3 크기의 필터로 분해하는 단계(610), 구성 비트에 따라 해당하는 컨볼루션을 진행하기 위해 최상위 제어신호를 조절하는 단계(620), 최상위 제어신호에 따라 PE 배열을 이용하여 입력된 3x3 필터 가중치와 이미지에 대해 컨볼루션을 수행하는 단계(630) 및 수행된 컨볼루션 결과 데이터에 ReLU를 이용하여 활성 함수를 적용시키는 단계(640)를 포함한다.

[0051] 본 발명의 실시예에 따른 필터 분해 기법은 모든 필터의 크기를 3x3 크기의 필터로 나누는 방법을 의미하고, 3x3 크기의 필터만으로 다양한 크기의 컨볼루션 연산이 가능하여 PE 구조에 대해 유연성이 낮은 SR 구조 사용이 가능하도록 하는 기반 기술이다.

[0053] 도 7은 본 발명의 일 실시예에 따른 3-병렬 SR구조 기반 PE의 동작 상태도를 나타낸다.

[0054] 도 8은 본 발명의 일 실시예에 따른 1 채널 PE의 컨볼루션 연산 과정을 설명하기 위한 도면이다.

[0055] 도 7은 3-병렬 SR구조 기반 PE의 동작 상태도를 나타낸다. en과 op는 최상위 제어 신호이며, 각각의 동작 상태에 대한 신호는 하위 제어 장치에 의해 생성된다.

[0056] 도 8은 1 채널(3개의 행) PE의 컨볼루션 연산 동작을 나타낸다. PE1이 제일 먼저 가중치를 FIFO 버퍼에 쓰고, PE3이 가중치를 쓸 때까지 PE 1,2는 대기상태이다. 가중치가 모든 PE의 FIFO 버퍼에 쓰였다면, 바로 이미지의 값을 써준다. 차례대로 이미지 데이터의 쓰기가 완료되면 각 PE는 버퍼로부터 스크래치패드로 데이터를 옮긴다. 옮겨진 데이터는 바로 컨볼루션 연산을 수행한다. PE1,2의 경우 PE3의 연산이 완료된 뒤에, 누적연산 과정을 거쳐야 하기 때문에 연산 값을 스크래치패드에 저장하여 놓는다. PE3의 연산 결과는 바로 출력 버퍼로 이동시킨다. 얻어진 PE3의 출력 값을 각각의 PE에서 누적하여, PE1에서 최종적인 컨볼루션 결과를 얻을 수 있다.

[0058] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로

세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0059] 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 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

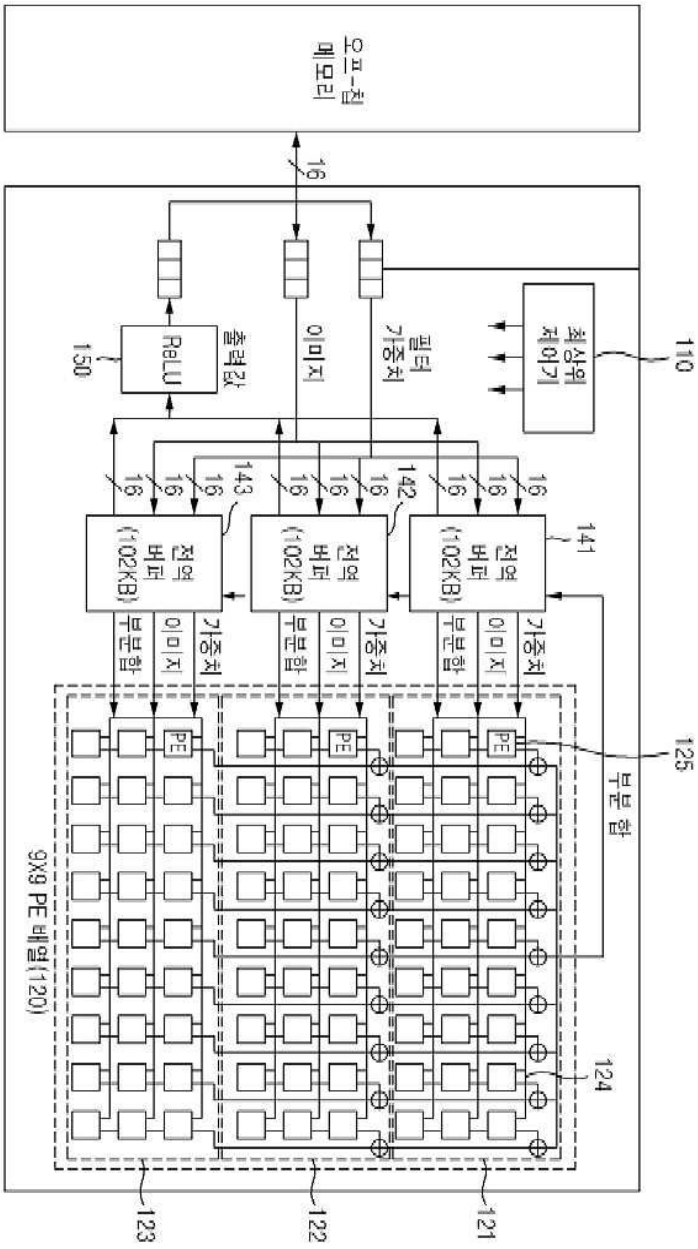
[0060] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic 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.

[0061] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

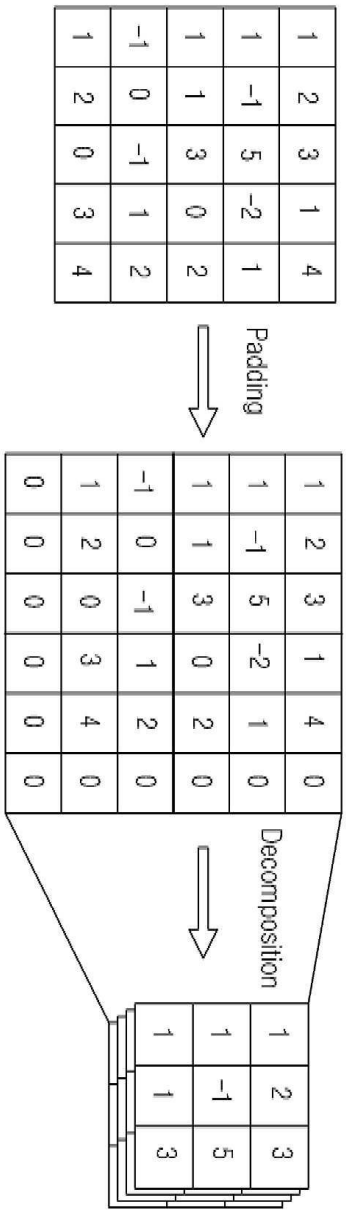
[0062] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

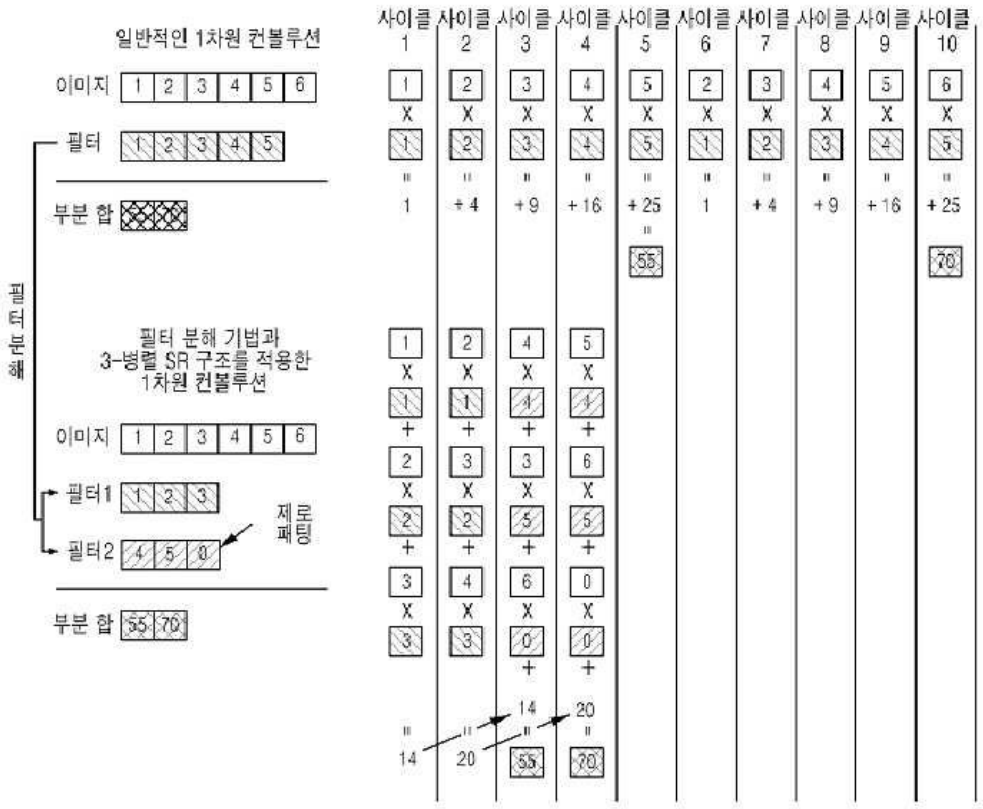
도면1


온-칩 메모리
16
ReLU
150
출력값
이미지
필터 가중치
최상위 제어기
110
143
142
141
전역 버퍼 (1102KB)
부분합
이미지
가중치
PE
125
부분합
9X9 PE 배열(120)
124
123
122
121

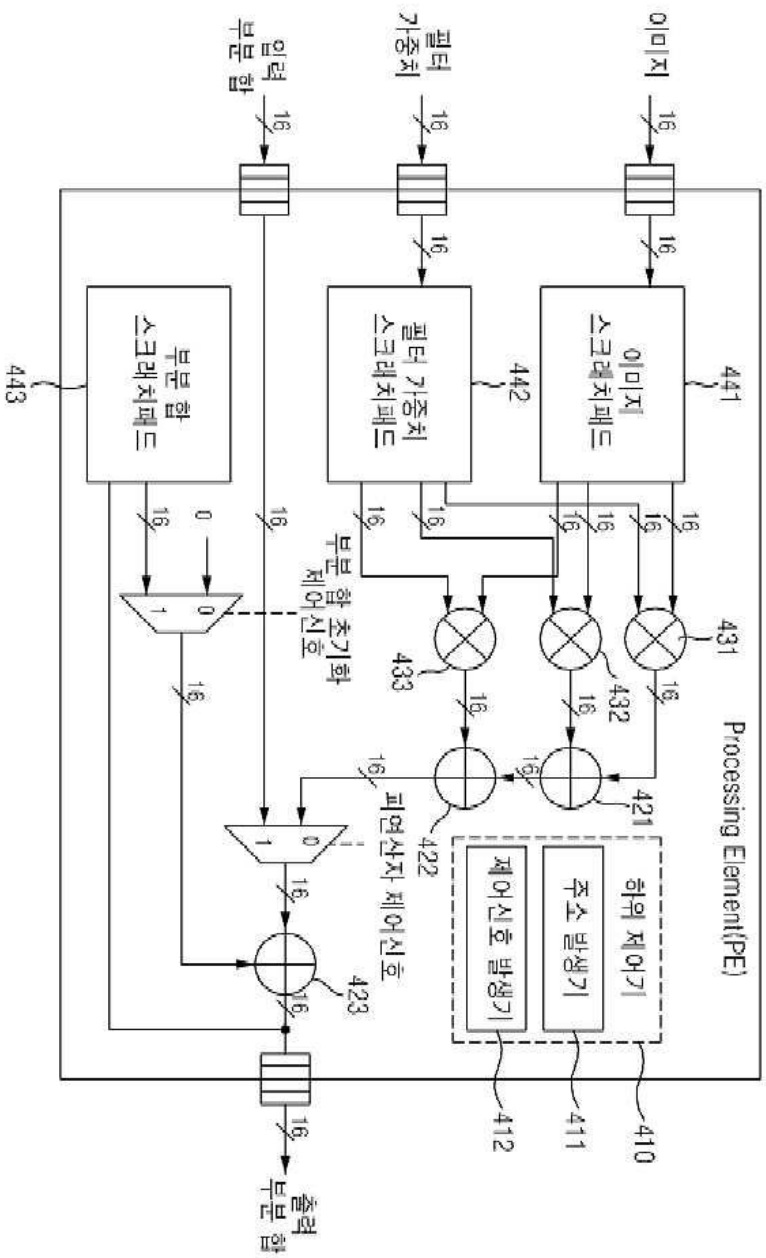
도면2



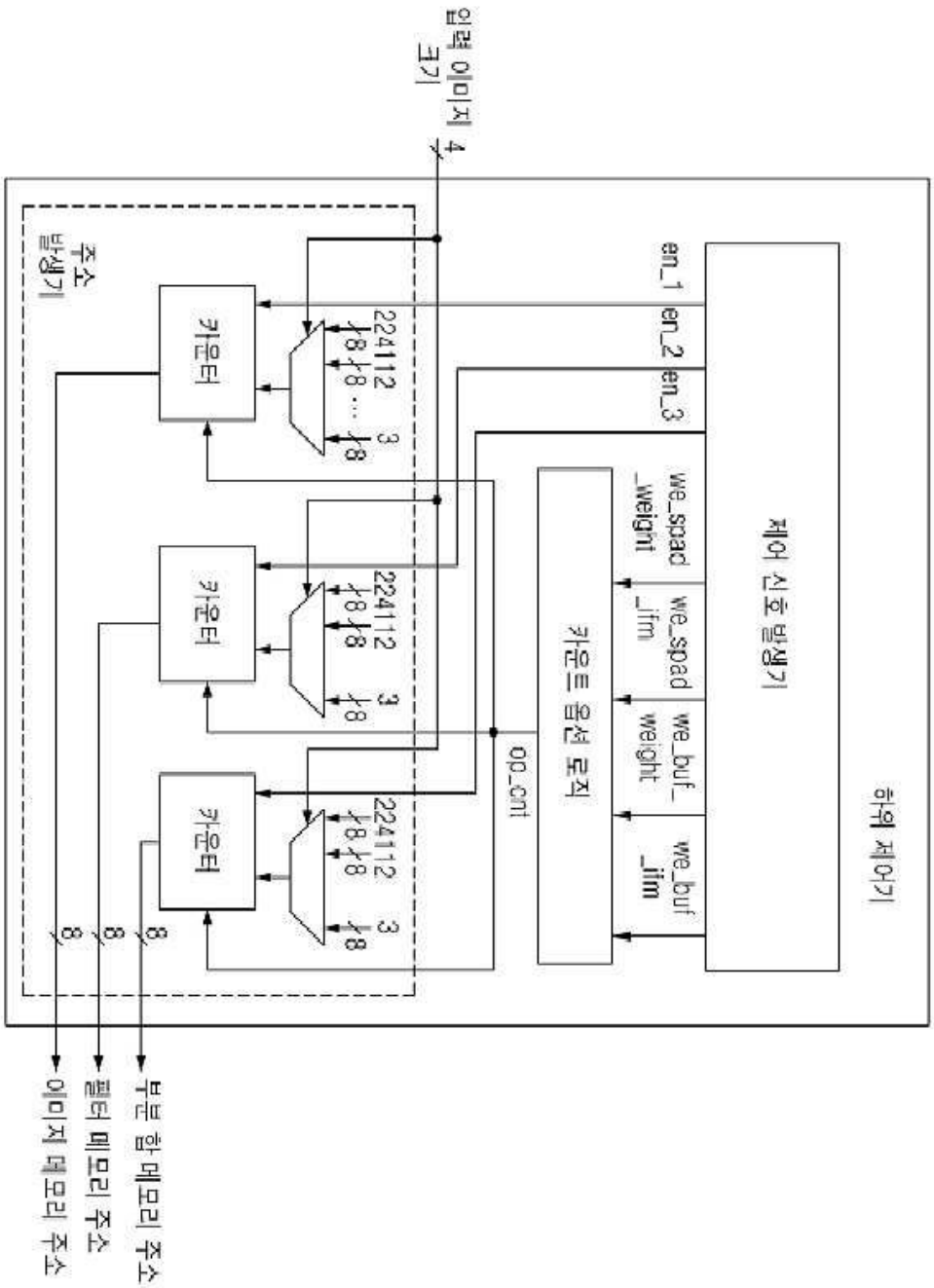
도면3


일반적인 1차원 컨볼루션
이미지
필터
부분 합
필터 분해
필터 분해 기법과
3-병렬 SR 구조를 적용한
1차원 컨볼루션
이미지
필터1
필터2
제로
패팅
부분 합
사이클 1
사이클 2
사이클 3
사이클 4
사이클 5
사이클 6
사이클 7
사이클 8
사이클 9
사이클 10

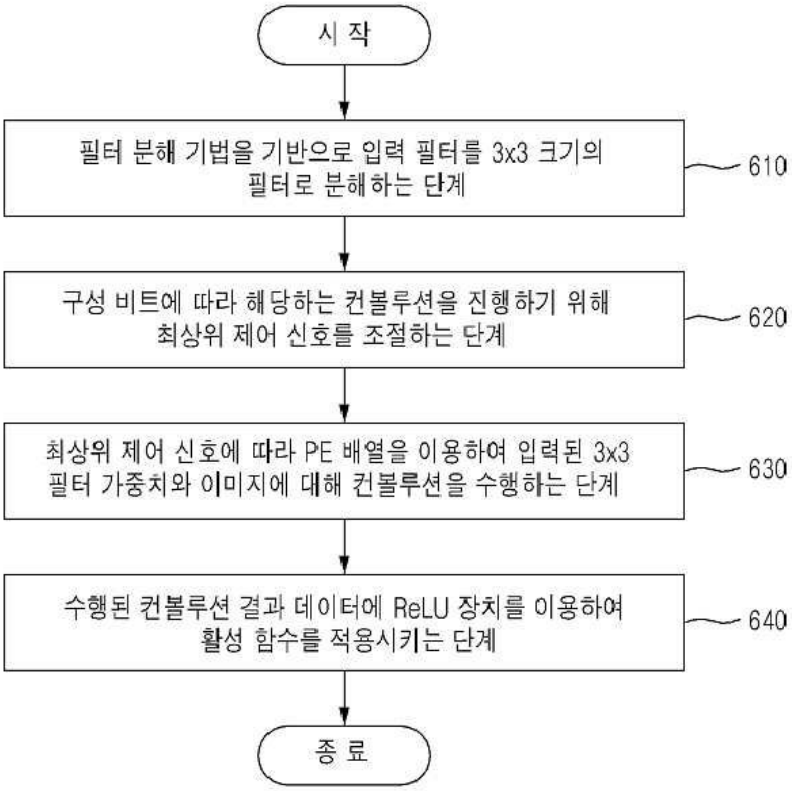
도면4


이미지
필터 가중치
입력 부분 합
이미지 스크래치패드
필터 가중치 스크래치패드
부분 합 스크래치패드
441
442
443
부분 합 초기화 제어신호
피연산자 제어신호
하위 제어기
주소 발생기
제어신호 발생기
410
411
412
421
422
423
431
432
433
Processing Element(PE)
출력 부분 합
16
0
1

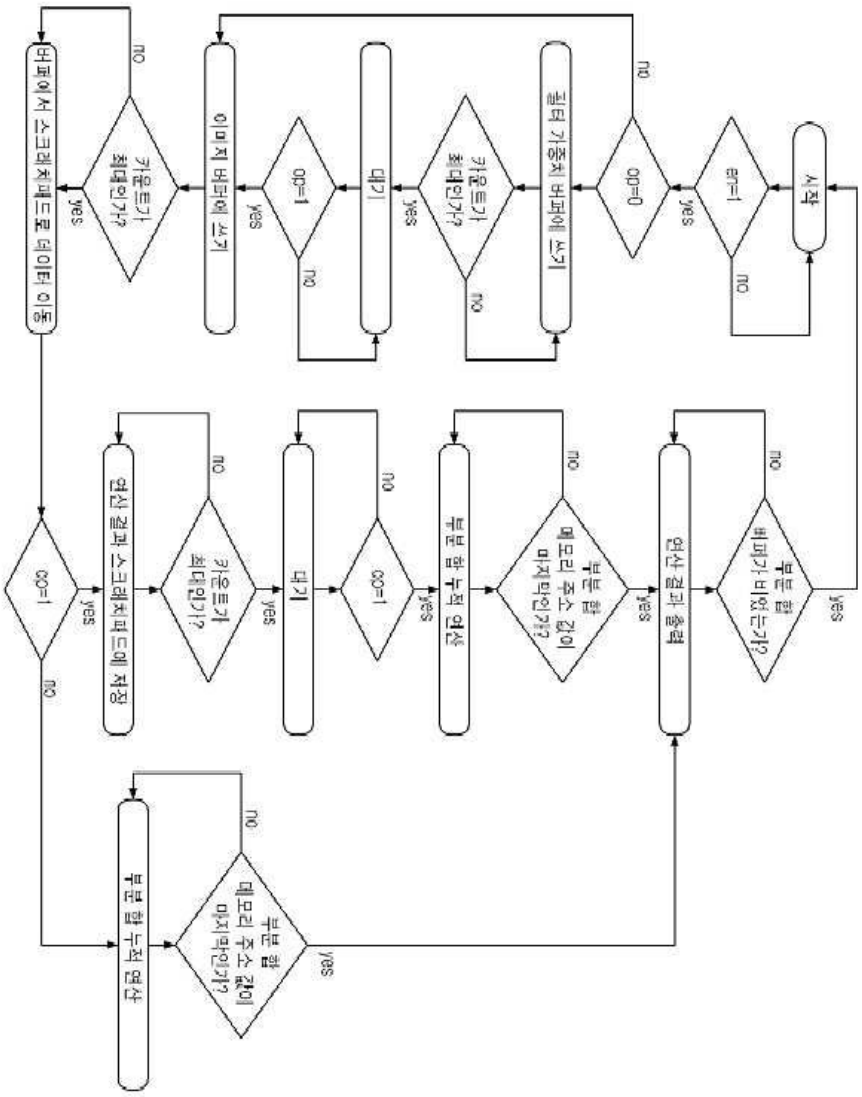
도면5


입력 이미지 크기
4
주소 발생기
카운터
카운터
카운터
224112
3
8
8
8
8
8
8
en_1
en_2
en_3
we_spad_weight
we_spad_ifm
we_buf_weight
we_buf_ifm
제어 신호 발생기
하위 제어기
카운트 옵션 로직
op_cnt
부분 합 메모리 주소
필터 메모리 주소
이미지 메모리 주소

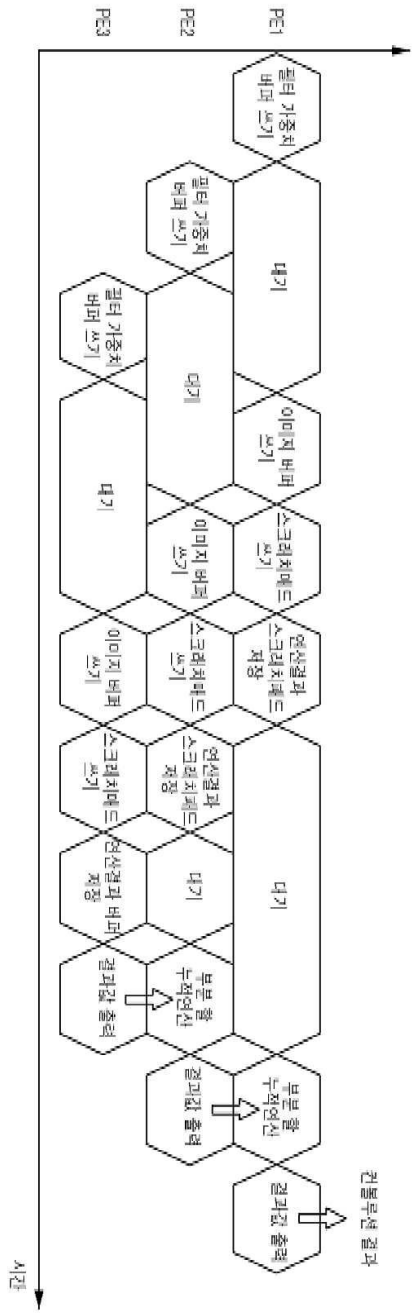
도면6


시 작
필터 분해 기법을 기반으로 입력 필터를 3x3 크기의 필터로 분해하는 단계
610
구성 비트에 따라 해당하는 컨볼루션을 진행하기 위해 최상위 제어 신호를 조절하는 단계
620
최상위 제어 신호에 따라 PE 배열을 이용하여 입력된 3x3 필터 가중치와 이미지에 대해 컨볼루션을 수행하는 단계
630
수행된 컨볼루션 결과 데이터에 ReLU 장치를 이용하여 활성 함수를 적용시키는 단계
640
종 료

도면7


시작
en=1
op=0
필터 가중치 버퍼에 쓰기
카운트가 최대인가?
대기
op=1
이미지 버퍼에 쓰기
카운트가 최대인가?
버퍼에서 스크래치패드로 데이터 이동
co=1
연산 결과 스크래치패드에 저장
카운트가 최대인가?
대기
co=1
부분 합 누적 연산
부분 합 메모리 주소 값이 마지막인가?
연산 결과 출력
부분 합 버퍼가 비었는가?
부분 합 누적 연산
부분 합 메모리 주소 값이 마지막인가?
yes
no

도면8


PE3
PE2
PE1
필터 가중치 버퍼 쓰기
대기
이미지 버퍼 쓰기
스크래치패드 쓰기
연산결과 스크래치패드 저장
대기
블록 합 누적연산
결과값 출력
필터 가중치 버퍼 쓰기
대기
이미지 버퍼 쓰기
스크래치패드 쓰기
연산결과 스크래치패드 저장
대기
블록 합 누적연산
결과값 출력
필터 가중치 버퍼 쓰기
대기
이미지 버퍼 쓰기
스크래치패드 쓰기
연산결과 저장
대기
결과값 출력
컨볼루션 결과
시간