(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0149907
(43) 공개일자 2024년10월15일

(51) 국제특허분류(Int. Cl.)
G06N 3/0464 (2023.01) *G06N 3/0495* (2023.01)
G06N 3/082 (2023.01)
(52) CPC특허분류
G06N 3/0464 (2023.01)
G06N 3/0495 (2023.01)
(21) 출원번호 10-2024-7028942
(22) 출원일자(국제) 2023년02월13일
심사청구일자 없음
(85) 번역문제출일자 2024년08월28일
(86) 국제출원번호 PCT/CN2023/075661
(87) 국제공개번호 WO 2023/155748
국제공개일자 2023년08월24일
(30) 우선권주장
17/673,490 2022년02월16일 미국(US)

(71) 출원인
모펫 인터내셔널 컴퍼니 리미티드
중국 홍콩 999077 산 포 콩 29 루크 호프 스트리트 왕 페이 인더스트리알 빌딩 11/에프 룸 8
(72) 발명자
장 샤오치엔
미국 캘리포니아주 94022 로스앨토스 스위트 200 949 셔우드 애비뉴
옌 은쉬
미국 캘리포니아주 94022 로스앨토스 스위트 200 949 셔우드 애비뉴
샤오 쥐빈
미국 캘리포니아주 94022 로스앨토스 스위트 200 949 셔우드 애비뉴
(74) 대리인
특허법인아주김장리

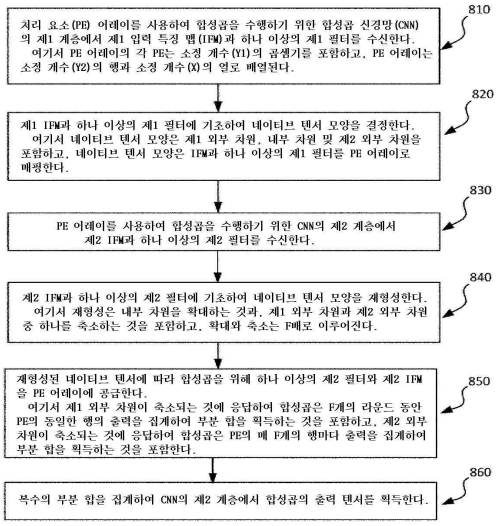
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **희소 신경망을 위한 적응형 텐서 계산 커널**

(57) 요 약

적응형 텐서 계산 커널을 사용한 신경망 계산의 효율성을 개선하기 위한 방법, 시스템 및 장치(컴퓨터 저장 매체에 인코딩된 컴퓨터 프로그램을 포함함). 첫째, 적응형 텐서 계산 커널은 병렬 처리를 위한 처리 요소(PE) 어레이에 가중치와 입력 값을 분배하기 위해 입력/가중치 텐서의 다양한 모양에 따라 형태를 조정할 수 있다. 텐서 계산 커널의 모양에 따라 합성곱 계산을 수행하기 위해 추가적인 클러스터 간 또는 클러스터 내 덧셈기가 필요할 수 있다. 둘째, 적응형 텐서 계산 커널은 모든 유형의 합성곱 계산을 포괄하기 위해 두 가지 다른 텐서 연산 모드, 즉, 1×1 텐서 연산 모드 및 3×3 텐서 연산 모드를 지원할 수 있다. 셋째, 기본 PE 어레이는 희소 신경망의 다양한 압축 비율과 희소성 세분성을 지원하기 위해 각 PE 내부 버퍼(예를 들어, 레지스터 파일)를 다르게 구성할 수 있다.

대 표 도 - 도8



(52) CPC특허분류
G06N 3/082 (2023.01)

명 세 서

청구범위

청구항 1

컴퓨터 구현 방법으로서,

처리 요소(PE) 어레이를 사용하여 합성곱을 수행하기 위한 합성곱 신경망(convolutional neural network: CNN)의 제1 계층에서 제1 입력 특징 맵(input feature map: IFM)과 하나 이상의 제1 필터를 수신하는 단계로서, 상기 PE 어레이의 각 PE는 소정 개수(Y1)의 곱셈기를 포함하고, 상기 PE 어레이는 소정 개수(Y2)의 행과 소정 개수(X)의 열로 배열된, 상기 수신하는 단계;

상기 제1 IFM과 상기 하나 이상의 제1 필터에 기초하여 네이티브 텐서 모양(native tensor shape)을 결정하는 단계로서, 상기 네이티브 텐서 모양은 제1 외부 차원, 내부 차원 및 제2 외부 차원을 포함하고, 상기 네이티브 텐서 모양은 상기 제1 IFM과 상기 하나 이상의 제1 필터를 상기 PE 어레이에 매핑하는, 상기 결정하는 단계;

상기 PE 어레이를 사용하여 합성곱을 수행하기 위한 CNN의 제2 계층에서 제2 IFM과 하나 이상의 제2 필터를 수신하는 단계;

상기 제2 IFM과 상기 하나 이상의 제2 필터에 기초하여 상기 네이티브 텐서 모양을 재형성(reshaping)하는 단계로서, 상기 재형성은 내부 차원을 확대하는 단계, 및 상기 제1 외부 차원과 상기 제2 외부 차원 중 하나를 축소하는 단계를 포함하고, 상기 확대와 축소는 F배로 이루어지는, 상기 재형성하는 단계;

재형성된 네이티브 텐서에 따라 합성곱을 위해 상기 하나 이상의 제2 필터와 상기 제2 IFM을 상기 PE 어레이에 공급하는 단계로서,

상기 제1 외부 차원이 축소되는 것에 응답하여 상기 합성곱은 F개의 라운드 동안 PE의 동일한 행의 출력을 집계하여 부분 합을 얻는 것을 포함하고,

상기 제2 외부 차원이 축소되는 것에 응답하여 상기 합성곱은 PE의 매 F개의 행마다 출력을 집계하여 부분 합을 얻는 것을 포함하는, 상기 공급하는 단계; 및

복수의 상기 부분 합을 집계하여 상기 CNN의 제2 계층에서 상기 합성곱의 출력 텐서를 얻는 단계

를 포함하되, Y1, Y2, X 및 F는 모두 1보다 큰 정수인, 방법.

청구항 2

청구항 1에 있어서, 상기 CNN의 제2 계층은 상기 CNN의 제1 계층 뒤에 있고, 상기 제2 IFM은 상기 제1 IFM보다 많은 입력 채널을 포함하고, 상기 제1 IFM보다 낮은 해상도를 포함하는, 방법.

청구항 3

청구항 1에 있어서, 상기 하나 이상의 제2 필터 각각은 2차원(2D) 커널(kernel)의 복수의 채널을 포함하고, 각 2D 커널은 1×1 또는 3×3의 차원을 갖는, 방법.

청구항 4

청구항 3에 있어서, 상기 재형성된 네이티브 텐서에 따라 상기 하나 이상의 제2 필터를 상기 PE 어레이에 공급하는 단계는,

상기 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 상기 하나 이상의 제2 필터를 행렬로 변환하는 단계로서, 상기 하나 이상의 제2 필터의 각 2D 커널이 1×1의 차원을 갖는 것에 응답하여, 상기 행렬의 각 행은 상기 하나 이상의 제2 필터의 서로 다른 입력 채널의 가중치를 포함하는, 상기 변환하는 단계; 및

상기 복수의 입력 채널이 한 번에 동시에 처리되도록 상기 행렬의 각 행의 가중치를 PE의 서로 다른 열에 분배하는 단계

를 포함하는, 방법.

청구항 5

청구항 3에 있어서, 상기 재형성된 네이티브 텐서에 따라 상기 하나 이상의 제2 필터를 상기 PE 어레이에 공급하는 단계는,

상기 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 상기 하나 이상의 제2 필터를 행렬로 변환하는 단계로서, 상기 하나 이상의 제2 필터의 각 2D 커널이 3×3의 차원을 갖고 9개의 가중치를 포함하는 것에 응답하여, 상기 9개의 가중치는 상기 행렬의 동일한 행에 배치되는, 상기 변환하는 단계; 및

동일한 채널의 가중치가 한 번에 동시에 처리되도록 상기 행렬의 동일한 행의 9개의 가중치를 PE의 서로 다른 열에 분배하는 단계

를 포함하는, 방법.

청구항 6

청구항 5에 있어서, 상기 재형성된 네이티브 텐서에 따라 상기 IFM을 상기 PE 어레이에 공급하는 단계는,

상기 재형성된 네이티브 텐서의 내부 차원과 제2 외부 차원에 따라 상기 IFM을 행렬로 변환하는 단계; 및

상기 행렬의 열에 대응하는 상기 IFM의 입력 값을 PE의 행의 버퍼에 공급하는 단계

를 포함하는, 방법.

청구항 7

청구항 1에 있어서,

상기 하나 이상의 필터의 채널을 복수의 채널 그룹으로 분할하는 단계로서, 각 채널 그룹은 1보다 큰 정수인 고정된 수의 채널을 포함하는, 상기 분할하는 단계; 및

각 채널 그룹 내의 고정된 백분율의 가중치가 0이 아니도록 상기 복수의 채널 그룹 각각을 가지치기(pruning)하는 단계

를 추가로 포함하는, 방법.

청구항 8

청구항 7에 있어서,

상기 PE 어레이의 각 PE와 연관된 버퍼의 깊이를 결정하는 단계;

상기 버퍼의 깊이가 상기 고정된 수보다 큰 것에 응답하여 상기 버퍼를 각 PE에 대한 개인 메모리로 구성하는 단계; 및

상기 버퍼의 깊이가 상기 고정된 수보다 작은 것에 응답하여 상기 PE의 버퍼와 이웃 PE의 하나 이상의 버퍼를 공유 메모리로 결합하는 단계

를 추가로 포함하는, 방법.

청구항 9

청구항 8에 있어서, 각 PE의 개인 메모리는 상기 PE 내 소정 개수(Y1)의 곱셈기에 의해 검색 가능한 입력 값을 저장하는, 방법.

청구항 10

청구항 8에 있어서, 상기 공유 메모리는 상기 PE 내 소정 개수(Y1)의 곱셈기와 상기 하나 이상의 이웃 PE에 의해 검색 가능한 입력 값을 저장하는, 방법.

청구항 11

청구항 1에 있어서, PE의 각 행은 각 PE 내 곱셈기의 수(Y1)에 대응하는 수(Y1)의 덧셈기 트리(adder tree)와 결합되고, 각 PE 내의 각 곱셈기는 집계를 위해 대응하는 덧셈기 트리에 곱셈 출력을 보내는, 방법.

청구항 12

청구항 1에 있어서, 상기 하나 이상의 제2 필터 각각은 복수의 0이 아닌 가중치를 포함하고,

상기 하나 이상의 제2 필터를 합성곱을 위해 상기 PE 어레이에 공급하는 단계는,

각 0이 아닌 가중치를 대응하는 PE의 곱셈기에 상기 0이 아닌 가중치와 대응 인덱스를 포함하는 인덱스-값 쌍으로 공급하는 단계

를 포함하고; 상기 합성곱은,

상기 인덱스에 따라 상기 대응하는 PE의 버퍼로부터 입력 값을 검색하는 단계; 및

검색된 값과 상기 0이 아닌 가중치를 상기 곱셈기에 보내어 출력을 얻는 단계; 및

상기 대응하는 PE와 같은 행에 있는 다른 PE의 다른 곱셈기에 의해 생성된 출력과 함께 집계를 위해 대응하는 덧셈기 트리에 상기 출력을 보내는 단계

를 포함하는, 방법.

청구항 13

청구항 1에 있어서, 각 PE 내 소정 개수(Y1)의 곱셈기는 병렬로 데이터를 처리하고, 상기 PE 어레이 내의 PE는 병렬로 데이터를 처리하는, 방법.

청구항 14

시스템으로서,

하나 이상의 프로세서와 하나 이상의 비일시적 컴퓨터 판독 가능 메모리를 포함하고, 상기 비일시적 컴퓨터 판독 가능 메모리는 상기 하나 이상의 프로세서에 결합되고, 상기 하나 이상의 프로세서에 의해 실행 가능한 명령어를 포함하도록 구성되고, 상기 명령어는, 상기 시스템으로 하여금,

처리 요소(PE) 어레이를 사용하여 합성곱을 수행하기 위한 합성곱 신경망(CNN)의 제1 계층에서 제1 입력 특징 맵(IFM)과 하나 이상의 제1 필터를 수신하는 단계로서, 상기 PE 어레이의 각 PE는 소정 개수(Y1)의 곱셈기를 포함하고, 상기 PE 어레이는 소정 개수(Y2)의 행과 소정 개수(X)의 열로 배열된, 상기 수신하는 단계;

상기 제1 IFM과 상기 하나 이상의 제1 필터에 기초하여 네이티브 텐서 모양을 결정하는 단계로서, 상기 네이티브 텐서 모양은 제1 외부 차원, 내부 차원 및 제2 외부 차원을 포함하고, 상기 네이티브 텐서 모양은 상기 제1 IFM과 상기 하나 이상의 제1 필터를 상기 PE 어레이에 매핑하는, 상기 결정하는 단계;

상기 PE 어레이를 사용하여 합성곱을 수행하기 위한 CNN의 제2 계층에서 제2 IFM과 하나 이상의 제2 필터를 수신하는 단계;

상기 제2 IFM과 상기 하나 이상의 제2 필터에 기초하여 상기 네이티브 텐서 모양을 재형성하는 단계로서, 상기 재형성은 내부 차원을 확대하는 단계, 및 상기 제1 외부 차원과 상기 제2 외부 차원 중 하나를 축소하는 단계를 포함하고, 상기 확대와 축소는 F배로 이루어지는, 상기 재형성하는 단계;

재형성된 네이티브 텐서에 따라 합성곱을 위해 상기 하나 이상의 제2 필터와 상기 제2 IFM을 상기 PE 어레이에 공급하는 단계로서,

상기 제1 외부 차원이 축소되는 것에 응답하여 상기 합성곱은 F개의 라운드 동안 PE의 동일한 행의 출력을 집계하여 부분 합을 얻는 것을 포함하고,

상기 제2 외부 차원이 축소되는 것에 응답하여 상기 합성곱은 PE의 매 F개의 행마다 출력을 집계하여 부분 합을 얻는 것을 포함하는, 상기 공급하는 단계; 및

복수의 상기 부분 합을 집계하여 상기 CNN의 제2 계층에서 상기 합성곱의 출력 텐서를 얻는 단계

를 포함하는 동작을 수행하게 하고, Y1, Y2, X 및 F는 모두 1보다 큰 정수인, 시스템.

청구항 15

청구항 14에 있어서, 상기 CNN의 제2 계층은 상기 CNN의 제1 계층 뒤에 있고, 상기 제2 IFM은 상기 제1 IFM보다 많은 입력 채널을 포함하고, 상기 제1 IFM보다 낮은 해상도를 포함하는, 시스템.

청구항 16

청구항 14에 있어서, 상기 동작은,

상기 하나 이상의 필터의 채널을 복수의 채널 그룹으로 분할하는 단계로서, 각 채널 그룹은 1보다 큰 정수인 고정된 수의 채널을 포함하는, 상기 분할하는 단계; 및

상기 복수의 채널 그룹 각각에서 단 하나의 채널만이 0이 아닌 입력 값을 포함하고, 상기 각 채널 그룹의 다른 채널은 모두 0을 포함하도록 상기 하나 이상의 필터 각각을 가지치기하는 단계

를 추가로 포함하는, 시스템.

청구항 17

청구항 16에 있어서, 상기 동작은,

상기 PE 어레이 내 각 PE와 연관된 버퍼의 깊이를 결정하는 단계;

상기 버퍼의 깊이가 상기 고정된 수보다 큰 것에 응답하여 상기 버퍼를 각 PE에 대한 개인 메모리로 구성하는 단계; 및

상기 버퍼의 깊이가 상기 고정된 수보다 작은 것에 응답하여 상기 PE의 버퍼와 이웃 PE의 하나 이상의 버퍼를 공유 메모리로 결합하는 단계

를 추가로 포함하는, 시스템.

청구항 18

청구항 14에 있어서, 상기 하나 이상의 제2 필터 각각은 2차원(2D) 커널의 복수의 채널을 포함하고, 각 2D 커널은 1×1 또는 3×3의 차원을 갖는, 시스템.

청구항 19

청구항 18에 있어서, 상기 재형성된 네이티브 텐서에 따라 상기 하나 이상의 제2 필터를 상기 PE 어레이에 공급하는 단계는,

상기 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 상기 하나 이상의 제2 필터를 행렬로 변환하는 단계로서, 상기 하나 이상의 제2 필터의 각 2D 커널이 1×1의 차원을 갖는 것에 응답하여, 상기 행렬의 각 행은 상기 하나 이상의 제2 필터의 서로 다른 입력 채널의 가중치를 포함하는, 상기 변환하는 단계; 및

상기 복수의 입력 채널이 한 번에 동시에 처리되도록 상기 제1 행렬의 각 행의 가중치를 PE의 서로 다른 열에 분배하는 단계

를 포함하는, 시스템.

청구항 20

비일시적 컴퓨터 판독 가능 저장 매체로서, 하나 이상의 프로세서가 실행 가능한 명령어를 포함하도록 구성되고, 상기 명령어는, 상기 하나 이상의 프로세서로 하여금,

처리 요소(PE) 어레이를 사용하여 합성곱을 수행하기 위한 합성곱 신경망(CNN)의 제1 계층에서 제1 입력 특징 맵(IFM)과 하나 이상의 제1 필터를 수신하는 단계로서, 상기 PE 어레이의 각 PE는 소정 개수(Y1)의 곱셈기를 포함하고, 상기 PE 어레이는 소정 개수(Y2)의 행과 소정 개수(X)의 열로 배열된, 상기 수신하는 단계;

상기 제1 IFM과 상기 하나 이상의 제1 필터에 기초하여 네이티브 텐서 모양을 결정하는 단계로서, 상기 네이티브 텐서 모양은 제1 외부 차원, 내부 차원 및 제2 외부 차원을 포함하고, 상기 네이티브 텐서 모양은 상기 제1 IFM과 상기 하나 이상의 제1 필터를 상기 PE 어레이에 매핑하는, 상기 결정하는 단계;

상기 PE 어레이를 사용하여 합성곱을 수행하기 위한 CNN의 제2 계층에서 제2 IFM과 하나 이상의 제2 필터를 수신하는 단계;

상기 제2 IFM과 상기 하나 이상의 제2 필터에 기초하여 상기 네이티브 텐서 모양을 재형성하는 단계로서, 상기 재형성은 내부 차원을 확대하는 단계, 및 상기 제1 외부 차원과 상기 제2 외부 차원 중 하나를 축소하는 단계를 포함하고, 상기 확대와 축소는 F배로 이루어지는, 상기 재형성하는 단계;

재형성된 네이티브 텐서에 따라 합성곱을 위해 상기 하나 이상의 제2 필터와 상기 제2 IFM을 상기 PE 어레이에 공급하는 단계로서,

상기 제1 외부 차원이 축소되는 것에 응답하여 상기 합성곱은 F개의 라운드 동안 PE의 동일한 행의 출력을 집계하여 부분 합을 얻는 것을 포함하고,

상기 제2 외부 차원이 축소되는 것에 응답하여 상기 합성곱은 PE의 매 F개의 행마다 출력을 집계하여 부분 합을 얻는 것을 포함하는, 상기 공급하는 단계; 및

복수의 상기 부분 합을 집계하여 상기 CNN의 제2 계층에서 상기 합성곱의 출력 텐서를 얻는 단계

를 포함하는 동작을 수행하게 하고, Y1, Y2, X 및 F는 모두 1보다 큰 정수인, 비일시적 컴퓨터 판독 가능 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 일반적으로 신경망 계산 효율성을 개선하는 것에 관한 것으로, 보다 상세하게는 네이티브 텐서 차원(native tensor dimension)과 연산 모드(operation mode)를 동적으로 조정하여 희소 신경망에 대한 다양한 입력 텐서 모양(input tensor shape)과 연산에 적응하는 것에 관한 것이다.

배 경 기 술

[0002] 거의 모든 딥 러닝 처리 요소(processing element: PE) 어레이는 네이티브 텐서 차원과 연산 모드에 있어서 고정되어 있으며, 일반적으로 컴파일러에 의존하여 다양한 텐서 모양(예를 들어, 입력 특징 맵 또는 필터)과 연산을 처리하기 위해 다양한 중첩 루프 매핑 방식을 사용한다. PE 어레이를 사용하여 PE 어레이의 네이티브 텐서 차원 또는 연산 모드와 호환되지 않는 텐서 모양 또는 연산에 대한 계산을 수행하는 것은 분명히 비효율적이다. 이러한 비호환성은 희소성 특징이 있는 희소 신경망의 경우 PE 어레이의 유연하지 않은 네이티브 텐서 모양과 모드가 많은 수의 0이 있는 텐서를 효율적으로 표현하고 처리할 수 없기 때문에 더욱 악화된다.

발명의 내용

[0003] 본 명세서의 다양한 실시형태는 신경망 계산에서 적응형 텐서 계산 커널(adaptive tensor compute kernel)을 사용하기 위한 시스템, 방법 및 비일시적 컴퓨터 판독 가능 매체를 포함할 수 있다.

[0004] 일 양태에 따르면, 신경망 계산에서 적응형 텐서 계산 커널을 사용하는 방법은, 처리 요소(PE) 어레이를 사용하여 합성곱을 수행하기 위한 합성곱 신경망(convolutional neural network: CNN)의 제1 계층에서 제1 입력 특징 맵(input feature map: IFM)과 하나 이상의 제1 필터를 수신하는 단계로서, PE 어레이의 각 PE는 소정 개수(Y1)의 곱셈기를 포함하고, PE 어레이는 소정 개수(Y2)의 행과 소정 개수(X)의 열로 배열된, 수신하는 단계; 제1 IFM과 하나 이상의 제1 필터에 기초하여 네이티브 텐서 모양을 결정하는 단계로서, 네이티브 텐서 모양은 제1 외부 차원, 내부 차원 및 제2 외부 차원을 포함하고, 네이티브 텐서 모양은 제1 IFM과 하나 이상의 제1 필터를 PE 어레이에 매핑하는, 결정하는 단계; PE 어레이를 사용하여 합성곱을 수행하기 위한 CNN의 제2 계층에서 제2 IFM과 하나 이상의 제2 필터를 수신하는 단계; 제2 IFM과 하나 이상의 제2 필터에 기초하여 네이티브 텐서 모양을 재형성(reshaping)하는 단계로서, 재형성은 내부 차원을 확대하는 단계, 및 제1 외부 차원과 제2 외부 차원 중 하나를 축소하는 단계를 포함하고, 확대와 축소는 F배로 이루어지는, 재형성하는 단계; 재형성된 네이티브 텐서에 따라 합성곱을 위해 하나 이상의 제2 필터와 제2 IFM을 PE 어레이에 공급하는 단계로서, 제1 외부 차원이 축소되는 것에 응답하여 합성곱은 F개의 라운드 동안 PE의 동일한 행의 출력을 집계하여 부분 합을 얻는 것을 포함하고, 제2 외부 차원이 축소되는 것에 응답하여 합성곱은 PE의 매 F개의 행마다 출력을 집계하여 부분 합을 얻는 것을 포함하는, 공급하는 단계; 및 복수의 부분 합을 집계하여 CNN의 제2 계층에서 합성곱의 출력 텐

서를 얻는 단계를 포함할 수 있고, Y1, Y2, X 및 F는 모두 1보다 큰 정수이다.

[0005] 일부 실시형태에서, CNN의 제2 계층은 CNN의 제1 계층 뒤에 있고, 제2 IFM은 제1 IFM보다 많은 입력 채널을 포함하고, 제1 IFM보다 낮은 해상도를 포함한다.

[0006] 일부 실시형태에서, 하나 이상의 제2 필터 각각은 2차원(2D) 커널의 복수의 채널을 포함하고, 각 2D 커널은 1×1 또는 3×3의 차원을 갖는다.

[0007] 일부 실시형태에서, 재형성된 네이티브 텐서에 따라 하나 이상의 제2 필터를 PE 어레이에 공급하는 단계는, 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 하나 이상의 제2 필터를 행렬로 변환하는 단계로서, 하나 이상의 제2 필터의 각 2D 커널이 1×1의 차원을 갖는 것에 응답하여, 행렬의 각 행은 하나 이상의 제2 필터의 서로 다른 입력 채널의 가중치를 포함하는, 변환하는 단계; 및 복수의 입력 채널이 한 번에 동시에 처리되도록 행렬의 각 행의 가중치를 PE의 서로 다른 열에 분배하는 단계를 포함한다.

[0008] 일부 실시형태에서, 재형성된 네이티브 텐서에 따라 하나 이상의 제2 필터를 PE 어레이에 공급하는 단계는, 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 하나 이상의 제2 필터를 행렬로 변환하는 단계로서, 하나 이상의 제2 필터의 각 2D 커널이 3×3의 차원을 갖고 9개의 가중치를 포함하는 것에 응답하여, 9개의 가중치는 행렬의 동일한 행에 배치되는, 변환하는 단계; 및 동일한 채널의 가중치가 한 번에 동시에 처리되도록 행렬의 동일한 행의 9개의 가중치를 PE의 서로 다른 열에 분배하는 단계를 포함한다.

[0009] 일부 실시형태에서, 재형성된 네이티브 텐서에 따라 IFM을 PE 어레이에 공급하는 단계는, 재형성된 네이티브 텐서의 내부 차원과 제2 외부 차원에 따라 IFM을 행렬로 변환하는 단계; 및 행렬의 열에 대응하는 IFM의 입력 값을 PE의 행의 버퍼에 공급하는 단계를 포함한다.

[0010] 일부 실시형태에서, 방법은 하나 이상의 필터의 채널을 복수의 채널 그룹으로 분할하는 단계로서, 각 채널 그룹은 1보다 큰 정수인 고정된 수의 채널을 포함하는, 분할하는 단계; 및 복수의 채널 그룹 각각에서 몇 개의 채널만이 0이 아닌 입력 값을 포함하고, 각 채널 그룹의 다른 채널은 모두 0을 포함하도록 하나 이상의 필터 각각을 가지치기(pruning)하는 단계를 추가로 포함할 수 있다.

[0011] 일부 실시형태에서, 방법은 PE 어레이의 각 PE와 연관된 버퍼의 깊이를 결정하는 단계; 버퍼의 깊이가 고정된 수보다 큰 것에 응답하여 버퍼를 각 PE에 대한 개인 메모리(private memory)로 구성하는 단계; 및 버퍼의 깊이가 고정된 수보다 작은 것에 응답하여 PE의 버퍼와 이웃 PE의 하나 이상의 버퍼를 공유 메모리로 결합하는 단계를 추가로 포함할 수 있다.

[0012] 일부 실시형태에서, 각 PE의 개인 메모리는 PE 내 소정 개수(Y1)의 곱셈기에 의해 검색 가능한 입력 값을 저장한다.

[0013] 일부 실시형태에서, 공유 메모리는 PE 내 소정 개수(Y1)의 곱셈기와 하나 이상의 이웃 PE에 의해 검색 가능한 입력 값을 저장한다.

[0014] 일부 실시형태에서, PE의 각 행은 각 PE 내의 곱셈기의 수(Y1)에 대응하는 수(Y1)의 덧셈기 트리(adder tree)와 결합되고, 각 PE 내의 각 곱셈기는 집계를 위해 대응하는 덧셈기 트리에 곱셈 출력을 보낸다.

[0015] 일부 실시형태에서, 하나 이상의 제2 필터 각각은 복수의 0이 아닌 가중치를 포함하고, 합성곱을 위해 하나 이상의 제2 필터를 PE 어레이에 공급하는 단계는, 각 0이 아닌 가중치를 대응하는 PE의 곱셈기에 0이 아닌 가중치와 대응 인덱스를 포함하는 인덱스-값 쌍으로 공급하는 단계를 포함하고; 합성곱은, 인덱스에 따라 대응하는 PE의 버퍼로부터 입력 값을 검색하는 단계; 및 검색된 값과 0이 아닌 가중치를 곱셈기에 보내어 출력을 얻는 단계; 및 대응하는 PE와 같은 행에 있는 다른 PE의 다른 곱셈기에 의해 생성된 출력과 함께 집계를 위해 대응하는 덧셈기 트리에 출력을 보내는 단계를 포함한다.

[0016] 일부 실시형태에서, 각 PE 내 소정 개수(Y1)의 곱셈기는 병렬로 데이터를 처리하고, PE 어레이 내의 PE는 병렬로 데이터를 처리한다.

[0017] 또 다른 양태에 따르면, 시스템은 하나 이상의 프로세서와 하나 이상의 비일시적 컴퓨터 판독 가능 메모리를 포함할 수 있고, 비일시적 컴퓨터 판독 가능 메모리는 하나 이상의 프로세서에 결합되고, 하나 이상의 프로세서에 의해 실행 가능한 명령어를 포함하도록 구성되고, 명령어는 시스템으로 하여금 본 명세서에 설명된 방법 중 임의의 방법을 수행하도록 할 수 있다.

[0018] 또 다른 양태에 따르면, 비일시적 컴퓨터 판독 가능 저장 매체는 하나 이상의 프로세서에 의해 실행 가능한 명

령어를 포함하도록 구성될 수 있고, 명령어는 하나 이상의 프로세서로 하여금 본 명세서에 설명된 방법 중 임의의 방법을 수행하도록 할 수 있다.

[0019] 본 명세서에 개시된 시스템, 방법 및 비일시적 컴퓨터 판독 가능 매체의 이러한 특징 및 기타 특징과, 구조의 관련 요소의 동작 방법과 기능, 및 부품의 조합과 제조 경제성은 첨부된 도면을 참조하는 다음 설명 및 첨부된 청구범위를 고려하면 더욱 명확해질 것이며, 도면에서 동일한 참조 부호는 여러 도면에서 대응하는 부분을 나타낸다. 그러나, 도면은 예시 및 설명의 목적만을 위한 것일 뿐, 본 발명의 한계를 정하려고 의도된 것이 아니라는 점을 명확히 이해해야 한다.

도면의 간단한 설명

[0020] 도 1은 다양한 실시형태에 따라 PE 어레이에서 신경망 계산을 처리하기 위한 예시적인 시스템 다이어그램을 도시한다.

도 2는 다양한 실시형태에 따라 PE 어레이의 예시적인 아키텍처 다이어그램을 도시한다.

도 3은 다양한 실시형태에 따라 네이티브 텐서 모양을 사용하는 PE 어레이에서 예시적인 신경망 계산을 도시한다.

도 4a는 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 PE 어레이에서 예시적인 신경망 계산을 도시한다.

도 4b는 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 신경망 계산을 위한 클러스터 간 덧셈기를 갖는 예시적인 PE 어레이를 도시한다.

도 5a는 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 PE 어레이에서 또 다른 예시적인 신경망 계산을 도시한다.

도 5b는 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 신경망 계산을 위한 클러스터 내 덧셈기를 갖는 또 다른 예시적인 PE 어레이를 도시한다.

도 6a는 다양한 실시형태에 따라 1×1 텐서 연산 모드를 사용하는 예시적인 신경망 계산을 도시한다.

도 6b는 다양한 실시형태에 따라 3×3 텐서 연산 모드를 사용하는 예시적인 신경망 계산을 도시한다.

도 7은 다양한 실시형태에 따라 PE 어레이에서 적응형 텐서 모양과 3×3 텐서 연산 모드를 사용하는 신경망 계산을 위한 예시적인 방법을 도시한다.

도 8은 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 신경망 계산을 위한 예시적인 방법을 도시한다.

도 9는 본 명세서에 설명된 실시형태 중 임의의 실시형태를 구현할 수 있는 예시적 컴퓨터 시스템을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 명세서에 설명된 실시형태는 적응형 텐서 모양과 연산 모드를 사용하는 PE 어레이에서 신경망 계산을 위한 방법, 시스템, 장치를 제공한다. 다음 설명에서 적응형 텐서 계산 커널은 다양한 모양의 입력 특징 맵(IFM)과 가중치 텐서(예를 들어, 필터)를 처리하기 위해 다수의 네이티브 텐서 차원과 연산 모드를 갖는 것으로 설명된다. 입력 및 출력 텐서 모양과 연산 모드에 따라 적응형 텐서 계산 커널(적응형 네이티브 텐서라고도 함)의 차원과 연산 모드는 병렬 처리를 위해 PE 어레이의 기본 하드웨어 자원을 완전히 활용하도록 동적으로 조정될 수 있다.

[0022] 이러한 적응형 텐서 계산 커널은 세 가지 기술적 솔루션을 제공하여 (배경 기술 란에서 언급된) 신경망 계산의 기술적 과제를 해결한다. 첫째, 적응형 텐서 계산 커널은 입력/가중치 텐서의 다양한 모양에 따라 형태를 조정할 수 있다. 입력/가중치 텐서의 다양한 모양은 다른 신경망에도 존재할 뿐만 아니라 동일한 신경망 파이프라인 내에도 존재할 수 있다. 예를 들어, 신경망에서 처음 몇 개의 계층을 처리할 때 텐서는 일반적으로 고해상도(더 많은 높이와 폭)로 구성되지만 더 적은 입력 및 출력 채널로 구성되고; 신경망에서 마지막 몇 개의 계층을 처리할 때 텐서는 저해상도(더 적은 높이와 폭)로 구성되지만 더 많은 입력 및 출력 채널로 구성될 수 있다. 이는 신경망의 처음 몇 개의 계층이 입력 특징 맵에서 특징을 추출하는 데 보다 중점을 두는 반면, 신경망의 마지막 몇 개의 계층은 추출된 특징 간의 기본 상관 관계를 학습하는 데 보다 중점을 두기 때문일 수 있다.

[0023] 둘째, 적응형 텐서 계산 커널은 1×1 텐서 연산 모드와 3×3 텐서 연산 모드의 두 가지 다른 텐서 연산 모드를 지원할 수 있다. 행렬 곱셈이 1×1 커널 합성곱을 포함하는(예를 들어, 가중치 텐서의 각 커널은 1×1 모양을 가짐) 신경망 계층은 1×1 텐서 연산 모드에 매핑될 수 있고, 임의의 다른 합성곱(3×3, 5x5, 7x7 등)은 3×3 텐서 연산 모드에 매핑될 수 있다. 이러한 다른 텐서 연산 모드는 가중치 텐서 모양에 기초하여 런타임 동안 동적으로 결정될 수 있다.

[0024] 셋째, 기본 PE 어레이는 희소 신경망의 다른 압축 비율과 희소성 세분성(sparsity granularity)을 지원하기 위해 각 PE 내부 버퍼(예를 들어, 레지스터 파일)를 다르게 구성할 수 있다. 희소 신경망이 보다 미세한 세분성(finer granularity)으로 가지치기되는 경우(예를 들어, 작은 수의 입력 채널에서 하나 이상의 0이 아닌 입력 채널을 선택하는 것, 작은 수는 임계값보다 작은 것임), 각 PE 내의 레지스터 파일은 (예를 들어, 대응하는 PE에 의해서만 사용되는) 개인 메모리로 구성될 수 있다. 희소 신경망이 거친 세분성(coarse granularity)으로 가지치기되는 경우(예를 들어, 많은 수의 입력 채널에서 하나 이상의 0이 아닌 입력 채널을 선택하는 것, 큰 수는 임계값보다 큰 것임), 이웃하는 PE의 다수의 레지스터 파일은 이웃하는 PE에서 공유하는 다중 포트 메모리로 구성될 수 있다.

[0025] 다음 설명에서는 본 발명의 구체적이고 비-제한적인 실시형태를 도면을 참조하여 설명한다. 본 명세서에 개시된 임의의 실시형태의 특정 특징과 양태는 본 명세서에 개시된 임의의 다른 실시형태의 특정 특징 및 양태와 함께 사용 및/또는 결합될 수 있다. 또한 이러한 실시형태는 예시적인 것일 뿐, 본 발명의 범위 내에 작은 수의 실시형태를 예시하는 것에 불과하다는 점도 이해해야 한다. 본 발명이 속하는 기술 분야의 당업자에게 명백한 다양한 변경과 수정은 첨부된 청구범위에서 추가로 한정되는 본 발명의 정신, 범위 및 구상 내에 있는 것으로 간주된다.

[0026] 도 1은 다양한 실시형태에 따라 PE 어레이에서 신경망 계산을 처리하기 위한 예시적인 시스템 다이어그램을 도시한다. 도 1의 다이어그램은 PE 어레이가 있는 파이프라인에서 실행되는 일반적인 신경망 계산 작업 흐름을 도시한다. 본 명세서에 설명된 실시형태는 도 1의 신경망 계산 또는 다른 적합한 환경의 일부로 구현될 수 있다.

[0027] 신경망(예를 들어, CNN) 내의 주어진 (예를 들어, 합성곱) 계층에서, 하나 이상의 입력 특징 맵(IFM)(120)은 입력 소스(예를 들어, 입력 이미지) 또는 이전 계층(예를 들어, 이전 계층의 텐서 출력)으로부터 얻어질 수 있고, 하나 이상의 가중치 텐서(110)는 IFM을 통해 합성곱을 수행하여 다양한 특징을 추출하는 데 사용될 수 있다. 합성곱 프로세스는 PE 어레이(160)라고 하는 처리 요소(PE)의 어레이에서 병렬로 수행될 수 있다. 각 PE는 처리 능력과 저장 용량(예를 들어, 버퍼 또는 캐시)을 갖는 프로세서를 의미할 수 있다. PE는 상호 연결 와이어로 PE 어레이에 특정 방식으로 배열될 수 있으며 런타임에 동적으로 재배열되지 않을 수 있다. PE 어레이는 신경망의 다른 계층이나 다른 신경망과 다른 사용 사례에 걸쳐 계산하는 데 재사용되고 관련될 수 있다. PE 어레이의 고정된 내부 PE 배열과 (IFM 및/또는 가중치 텐서에서) 잠재적으로 무한하게 다양한 텐서 모양 간의 비호환성은 일반적으로 비효율적인 자원 활용과 최적이 아닌 병렬 처리로 이어진다.

[0028] 도 1을 참조하면, 일부 실시형태에서, IFM(120)은 IFM 캐시(140)에 저장될 수 있고, 가중치 텐서(110)는 PE 어레이(160)의 소비를 위해 가중치 캐시(130)에 저장될 수 있다. PE 어레이(160)는 PE의 행렬(예를 들어, X x Y)을 포함할 수 있고, 각 PE는 병렬 처리를 위한 복수의 곱셈기를 포함할 수 있다. 일부 실시형태에서, IFM 캐시(140)의 각 IFM은 PE 어레이(160)에서 계산을 용이하게 하기 위해 행렬 변환 계층(150)을 거칠 수 있다. 행렬 변환은 im2col 도구를 사용하여 IFM을 원래 HWC 형식(H는 높이를 나타내고, W는 가중치를 나타내고, C는 채널을 나타냄)으로부터 RSC 형식(R은 행을 나타내고, S는 열을 나타내고, C는 채널을 나타냄)으로 재형성하는 Toeplitz 행렬 변환을 포함할 수 있고, 여기서 RSC 형식은 가중치 텐서 모양에 기초하여 결정된다. 여기서, 변환은 IFM의 입력 값을 복제하고 배열하여 변환된 IFM을 형성하여, 변환된 IFM과 가중치 텐서 간의 행렬 곱셈이 PE 어레이(160)에서 PE 간의 종속성을 최소화하여 병렬 방식으로 PE 어레이(160)에서 실행될 수 있도록 하는 것이다. 일부 실시형태에서, PE 어레이(160)에서 병렬 합성곱의 각 라운드는 복수의 부분 합을 생성할 수 있으며, 이 복수의 부분 합은 누적 버퍼(170)에서 집계되어 하나 이상의 출력 값을 생성할 수 있다. 출력 값은 결국 현재 계층에서 생성된 출력 텐서의 일부가 될 수 있다.

[0029] 도 2는 다양한 실시형태에 따라 PE 어레이의 예시적인 아키텍처 다이어그램을 보여준다. 도 2의 PE 어레이에서 PE의 배열은 설명을 위한 것이며, 사용 사례에 따라 다른 방식으로 구현될 수 있다.

[0030] 도 2의 좌측 부분에 도시된 바와 같이, PE 어레이(200)는 PE의 행렬을 포함할 수 있다. 도 2의 우측 부분에 도시된 바와 같이, 각 PE(240)는 복수의 곱셈기(MUL 게이트)를 포함할 수 있다. 각 PE(240) 내의 곱셈기는 병렬로 작동할 수 있고, PE 어레이(220) 내의 PE는 병렬로 작동할 수 있다. 참조의 편의를 위해, 다음 설명은 PE 어레

이(200) 내의 PE의 열(220)의 수를 X로 표시하고, PE 어레이(200) 내의 PE의 행(210)의 수를 Y2로 표시하고, 각 PE(240) 내의 곱셈기의 수를 Y1로 표시한다. PE(210)의 각 행은 PE 클러스터라고 할 수 있으며, 각 PE 클러스터는 PE 클러스터 내의 곱셈기에서 생성된 부분 합을 집계하기 위해 Y1개의 덧셈기 트리(230)에 결합될 수 있다. 즉, PE 클러스터 내의 각 PE(240)의 제1 곱셈기는 집계를 위해 제1 덧셈기 트리(230)에 결합되고, PE 클러스터 내의 각 PE(240)의 제2 곱셈기는 집계를 위해 제2 덧셈기 트리(230)에 결합되고, 이와 같이 계속된다. 모든 PE 클러스터(총 Y1 x Y2 덧셈기 트리)에 걸친 덧셈기 트리(230)의 집계 결과는 집계를 위해 덧셈기(250)에 공급될 수 있다. 덧셈기(250)는 네트워크 온칩(NoC) 서브시스템의 일부인 숫자의 덧셈을 수행하는 디지털 회로를 지칭할 수 있다.

[0031] 일부 실시형태에서, PE 어레이(200)는 가중치를 PE에 브로드캐스트(broadcast)할 수 있다. 희소 신경망의 경우, 가중치의 대부분이 0이므로 PE에 브로드캐스트되는 가중치는 모두 0이 아닌 가중치이다. 0이 아닌 가중치는 가중치 텐서 내의 임의의 위치에서 올 수 있으므로, 브로드캐스트되는 각 가중치는 가중치 값을 포함할 뿐만 아니라 가중치 값의 위치 정보를 나타내는 인덱스, 즉, (인덱스, 가중치 값)과 같은 인덱스-값 쌍을 포함할 수 있다. 인덱스에 기초하여, 각 PE(240)는 IFM으로부터 대응하는 입력 값을 검색하여 가중치 값과 곱셈을 수행할 수 있다. 곱셈 결과는 대응하는 덧셈기 트리에 공급될 수 있다. 도 1에 도시된 바와 같이, 제1 곱셈기(MUL1)는 (인덱스1, 값1)의 형태로 가중치를 수신하고, (IFM을 저장하는) IBUF(260)으로부터 인덱스1에 기초하여 입력 값(IFM1)을 검색하고, 입력 값(IFM1)과 값1에 기초하여 곱셈을 수행하고, 집계를 위해 결과를 덧셈기 트리 1(예를 들어, PE가 위치된 PE 클러스터에 대한 Y1 덧셈기 트리(230)의 제1 덧셈기 트리)로 보낼 수 있다.

[0032] 도 3은 다양한 실시형태에 따른 네이티브 텐서 모양을 사용하는 PE 어레이에서 예시적인 신경망 계산을 도시한다. 도 3의 예시적인 계산은 변환된 가중치 텐서 A(310)와 변환된 IFM 텐서 B(320) 간의 행렬 곱셈을 포함하고, 이는 출력 특징 맵(OFM) 텐서 C(330)를 생성한다. 행렬 곱셈은 X 및 Y 차원을 갖는 PE 어레이(340)에 대응하는 네이티브 텐서 모양을 사용한다.

[0033] 일부 실시형태에서, 변환된 가중치 텐서 A(310)는 RSC 형식(3차원)의 모든 가중치 텐서를 m'*k'로 표시된 2차원 행렬(예를 들어, 서로 다른 채널의 가중치가 동일한 채널에 재배열됨)로 집계하여 얻어질 수 있고, 여기서 m'은 가중치 텐서의 수(일반적으로 K로 표시됨)에 의해 결정되는 출력 채널의 수이고, k'은 각 가중치 텐서의 R, S 및 C 차원(R 및 S는 가중치 텐서의 각 커널의 차원을 나타내고, C는 입력 채널의 수를 나타냄)의 곱이다.

[0034] 일부 실시형태에서, 변환된 IFM 텐서 B(320)는 RSC 형식에 기초하여 HWC 형식(3차원)의 모든 IFM을 k'*n'으로 표시되는 2차원 행렬로 집계하여 얻어질 수 있고, 여기서 k'은 여전히 각 가중치 텐서의 R, S 및 C 차원의 곱이고, n'은 IFM의 H 및 W 차원(H는 IFM의 높이를 나타내고, W는 IFM의 폭을 나타냄)의 곱이다. 행렬(m'*k')(가중치 텐서 A(310))과 행렬(k'*n')(IFM B(320))의 행렬 곱은 m'*n'의 행렬인 OFM 텐서 C(330)를 생성할 수 있다.

[0035] 위의 변환을 통해, 변환된 가중치 텐서 A(310) 및 변환된 IFM 텐서 B(320)는 병렬 처리를 위해 PE 어레이(340)의 PE에 매핑될 수 있다. PE 어레이(340)가 PE의 Y2개의 행, PE의 X개의 열을 포함하고, 각 PE가 Y1개의 곱셈기를 포함한다고 가정하면, 텐서(A 및 B)는, 텐서 A(310)와 텐서 B(320)의 내부 차원(k')이 PE 어레이(340)의 X(행) 차원에 매핑될 수 있고, 즉 X=k'= R * S * C이고, 텐서 A 및 B의 외부 차원의 곱셈(m'* n')이 PE 어레이(340)의 Y (열) 차원에 매핑될 수 있는 방식으로 PE 어레이(340)에 매핑될 수 있다. PE의 각 열이 Y1 * Y2개의 곱셈기를 포함하고 있으므로, 위의 매핑은 Y = m'*n'=K*H*W 곱셈이 Y1 * Y2개의 곱셈기에 의해 병렬로 처리됨을 의미한다. 예를 들어, 각 PE 내에서 하나의 곱셈기는 동일한 출력 채널(예를 들어, 모든 가중치 텐서의 동일한 위치의 가중치)에 대응하는 가중치를 처리하고, 즉, Y1= K= m'이고, PE의 각 열은 H*W 가중치를 병렬로 처리하고, 즉, Y2= H*W= n'이다.

[0036] 위 설명에서 네이티브 텐서 모양 m'*k'*n'은 PE 어레이(340) 내의 PE에 작업 부하(예를 들어, 곱셈을 위한 대응하는 입력 값과 가중치의 쌍)를 매핑하기 위해 고정되고, 여기서 X=k', Y1*Y2=m'*n'이다. 이것은 네이티브 텐서 모양이 PE 어레이 내의 PE의 레이아웃에 기초하여 결정된다는 것을 의미한다. PE 어레이의 레이아웃이 고정되면 네이티브 텐서 모양이 고정된다. 모든 수신 텐서(예를 들어, IFM와 필터/가중치 텐서)는 고정된 네이티브 텐서 모양에 따라 변환되어야 한다. 그러나, 실제 응용에서 수신 텐서는 모양이 다를 수 있으며, 변환이 PE 어레이(340) 내의 PE 레이아웃이 아닌 수신 텐서의 모양에 기초할 때 최적의 병렬성을 달성할 수 있다. 많은 경우, PE 레이아웃에 기초하여 결정된 고정된 네이티브 텐서 모양을 사용한 변환이 작업 부하를 PE에 매핑할 수 있다 하더라도, 이것은 특정 PE 간에 직렬화된 종속성(예를 들어, 하나의 PE가 다른 PE의 출력을 기다려야 함)을 일으킬 수도 있다. 다음 설명은 IFM과 필터의 차원에 기초하여 결정되고 동시에 병렬성을 극대화하기 위해 작업 부하를 PE에 매핑하는 적응형 네이티브 텐서 모양을 사용한 변환을 설명한다.

[0037] 도 4a는 다양한 실시형태에 따라 적응형 텐서 모양을 사용한 PE 어레이에서 예시적인 신경망 계산을 도시한다. (도 3에서) 위에서 설명된 바와 같이, 입력 텐서(IFM)와 가중치 텐서가 고정된 네이티브 텐서 모양 m'*k'*n'을 사용하여 (즉, 행렬 A와 B를 바라는) 행렬 A(410)와 행렬 B(420)로 변환될 수 있는 경우, 변환된 텐서는 대응하는 PE 어레이에 분배될 수 있다. 그러나, 실제 응용에서, 곱셈을 위한 IFM과 가중치 텐서(예를 들어, CNN의 다른 계층에서, 다른 수준의 희소화를 거친 텐서)는 PE 어레이에 완벽하게 매핑되지 않을 수 있는 다양한 모양을 가질 수 있다. 고정된 네이티브 텐서 모양을 사용하여 텐서를 강제로 변환하면 일부 PE가 유휴 상태가 되거나 계산 동안 순차적 종속성이 발생할 수 있다. 예를 들어, 동일한 합성곱 신경망(CNN) 내에서, 처음 몇 개의 CNN 계층의 텐서는 더 적은 입력 채널 수(C=16)와 함께 높은 해상도(예를 들어, H*W=64)를 가질 수 있고, 마지막 몇 개의 CNN 계층의 텐서는 더 많은 입력 채널 수(C=64)와 함께 낮은 해상도(예를 들어, H*W=16)를 가질 수 있다. 여기서, "더 적은"과 "더 많은"은 임계값을 기준으로 결정된다. 이것은 동일한 CNN 내의 합성곱 프로세스라도 서로 다른 모양의 텐서를 경험할 수 있다는 것을 의미한다.

[0038] 일부 실시형태에서, 네이티브 텐서 모양은 입력 텐서와 가중치 텐서의 변하는 모양을 수용하기 위해 동적으로 재형성될 수 있다. 예를 들어, 입력 텐서가 (예를 들어, 처음 몇 개의 CNN 계층에서) 더 적은 입력 채널과 고해상도(더 많은 픽셀)로부터 (예를 들어, 마지막 몇 개의 CNN 계층에서) 더 많은 입력 채널과 저해상도로 변경되는 경우, 네이티브 텐서 모양은 그에 따라 재형성될 수 있다. 일부 실시형태에서, 네이티브 텐서 모양은 제1_외부_차원, 내부_차원 및 제2_외부_차원으로 표시되는 세 개의 차원을 갖는다. 처음 두 개의 차원(제1_외부_차원 및 내부_차원)은 가중치 텐서를 행렬로 변환하는 데 사용될 수 있고, 마지막 두 개의 차원(내부_차원 및 제2_외부_차원)은 IFM을 행렬로 변환하는 데 사용될 수 있다. 변환된 행렬은 가중치와 입력 값이 PE 어레이에 매핑되는 방식(예를 들어, 최적의 병렬성에 도달하기 위해 가중치와 입력 값을 분배하는 방식)에 대한 가이드라인을 제공할 수 있다.

[0039] 일부 실시형태에서, 이전 텐서가 매핑과 변환을 위해 네이티브 텐서 모양 m'*k'*n'을 사용했고, 수신 텐서가 이전 텐서에 비해 더 많은 입력 채널과 함께 더 낮은 해상도를 갖는다고 가정하면, 네이티브 텐서 모양의 세 개의 차원은 m'* (F*k') * (n'/F)로 재형성될 수 있고, 여기서 F는 1보다 큰 정수이고, 스케일 조정 인수(scaling factor)를 나타내고, 처음 두 개의 차원(즉, 제1 외부 차원(m')과 내부 차원(F*k'))은 가중치 텐서 행렬(420)을 나타내고, 그 다음 두 개의 차원(즉, 내부 차원(F*k')과 제2 외부 차원(n'/F))은 합성곱을 위한 IFM 텐서 행렬(422)을 나타낸다. 즉, 네이티브 텐서 모양은 내부 차원을 F배로 확대할 수 있고, (IFM 텐서 행렬에 대응하는) 제2 외부 차원을 F배로 축소할 수 있다. 이러한 재형성 방식은 다음 설명에서 k'&n' 재형성이라고 지칭될 수 있다. 일부 실시형태에서, F는 2, 4, 8 등 중 하나일 수 있다.

[0040] 일부 실시형태에서, 재형성된 텐서 모양의 내부 차원(F*k')은 가중치 텐서 행렬(420)과 IFM 텐서 행렬(422)에 의해 공유되고(예를 들어, 이들 행렬은 동일한 내부 차원을 가짐), PE 어레이의 PE의 열의 수에 대응하고; 제1 외부 차원(예를 들어, 가중치 텐서 행렬(420)의 외부 차원(m'))은 각 PE 내의 곱셈기의 수에 대응하고; 제2 외부 차원(예를 들어, IFM 텐서 행렬(422)의 외부 차원(n'/F))은 PE 어레이의 PE의 행의 수에 대응한다. 여기서, "대응하는"은 변환된 텐서 행렬의 가중치와 입력 값이 PE 어레이에 분배되는 방식을 나타내는 매핑 관계를 의미한다. 예를 들어, 가중치 텐서 행렬(420)의 각 외부 차원(예를 들어, 각 열)의 가중치는 병렬 처리를 위해 단일 PE 내의 곱셈기에 분배될 수 있고; IFM 텐서 행렬(422)의 각 외부 차원(예를 들어, 각 열)의 입력 값은 PE 어레이의 PE의 행에 걸쳐 분배될 수 있다.

[0041] 도 4a에 도시된 바와 같이, 이러한 재형성된 네이티브 텐서 모양을 사용하면, 가중치 텐서 행렬(410)은 내부 차원을 F배로 확대하고, 외부 차원(m')을 동일하게 유지하여 새로운 가중치 텐서 행렬(420)을 형성함으로써 재형성될 수 있다. 즉, 가중치 텐서 행렬의 내부 차원은 k'=R*S*C(예를 들어, 410의 행렬 A)로부터 F * k'= R * S * (F * C)(예를 들어, 420의 행렬 A)로 변경되어, 새로운 행렬(420)은 (C로부터 F*C로) 더 많은 입력 채널을 지원할 수 있다. 유사하게, IFM 행렬(412)은 외부 차원을 F배로 축소하고, 가중치 텐서 행렬(420)의 확대된 내부 차원과 같은 방식으로 내부 차원을 확대하여, 새로운 IFM 텐서 행렬(422)을 형성할 수 있다. 즉, IFM 텐서 행렬의 내부 차원은 k'=R*S*C(예를 들어, 412의 행렬 B)로부터 F *k'=R * S * (F * C)(예를 들어, 422의 행렬 B)로 변경되고, IFM 텐서 행렬의 외부 차원은 n'(예를 들어, 412의 행렬 B)로부터 n'/F(예를 들어, 422의 행렬 B)로 변경되어, 새로운 행렬(422)은 더 적은 픽셀을 지원할 수 있다. 따라서, 새로운 행렬(420 및 422)은 더 많은 입력 채널과 함께 낮은 해상도를 갖는 마지막 몇 개의 CNN 계층의 텐서를 나타내는 데 보다 적합하다. 일부 실시형태에서, "처음 몇 개의 CNN 계층"과 "마지막 몇 개의 CNN 계층"은 각각 CNN 구조의 시작에서부터 CNN 계층의 제1 수와, CNN 구조의 끝에서부터 CNN 계층의 제2 수를 나타낼 수 있다.

[0042] 일례로서, 제1 새로운 CNN 계층의 텐서는 더 적은 수의 입력 채널(C=16)과 함께 고해상도(H*W=64)를 가질 수 있

다. 여기서, "더 적은 수"는 임계값보다 작은 수를 의미하며, 임계값은 기본 PE 어레이에 따라 구성된 컴파일러에 의해 결정될 수 있다. 처음 몇 개의 CNN 계층의 이러한 텐서에 대한 네이티브 텐서 모양은 합성곱이 1*1 커널에 기초한다고 가정할 때 m'=K=16, k'=1*1*16, 및 n'=64의 모양을 가질 수 있다. 합성곱이 마지막 몇 개의 CNN 계층으로 진행될 때, 텐서는 (예를 들어, 임계값보다 큰) 더 많은 입력 채널(C=64)과 함께 낮은 해상도(H*W=16)를 가질 수 있고, 네이티브 텐서 모양은 m'=K=16, k'=1*1*64, 및 n'=16으로 재형성될 수 있다.

[0043] 위에서 설명한 k'&n'재형성을 사용하여 텐서를 변환한 후, 변환된 텐서 행렬(420 및 422)은 병렬 처리를 위해 PE 어레이에 분배될 수 있다. 도 4b는 다양한 실시형태에 따라 k'&n' 재형성 기반 적응형 텐서 모양을 사용한 신경망 계산을 위한 클러스터 간 덧셈기를 갖는 예시적인 PE 어레이를 도시한다. 도 4b에 도시된 PE 어레이를 사용하는 병렬 처리 방식은 도 4a에 설명된 네이티브 텐서의 k'&n' 재형성에 대응할 수 있다. 일관성을 위해, PE 어레이는 Y2개의 행과 X개의 열의 PE를 갖고, 각 PE는 Y1개의 곱셈기를 갖는다고 여전히 가정한다.

[0044] k'&n' 재형성을 통해 가중치 텐서 행렬과 IFM 텐서 행렬의 내부 차원은 F배로 확대되고, IFM 텐서 행렬의 외부 차원은 F배로 축소된다. 가중치와 입력 값을 PE 어레이에 분배하는 것은 가중치 텐서 행렬의 동일한 행의 (즉, 확대된 내부 차원/행을 따른) 가중치와, IFM 텐서 행렬의 동일한 열의 (즉, 확대된 내부 차원/열을 따른) 입력 값이 행별로 PE에 할당되게 할 수 있다. 이것은 이러한 가중치와 입력 값 쌍이 PE의 F개의 행에 걸쳐 분배될 수 있다는 것을 의미한다. 따라서, PE 어레이는 PE의 각 행에서 생성된 출력을 집계하여 합성곱 프로세스의 부분 합을 얻기 위해 클러스터 간(즉, PE 클러스터 또는 행 사이에) 덧셈기(400)를 가질 수 있다. 각 클러스터 간 덧셈기(400)는 PE의 행에서 Y1개의 덧셈기 트리의 출력을 Y1개의 부분 합으로 집계할 수 있다. 그런 다음 이러한 부분 합은 집계되어 합성곱의 결과로 출력 텐서를 구성할 수 있다. 이 프로세스 동안, 부분 합의 총 수는 Y1 * (Y2/F)이며, 이는 출력 채널 수(예를 들어, 합성곱 프로세스의 출력 텐서의 채널의 수)가 Y1이고 출력 픽셀 수가 Y2/F=H*W/F임을 의미한다.

[0045] 도 5a는 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 PE 어레이에서의 또 다른 예시적인 신경망 계산을 나타낸다. 위에서 설명한 k'&n' 재형성과 비교하여, 네이티브 텐서 모양은 가중치 텐서의 희소성 정도에 기초하여 동적으로 재형성될 수도 있다. 많은 실제 응용에서, 합성곱 프로세스의 가중치 텐서는 계산 효율성을 개선하고 신경망의 풋프린트를 줄이기 위해 가지치기 또는 희소화될 수 있다. 신중하게 가지치기된 가중치 텐서는 0 값 가중치를 도입하여 총 계산 횟수를 줄임(예를 들어, 0 가중치는 건너뜀)으로써 특징 추출 정확도를 희생하지 않고도 합성곱 속도를 개선할 수 있다. 일부 실시형태에서, 가중치 텐서를 가지치기하는 것은 가중치 텐서(필터라고도 함)의 채널을 복수의 채널 그룹으로 분할하고, 여기서 모든 채널 그룹은 동일한 수의 채널을 갖고; 그런 다음 각 채널 그룹의 몇 개의 채널만을 0이 아닌 입력 채널(예를 들어, 0이 아닌 가중치)로 유지하고, 이 채널 그룹 내의 모든 다른 채널(예를 들어, 모든 0 가중치)을 0으로 만드는 것을 포함할 수 있다. 가지치기 프로세스 후, 각 채널 그룹은 동일한 백분율의 0이 아닌 가중치를 포함한다. 일부 실시형태에서, 가지치기를 위한 채널 그룹의 크기(예를 들어, 각 채널 그룹 내의 채널의 수)는 가중치 텐서(필터)의 수, 즉 출력 채널의 수에 기초하여 결정될 수 있다. 일반적으로, 가중치 텐서 가지치기는, 출력 채널의 수(예를 들어, 가중치 텐서의 수)가 제1 임계값보다 크고, 0이 아닌 입력 채널의 수가 제2 임계값보다 작은, 높은 가중치 희소성과; 출력 채널의 수(예를 들어, 가중치 텐서의 수)가 제1 임계값보다 작고, 0이 아닌 입력 채널의 수가 제2 임계값보다 큰, 낮은 가중치 희소성을 포함하는 두 가지 수준으로 분류될 수 있다. 예를 들어, 높은 가중치 희소성(16X) 사례의 네이티브 텐서 모양은 m'=K=16(예를 들어, 16개의 가중치 텐서 또는 필터), k'=3*3*4(예를 들어, 각 커널은 3*3이고, 하나의 필터 내의 0이 아닌 채널의 수는 4임), 및 n'=64일 수 있는 반면; 낮은 가중치 희소성(4X) 사례의 네이티브 텐서 모양은 m'=K=4(예를 들어, 4개의 가중치 텐서 또는 필터), k'=3*3*16(예를 들어, 각 커널은 3*3이고, 하나의 필터 내의 0이 아닌 채널의 수는 16임), 및 n'=64일 수 있다.

[0046] 일부 실시형태에서, 가중치 희소성이 높음에서 낮음으로 변경될 때, 제1_외부_차원 * 내부_차원 * 제2_외부_차원으로 표시되는 네이티브 텐서 모양은, (가중치 텐서 행렬과 IFM 텐서 행렬이 공유하는) 내부 차원을 F배로 확대하고, (가중치 텐서 행렬에 대응하는) 제1 외부 차원을 F배로 축소하여 재형성될 수 있다. 도 5a에 도시된 바와 같이, 원래의 네이티브 텐서 모양 m'*k'*n'은 (m'/F) * (F*k') * n'이 되고, 여기서 가중치 텐서 행렬(510)은 m'*k'로부터 차원((m'/F) * (F*k'))을 갖는 재형성된 텐서 행렬(520)로 변경되고, IFM 텐서 행렬(512)은 k'*n'으로부터 차원((F*k')*n')을 갖는 재형성된 IFM 행렬(522)로 변경된다. 이러한 재형성 방식은 다음 설명에서 k'&m' 재형성이라고 할 수 있다. F배로 확대된 내부 차원은 (C로부터 F*C로) 더 많은 입력 채널의 지원을 나타내고, 가중치 텐서 행렬의 축소된 외부 차원은 (K로부터 K/F로) 더 적은 출력 채널을 나타낸다.

[0047] 위에서 설명한 k'&m' 재형성을 사용하여 텐서를 변환한 후, 변환된 텐서 행렬(520 및 522)은 병렬 처리를 위해 PE 어레이에 분배될 수 있다. 도 5b는 다양한 실시형태에 따라 k'&m' 재형성 기반 적응형 텐서 모양을 사용하는

신경망 계산을 위한 클러스터 간 덧셈기를 갖는 또 다른 예시적인 PE 어레이를 도시한다. 도 5b에 도시된 PE 어레이를 사용하는 병렬 처리 방식은 도 5a에 설명된 네이티브 텐서의 k'&m' 재형성에 대응할 수 있다.

[0048] 일관성을 위해 PE 어레이는 Y2개의 행과 X개의 열의 PE를 갖고, 각 PE는 Y1개의 곱셈기를 갖는다고 여전히 가정한다. 또한, 가중치와 IFM 행렬(520 및 522)은 PE 어레이의 PE의 열의 수(X)에 대응하는 동일한 내부 차원을 갖고, 가중치 행렬(520)의 외부 차원은 PE 어레이의 각 PE 내의 곱셈기의 수(Y1)에 대응하고, IFM 행렬(522)의 외부 차원은 PE 어레이의 PE의 행의 수(Y2)에 대응한다.

[0049] 재형성된 네이티브 텐서는 (가중치 행렬(520)의 외부 차원에 대응하는) 제1 외부 차원을 m'/F로 가지므로, 가중치 텐서 행렬의 각 열의 가중치는 각 PE 내의 Y1/F개의 곱셈기에 공급될 수 있다. PE 어레이로부터 부분 합을 얻기 위해, 클러스터 내 덧셈기(500)는 F개의 라운드 동안 Y1개의 덧셈기 트리의 출력을 저장하고 집계하도록 구현될 수 있다. 여기서, "라운드"는 PE 내 곱셈기를 사용하여 곱셈을 수행하는 사이클을 말한다. 각 라운드 동안, Y1/F개의 곱셈기의 출력은 하나의 클러스터 내 덧셈기(500)에 일시적으로 저장될 수 있다. F개의 라운드 후, 클러스터 내 덧셈기(500)는 Y1개의 덧셈기 트리로부터 수집된 F*Y1/F=Y1개의 부분 합을 가질 수 있다. 그런 다음 이러한 부분 합을 집계하면 합성곱의 결과로 출력 텐서를 구성할 수 있다. 이 프로세스 동안, 부분 합의 총 수는 *Y1/F)*Y2이며, 이는 출력 채널 수(예를 들어, 합성곱 프로세스의 출력 텐서의 채널의 수)가 Y1/F이고 출력 픽셀 수가 Y2=H*W임을 의미한다.

[0050] 합성곱 신경망 분야에서, 가중치 텐서는 복수의 2D 커널을 포함하는 3D 필터로 지칭될 수 있다. 각 3D 필터 내의 2D 커널의 수는 필터 내 채널의 수로 지칭될 수 있으며, 각 2D 커널은 1×1 또는 3×3 행렬일 수 있다. 도 6a는 다양한 실시형태에 따라 (1×1 커널을 사용하는) 1×1 텐서 연산 모드를 통한 예시적인 신경망 계산을 도시하고, 도 6b는 다양한 실시형태에 따라 (3×3 커널을 사용하는) 3×3 텐서 연산 모드를 통한 예시적인 신경망 계산을 도시한다.

[0051] 일부 실시형태에서, 일반 행렬 곱셈(GEMM) 및 1×1 합성곱 연산은 (예를 들어, 1×1 커널을 사용하는) 1×1 연산 모드에 매핑될 수 있다. 도 6a에 도시된 바와 같이, 서로 다른 입력 채널(또는 희소화된 입력 텐서의 채널 그룹)의 2D 커널(즉, 가중치)은 복수의 입력 채널이 한 번에 동시에 처리되도록 PE의 서로 다른 열에 배치될 수 있고, 동일한 입력 채널의 2D 커널은 곱셈기가 다수의 출력 채널을 한 번에 동시에 처리할 수 있도록 하나의 PE 내의 곱셈기 중에 분산될 수 있다. 예를 들어, 채널 1(C=1)의 가중치의 수(Y1)와 필터의 1 내지 Y1개의 커널(즉, 다수의 필터의 동일한 입력 채널의 가중치)이 제1 PE에 공급될 수 있고, 채널 2(C=2)의 가중치의 수(Y1)와 필터의 1 내지 Y1개의 커널이 제2 PE에 공급될 수 있다. 이러한 방식으로, 서로 다른 입력 채널의 2D 커널이 PE의 열에 분산된다.

[0052] 희소화된 입력 텐서를 포함하는 일부 실시형태에서, 각 가중치는 인덱스-값 쌍으로 표현될 수 있다. 인덱스-값 쌍의 값은 0이 아닌 가중치의 값이고, 인덱스-값 쌍의 인덱스는 0이 아닌 가중치의 인덱스이며, 이는 하나의 곱셈기로 곱셈을 수행하기 위해 대응하는 입력 값을 식별하는 데 사용될 수 있다. 일부 실시형태에서, 채널의 수가 각 PE 클러스터(각 행) 내의 PE의 수보다 적으면, 나머지 PE는 다른 벡터 연산에 사용될 수 있다.

[0053] 일부 실시형태에서, 위에서 설명된 1×1 합성곱 연산 이외의 합성곱은 하나 이상의 3×3 합성곱으로 분해되어 (예를 들어, 3×3 커널을 사용하는) 3×3 네이티브 연산 모드에 매핑될 수 있다. 도 6b에 도시된 바와 같이, 각각의 2D 3*3 커널은 동일한 입력 채널에서 나온 (0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2)로 표시된 9개의 가중치를 가지며, 이는 동시 처리를 위해 PE의 동일한 행(다른 열)에 분배될 수 있다. 다른 입력 채널에서 나온 9개의 가중치는 PE의 다른 행에 분배될 수 있다.

[0054] 도 7은 다양한 실시형태에 따라 PE 어레이의 내부 버퍼의 예시적 아키텍처 다이어그램을 도시한다. 일부 실시형태에서, PE 어레이의 각 PE는 입력 값을 저장하기 위해 입력 버퍼(IBUF로 표시됨)(722)와 결합된다. 이러한 입력 값은 대응하는 입력 값을 찾기 위해 주어진 가중치 인덱스에 기초하여 PE에 의해 검색될 수 있다. 그런 다음 검색된 입력 값은 PE의 곱셈기 내의 가중치 값과 곱해질 수 있다. 실제 구현에서, IBUF의 깊이는 일반적으로 제한되어 있으며, 이는 하나의 IBUF(722)가 고정된 수의 입력 채널의 입력 값만을 저장할 수 있음을 의미한다. 이러한 설계는 0이 아닌 입력 채널의 수도 제한되어 있기 때문에 희소화된 입력 텐서에 적합하다. 그러나, 종종 0이 아닌 입력 채널의 수가 IBUF(722)의 깊이를 넘어갈 수 있는 경우가 있다. 이러한 경우에, IBUF(722)는 필요한 입력 값을 외부 메모리로부터 판독하기 위해 캐시 교체를 수행해야 할 수 있으며, 이는 비용이 많이 들고 비효율적이다.

[0055] 일부 실시형태에서, 가중치 텐서(필터)의 희소화 정도에 따라, 각 PE의 IBUF는 개인 메모리 또는 공유 메모리로

구성될 수 있다. 예를 들어, 하나 이상의 가중치 텐서를 희소화하는 것은 하나 이상의 가중치 텐서의 입력 채널을 복수의 채널 그룹으로 분할하고(각 채널 그룹은 1보다 큰 정수인 고정된 수의 채널을 포함함); 및 복수의 채널 그룹 각각에서 몇 개의 채널만이 0이 아닌 입력 값을 포함하고 각 채널 그룹의 다른 채널은 모두 0을 포함하도록 하나 이상의 가중치 텐서 각각을 가지치기하는 것을 포함할 수 있다. 가지치기 프로세스 후, 각 채널 그룹은 동일한 백분율의 0이 아닌 가중치를 포함한다. 희소화의 세분성은 미세 세분화(710)와 거친 세분화(750)로 분류될 수 있다. 미세 세분화 희소화(710)는 고정된 수보다 작은 수의 채널로부터 0이 아닌 입력 채널이 선택될 때 발생하고, 거친 세분화 희소화(750)는 고정된 수보다 큰 수의 채널로부터 0이 아닌 입력 채널이 선택될 때 발생한다. 예를 들어, 가중치 희소성이 15/16(16개 채널 중 하나의 0이 아닌 입력 채널)인 경우, 매 16개 입력 채널마다(예를 들어, 하나의 채널 그룹은 16개 입력 채널을 포함함) 하나의 0이 아닌 입력 채널을 선택하는 것은 미세 세분화 희소화(710)로 결정될 수 있는 반면, 매 64개 입력 채널마다(예를 들어, 하나의 채널 그룹은 64개 입력 채널을 포함함) 4개의 0이 아닌 입력 채널을 선택하는 것은 거친 세분화 희소화(750)로 결정될 수 있다.

[0056] 일부 실시형태에서, IBUF(722)는 미세 세분성을 갖는 희소 가중치 텐서에 대한 개인 메모리로 구성되거나 또는 거친 세분성을 갖는 희소 텐서에 대한 공유 메모리로 구성될 수 있다. 이것은 IBUF(722)의 깊이가 미세 세분성 희소화와 거친 세분성 희소화를 분류하는 데 사용되는 고정된 숫자와 비교될 수 있다. IBUF(722)의 깊이가 고정된 숫자보다 큰 경우, 이것은 IBUF(722)가 필요한 입력 값을 저장하기에 충분하다는 것을 의미한다. 이러한 방식으로, 전용 개인 메모리로 인해 데이터 검색 성능이 최적이 된다. IBUF(722)의 깊이가 고정된 숫자보다 작은 경우, 다수의 이웃하는 PE가 IBUF를 공유하여 (공유 IBUF(780)로 표시됨) 이들이 검색할 수 있는 입력 값을 저장할 수 있다. 이러한 방식으로, 중복된 입력 값이 줄어들고 전반적인 저장 효율성이 향상된다.

[0057] 도 8은 다양한 실시형태에 따라 적응형 텐서 모양을 사용하는 신경망 계산을 위한 예시적인 방법(800)을 설명한다. 방법(800)은 도 1 내지 도 7에 설명된 디바이스, 장치 또는 시스템에 의해 수행될 수 있다. 아래에 제시된 방법(800)의 동작은 설명을 위해 의도된 것이다. 구현에 따라, 방법(800)은 다양한 순서로 또는 병렬로 수행되는 추가적인 단계, 더 적은 단계, 또는 대안적인 단계를 포함할 수 있다.

[0058] 블록(810)은 처리 요소(PE) 어레이를 사용하여 합성곱을 수행하기 위한 합성곱 신경망(CNN)의 제1 계층에서 제1 입력 특징 맵(IFM)과 하나 이상의 제1 필터를 수신하는 것을 포함하고, 여기서 PE 어레이의 각 PE는 소정 개수(Y1)의 곱셈기를 포함하고, PE 어레이는 소정 개수(Y2)의 행과 소정 개수(X)의 열로 배열된다. 일부 실시형태에서, PE의 각 행은 각 PE 내의 곱셈기의 수(Y1)에 대응하는 수(Y1)의 덧셈기 트리와 결합되며, 여기서 각 PE 내의 각 곱셈기는 집계를 위해 대응하는 덧셈기 트리에 곱셈 출력을 보낸다. 각 PE 내 소정 개수(Y1)의 곱셈기는 병렬로 데이터를 처리하고, PE 어레이의 PE는 병렬로 데이터를 처리한다.

[0059] 블록(820)은 제1 IFM과 하나 이상의 제1 필터에 기초하여 네이티브 텐서 모양을 결정하는 것을 포함하고, 여기서 네이티브 텐서 모양은 제1 외부 차원, 내부 차원 및 제2 외부 차원을 포함하고, 여기서 네이티브 텐서 모양은 제1 IFM과 하나 이상의 제1 필터를 PE 어레이에 매핑한다.

[0060] 블록(830)은 PE 어레이를 사용하여 합성곱을 수행하기 위한 CNN의 제2 계층에서 제2 IFM과 하나 이상의 제2 필터를 수신하는 것을 포함한다. 일부 실시형태에서, CNN의 제2 계층은 CNN의 제1 계층 뒤에 있고, 제2 IFM은 제1 IFM보다 많은 입력 채널과, 제1 IFM보다 낮은 해상도를 포함한다. 일부 실시형태에서, 하나 이상의 제2 필터 각각은 2차원(2D) 커널의 복수의 채널을 포함하고, 각 2D 커널은 1×1 또는 3×3의 차원을 갖는다.

[0061] 블록(840)은 제2 IFM과 하나 이상의 제2 필터에 기초하여 네이티브 텐서 모양을 재형성하는 것을 포함하고, 여기서 재형성은 내부 차원을 확대하는 것과, 제1 외부 차원과 제2 외부 차원 중 하나를 축소하는 것을 포함하고, 확대와 축소는 F배로 이루어진다.

[0062] 블록(850)은 재형성된 네이티브 텐서에 따라 합성곱을 위해 하나 이상의 제2 필터와 제2 IFM을 PE 어레이에 공급하는 것을 포함하고, 여기서 제1 외부 차원이 축소되는 것에 응답하여 합성곱은 F개의 라운드 동안 PE의 동일한 행의 출력을 집계하여 부분 합을 얻는 것을 포함하고, 제2 외부 차원이 축소되는 것에 응답하여 합성곱은 PE의 매 F개의 행마다 출력을 집계하여 부분 합을 얻는 것을 포함한다. 일부 실시형태에서, 재형성된 네이티브 텐서에 따라 하나 이상의 제2 필터를 PE 어레이에 공급하는 것은 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 하나 이상의 제2 필터를 행렬로 변환하는 것(여기서 하나 이상의 제2 필터의 각 2D 커널이 1×1 차원을 갖는 것에 응답하여, 행렬의 각 행은 하나 이상의 제2 필터의 서로 다른 입력 채널의 가중치를 포함함); 및 복수의 입력 채널이 한 번에 동시에 처리되도록 행렬의 각 행의 가중치를 PE의 서로 다른 열에 분배하는 것을 포함한다. 일부 실시형태에서, 재형성된 네이티브 텐서에 따라 하나 이상의 제2 필터를 PE 어레이에 공급하

는 것은 재형성된 네이티브 텐서의 제1 외부 차원과 내부 차원에 따라 하나 이상의 제2 필터를 행렬로 변환하는 것(여기서 하나 이상의 제2 필터의 각 2D 커널이 3×3의 차원을 갖고 9개의 가중치를 포함하는 것에 응답하여, 9개의 가중치가 행렬의 동일한 행에 배치됨); 및 동일한 채널의 가중치가 한 번에 동시에 처리되도록 행렬의 동일한 행의 9개의 가중치를 PE의 서로 다른 열에 분배하는 것을 포함한다. 일부 실시형태에서, 재형성된 네이티브 텐서에 따라 IFM을 PE 어레이에 공급하는 것은 재형성된 네이티브 텐서의 내부 차원과 제2 외부 차원에 따라 IFM을 행렬로 변환하는 것; 및 행렬의 열에 대응하는 IFM의 입력 값을 PE의 행의 버퍼에 공급하는 것을 포함한다.

[0063] 블록(860)은 복수의 부분 합을 집계하여 CNN의 제2 계층에서 합성곱의 출력 텐서를 얻는 것을 포함한다.

[0064] 위의 설명에서, Y1, Y2, X 및 F는 모두 1보다 큰 정수이다.

[0065] 일부 실시형태에서, 방법(800)은 하나 이상의 필터의 채널을 복수의 채널 그룹으로 분할하는 단계(여기서 각 채널 그룹은 1보다 큰 정수인 고정된 수의 채널을 포함함); 및 복수의 채널 그룹 각각에서 단 하나의 채널만이 0이 아닌 입력 값을 포함하고, 각 채널 그룹의 다른 채널은 모두 0을 포함하도록 하나 이상의 필터 각각을 가지치기하는 단계를 추가로 포함할 수 있다. 일부 실시형태에서, 방법(800)은 PE 어레이의 각 PE와 연관된 버퍼의 깊이를 결정하는 단계; 버퍼의 깊이가 고정된 수보다 큰 것에 응답하여, 버퍼를 각 PE에 대한 개인 메모리로 구성하는 단계; 및 버퍼의 깊이가 고정된 수보다 작은 것에 응답하여, PE의 버퍼와 이웃 PE의 하나 이상의 버퍼를 공유 메모리로 결합하는 단계를 추가로 포함할 수 있다. 일부 실시형태에서, 각 PE의 개인 메모리는 PE 내 소정 개수(Y1)의 곱셈기에 의해 검색 가능한 입력 값을 저장하고, 공유 메모리는 PE 내 소정 개수(Y1)의 곱셈기와 하나 이상의 이웃 PE에 의해 검색 가능한 입력 값을 저장한다.

[0066] 일부 실시형태에서, 하나 이상의 제2 필터 각각은 복수의 0이 아닌 가중치를 포함하고, 합성곱을 위해 하나 이상의 제2 필터를 PE 어레이에 공급하는 것은, 각 0이 아닌 가중치를 대응하는 PE의 곱셈기에 0이 아닌 가중치와 대응 인덱스를 포함하는 인덱스-값 쌍으로 공급하는 것을 포함하고; 합성곱은 인덱스에 따라 대응하는 PE의 버퍼로부터 입력 값을 검색하는 단계; 및 검색된 값과 0이 아닌 가중치를 곱셈기에 보내 출력을 얻는 단계; 및 대응하는 PE와 같은 행에 있는 다른 PE의 다른 곱셈기에서 생성된 출력과 함께 집계를 위해 대응하는 덧셈기 트리에 출력을 보내는 단계를 포함한다.

[0067] 도 9는 본 명세서에 설명된 실시형태 중 임의의 실시형태를 구현할 수 있는 예시적인 컴퓨팅 디바이스를 도시한다. 컴퓨팅 디바이스는 도 1 내지 도 8에 도시된 시스템 및 방법의 하나 이상의 구성 요소를 구현하는 데 사용될 수 있다. 컴퓨팅 디바이스(900)는 정보를 통신하기 위한 버스(902) 또는 기타 통신 메커니즘, 및 이 버스(902)와 결합되어 정보를 처리하기 위한 하나 이상의 하드웨어 프로세서(904)를 포함할 수 있다. 하드웨어 프로세서(들)(904)는 예를 들어 하나 이상의 범용 마이크로프로세서일 수 있다.

[0068] 컴퓨팅 디바이스(900)는 또한 버스(902)에 결합되어 프로세서(들)(904)가 실행할 명령어와 정보를 저장하기 위한 랜덤 액세스 메모리(RAM), 캐시 및/또는 기타 동적 저장 디바이스와 같은 주 메모리(907)를 포함할 수 있다. 주 메모리(907)는 프로세서(들)(904)가 실행할 명령어를 실행하는 동안 임시 변수나 다른 중간 정보를 저장하는 데에도 사용될 수 있다. 이러한 명령어는 프로세서(들)(904)가 액세스할 수 있는 저장 매체에 저장될 때 컴퓨팅 디바이스(900)를 명령어에 지정된 동작을 수행하도록 사용자 정의된 특수 목적의 기계로 만들 수 있다. 주 메모리(907)는 비휘발성 매체 및/또는 휘발성 매체를 포함할 수 있다. 비휘발성 매체는 예를 들어 광학 또는 자기 디스크를 포함할 수 있다. 휘발성 매체는 동적 메모리를 포함할 수 있다. 일반적인 매체 형태는 예를 들어 플로피 디스크, 플렉시블 디스크, 하드 디스크, 솔리드 스테이트 드라이브, 자기 테이프 또는 임의의 다른 자기 데이터 저장 매체, CD-ROM, 임의의 다른 광학 데이터 저장 매체, 구멍 패턴이 있는 임의의 물리적 매체, RAM, DRAM, PROM 및 EPROM, FLASH-EPROM, NVRAM, 임의의 다른 메모리 칩 또는 카트리지, 또는 이들이 네트워크로 연결된 형태를 포함할 수 있다.

[0069] 컴퓨팅 디바이스(900)는 컴퓨팅 디바이스와 결합하여 컴퓨팅 디바이스(900)를 특수 목적 기계로 만들거나 프로그래밍할 수 있는 사용자 정의 하드와이어 논리, 하나 이상의 ASIC 또는 FPGA, 펌웨어 및/또는 프로그램 논리를 사용하여 본 명세서에 설명된 기술을 구현할 수 있다. 일 실시형태에 따르면, 본 명세서에 설명된 기술은 프로세서(들)(904)가 주 메모리(907)에 포함된 하나 이상의 명령어의 하나 이상의 시퀀스를 실행하는 것에 응답하여 컴퓨팅 디바이스(900)에 의해 수행된다. 이러한 명령어는 저장 디바이스(909)와 같은 다른 저장 매체로부터 주 메모리(907)로 판독될 수 있다. 주 메모리(907)에 포함된 명령어 시퀀스를 실행하면 프로세서(들)(904)가 본 명세서에 설명된 프로세스 단계를 수행할 수 있다. 예를 들어, 본 명세서에 개시된 프로세스/방법은 주 메모리(907)에 저장된 컴퓨터 프로그램 명령어에 의해 구현될 수 있다. 이러한 명령어가 프로세서(들)(904)에 의해 실

행되면, 프로세서는 대응하는 도면에 도시되고 위에서 설명된 단계를 수행할 수 있다. 대안적인 실시형태에서, 하드와이어로 연결된 회로부는 소프트웨어 명령어 대신 또는 소프트웨어 명령어와 함께 사용될 수 있다.

[0070] 컴퓨팅 디바이스(900)는 또한 버스(902)에 결합된 통신 인터페이스(910)를 포함한다. 통신 인터페이스(910)는 하나 이상의 네트워크에 연결된 하나 이상의 네트워크 링크에 양방향 데이터 통신 결합을 제공할 수 있다. 또 다른 예로서, 통신 인터페이스(910)는 호환 LAN(또는 WAN과 통신하는 WAN 구성 요소)에 데이터 통신 연결을 제공하는 근거리 네트워크(LAN) 카드일 수 있다. 무선 링크도 구현될 수 있다.

[0071] 특정 동작의 수행은 단일 기계 내에 상주할 뿐만 아니라 여러 기계에 걸쳐 전개되는 프로세서 중에 분산될 수 있다. 일부 예시적인 실시형태에서, 프로세서 또는 프로세서 구현 엔진은 단일 지리적 위치(예를 들어, 가정 환경, 사무실 환경 또는 서버 팜)에 위치될 수 있다. 다른 예시적 실시형태에서, 프로세서 또는 프로세서 구현 엔진은 여러 지리적 위치에 분산될 수 있다.

[0072] 이전 부분에 설명된 각각의 프로세스, 방법 및 알고리즘은 컴퓨터 하드웨어를 포함하는 하나 이상의 컴퓨터 시스템 또는 컴퓨터 프로세서에 의해 실행되는 코드 모듈에 구현되고, 코드 모듈에 의해 완전히 또는 부분적으로 자동화될 수 있다. 프로세스와 알고리즘은 부분적으로 또는 전체적으로 응용별 회로에 구현될 수 있다.

[0073] 본 명세서에 개시된 기능이 소프트웨어 기능 단위의 형태로 구현되고 독립적인 제품으로 판매되거나 사용되는 경우, 이 기능은 프로세서 실행 가능 비휘발성 컴퓨터 판독 가능 저장 매체에 저장될 수 있다. 본 명세서에 개시된 특정 기술 솔루션(전체 또는 일부) 또는 현재 기술에 기여하는 양태는 소프트웨어 제품의 형태로 구현될 수 있다. 컴퓨팅 디바이스(개인용 컴퓨터, 서버, 네트워크 디바이스 등일 수 있음)가 본 출원의 실시형태의 방법의 모든 단계 또는 일부 단계를 실행하게 하는 다수의 명령어를 포함하는 소프트웨어 제품이 저장 매체에 저장될 수 있다. 저장 매체는 플래시 드라이브, 휴대용 하드 드라이브, ROM, RAM, 자기 디스크, 광 디스크, 프로그램 코드를 저장하도록 동작 가능한 다른 매체 또는 이들의 임의의 조합을 포함할 수 있다.

[0074] 특정 실시형태는 프로세서를 포함하는 시스템, 및 프로세서에 의해 실행되어 시스템으로 하여금 위에서 개시된 실시형태의 임의의 방법의 단계에 대응하는 동작을 수행하게 하는 명령어를 저장하는 비일시적 컴퓨터 판독 가능 저장 매체를 추가로 제공한다. 특정 실시형태는 하나 이상의 프로세서에 의해 실행되어 하나 이상의 프로세서가 위에서 개시된 실시형태의 임의의 방법의 단계에 대응하는 동작을 수행하게 하는 명령어를 포함하도록 구성된 비일시적 컴퓨터 판독 가능 저장 매체를 추가로 제공한다.

[0075] 본 명세서에 개시된 실시형태는 클라이언트와 상호 작용하는 클라우드 플랫폼, 서버 또는 서버 그룹(이하 통칭하여 "서비스 시스템")을 통해 구현될 수 있다. 클라이언트는 단말 디바이스이거나 플랫폼에서 사용자가 등록한 클라이언트일 수 있으며, 단말 디바이스는 플랫폼 응용 프로그램이 설치될 수 있는 모바일 단말, 개인용 컴퓨터(PC) 및 임의의 디바이스일 수 있다.

[0076] 위에서 설명된 다양한 특징 및 프로세스는 서로 독립적으로 사용되거나 다양한 방식으로 결합될 수 있다. 모든 가능한 조합 및 하위 조합은 본 발명의 범위 내에 포함되도록 의도된다. 또한, 일부 구현에서는 특정 방법 또는 프로세스 블록이 생략될 수 있다. 본 명세서에 설명된 방법 및 프로세스는 또한 임의의 특정 시퀀스로 제한되지 않으며, 블록 또는 이 블록과 관련된 상태는 적절한 다른 시퀀스로 수행될 수 있다. 예를 들어, 설명된 블록 또는 상태는 구체적으로 개시된 순서와는 다른 순서로 수행될 수 있고, 또는 다수의 블록 또는 상태는 단일 블록 또는 상태로 결합될 수 있다. 예시적인 블록 또는 상태는 직렬로, 병렬로 또는 일부 다른 방식으로 수행될 수 있다. 블록 또는 상태는 개시된 예시적인 실시형태에 추가되거나 이로부터 제거될 수 있다. 본 명세서에 설명된 예시적인 시스템 및 구성 요소는 설명된 것과 다르게 구성될 수 있다. 예를 들어, 요소는 개시된 예시적인 실시형태에 추가되거나, 이로부터 제거되거나, 이에 비해 재배열될 수 있다.

[0077] 본 명세서에 설명된 예시적인 방법의 다양한 동작은 적어도 부분적으로 알고리즘에 의해 수행될 수 있다. 알고리즘은 메모리(예를 들어, 위에서 설명된 비일시적 컴퓨터 판독 가능 저장 매체)에 저장된 프로그램 코드 또는 명령어로 구성될 수 있다. 이러한 알고리즘은 기계 러닝 알고리즘을 포함할 수 있다. 일부 실시형태에서, 기계 러닝 알고리즘은 기능을 수행하도록 컴퓨터를 명시적으로 프로그래밍하지는 않지만 훈련 샘플로부터 학습하여 기능을 수행하는 예측 모델을 만들 수 있다.

[0078] 본 명세서에 설명된 예시적인 방법의 다양한 동작은 관련 동작을 수행하도록 일시적으로 구성되거나(예를 들어, 소프트웨어) 영구적으로 구성된 하나 이상의 프로세서에 의해 적어도 부분적으로 수행될 수 있다. 이러한 프로세서는 일시적으로 구성되든 영구적으로 구성되든, 본 명세서에 설명된 하나 이상의 동작 또는 기능을 수행하도록 동작하는 프로세서 구현 엔진을 구성할 수 있다.

[0079] 유사하게, 본 명세서에 설명된 방법은 적어도 부분적으로 프로세서로 구현될 수 있으며, 여기서 특정 프로세서 또는 프로세서들은 하드웨어의 일례이다. 예를 들어, 방법의 적어도 일부 동작은 하나 이상의 프로세서 또는 프로세서 구현 엔진에 의해 수행될 수 있다. 또한, 하나 이상의 프로세서는 또한 "클라우드 컴퓨팅" 환경 또는 "서비스로서의 소프트웨어"(SaaS)로서 관련 동작의 성능을 지원하도록 동작할 수 있다. 예를 들어, 적어도 일부 동작은 컴퓨터 그룹(프로세서를 포함하는 기계의 예)에 의해 수행될 수 있으며, 이러한 동작은 네트워크(예를 들어, 인터넷) 및 하나 이상의 적절한 인터페이스(예를 들어, 응용 프로그램 인터페이스(API))를 통해 액세스 가능할 수 있다.

[0080] 특정 동작의 수행은, 단일 기계 내에 상주하는 것뿐만 아니라 여러 기계에 걸쳐 전개되는 프로세서 간에 분산될 수 있다. 일부 예시적인 실시형태에서, 프로세서 또는 프로세서 구현 엔진은 단일 지리적 위치(예를 들어, 가정 환경, 사무실 환경 또는 서버 팜)에 위치될 수 있다. 다른 예시적인 실시형태에서, 프로세서 또는 프로세서 구현 엔진은 여러 지리적 위치에 걸쳐 분산될 수 있다.

[0081] 본 명세서 전반에 걸쳐, 복수의 인스턴스는 단일 인스턴스로 설명된 구성 요소, 동작 또는 구조를 구현할 수 있다. 하나 이상의 방법의 개별 동작이 별도의 동작으로 예시되고 설명되지만, 하나 이상의 개별 동작이 동시에 수행될 수 있으며, 동작이 설명된 순서대로 수행되어야 한다는 요구 사항은 없다. 예시적인 구성에서 별도의 구성 요소로 제시된 구조 및 기능은 결합된 구조 또는 구성 요소로 구현될 수 있다. 유사하게, 단일 구성 요소로 제시된 구조 및 기능은 별도의 구성 요소로 구현될 수 있다. 이러한 및 기타 변형, 수정, 추가 및 개선은 본 발명의 범위에 속한다.

[0082] 본 명세서에서 사용되는 "또는"은 명시적으로 달리 지시되거나 문맥상 달리 지시되지 않는 한, 포괄적이며 배타적이지 않다. 따라서, 본 명세서에서 "A, B 또는 C"는 명시적으로 달리 지시되거나 문맥상 달리 지시되지 않는 한, "A, B, A 및 B, A 및 C, B 및 C 또는 A, B 및 C"를 의미한다. 또한, "및"은 문맥상 달리 명시적으로 지시되거나 달리 지시되지 않는 한, 함께 및 개별적인 것을 모두 의미한다. 따라서, 본 명세서에서 "A 및 B"는 문맥상 달리 명시적으로 지시되거나 달리 지시되지 않는 한, "함께 또는 개별적으로 A 및 B"를 의미한다. 또한, 본 명세서에서 단일 인스턴스로 설명된 자원, 동작 또는 구조에 대해 복수의 인스턴스가 제공될 수 있다. 추가로, 다양한 자원, 동작, 엔진 및 데이터 저장소 간의 경계는 다소 임의적이며, 특정 동작은 특정 예시적인 구성의 맥락에서 설명된다. 다른 기능 할당이 구상되고 본 발명의 다양한 실시형태의 범위 내에 포함될 수 있다. 일반적으로, 예시적인 구성에서 별도의 자원으로 제시된 구조 및 기능은 결합된 구조 또는 자원으로 구현될 수 있다. 유사하게, 단일 자원으로 제시된 구조 및 기능은 별도의 자원으로 구현될 수 있다. 이러한 및 다른 변형, 수정, 추가 및 개선은 첨부된 청구범위에 표현된 본 발명의 실시형태의 범위 내에 포함된다. 따라서, 본 명세서와 도면은 본 발명을 제한하는 의미가 아니라 본 발명을 설명하는 의미로 간주되어야 한다.

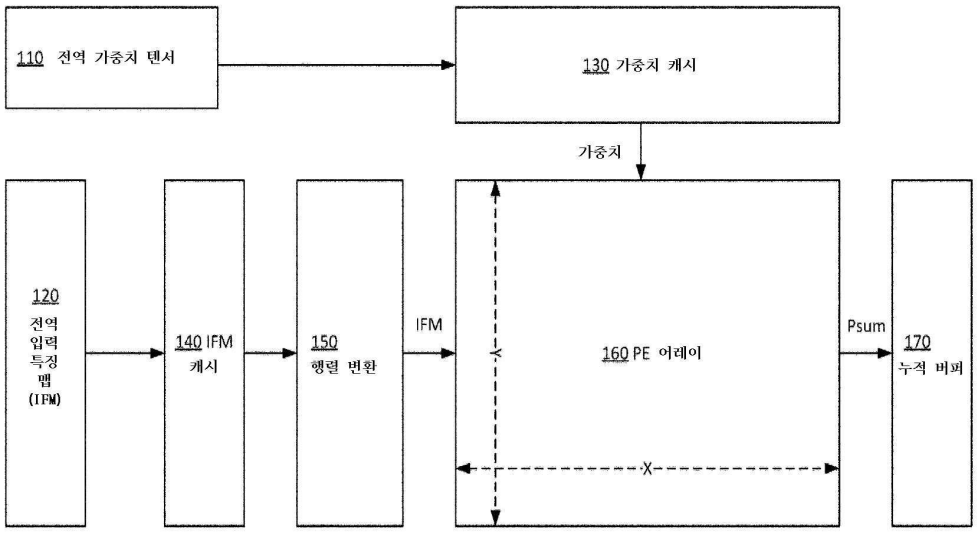
[0083] "구비하는" 또는 "포함하는"이라는 용어는 이후에 선언된 특징의 존재를 나타내는 데 사용되지만, 다른 특징의 추가를 배제하지는 않는다. 특히 "할 수 있는"과 같은 조건부 언어는 특별히 달리 명시되지 않거나 사용된 문맥 내에서 달리 이해되지 않는 한, 일반적으로 특정 기능, 요소 및/또는 단계를 특정 실시형태는 포함하고 다른 실시형태는 포함하지 않는다는 것을 전달하기 위해 의도된 것이다. 따라서, 이러한 조건부 언어는 일반적으로 하나 이상의 실시형태에 특징, 요소 및/또는 단계가 임의의 방식으로 필요하다는 것을 의미하도록 의도된 것도 아니고, 하나 이상의 실시형태에 사용자 입력이나 프롬프트 여부에 관계없이 이러한 특징, 요소 및/또는 단계가 임의의 특정 실시형태에 포함되거나 임의의 특정 실시형태에서 수행되어야 하는지 여부를 결정하기 위한 논리가 반드시 포함된다는 것을 의미하도록 의도된 것이 아니다.

[0084] 특정 예시적인 실시형태를 참조하여 주제에 대한 개요를 설명했지만, 본 발명의 실시형태의 보다 광범위한 범위를 벗어나지 않고 이러한 실시형태에 다양한 수정 및 변경이 이루어질 수 있다. 본 주제의 이러한 실시형태는 실제로 두 개 이상의 개시 또는 개념이 개시된 경우 본 출원의 범위를 임의의 단일 개시 또는 개념으로 자발적으로 제한하려는 의도 없이 단지 편의상 본 명세서에서는 개별적으로 또는 집합적으로 "발명"이라는 용어로 지칭될 수 있다.

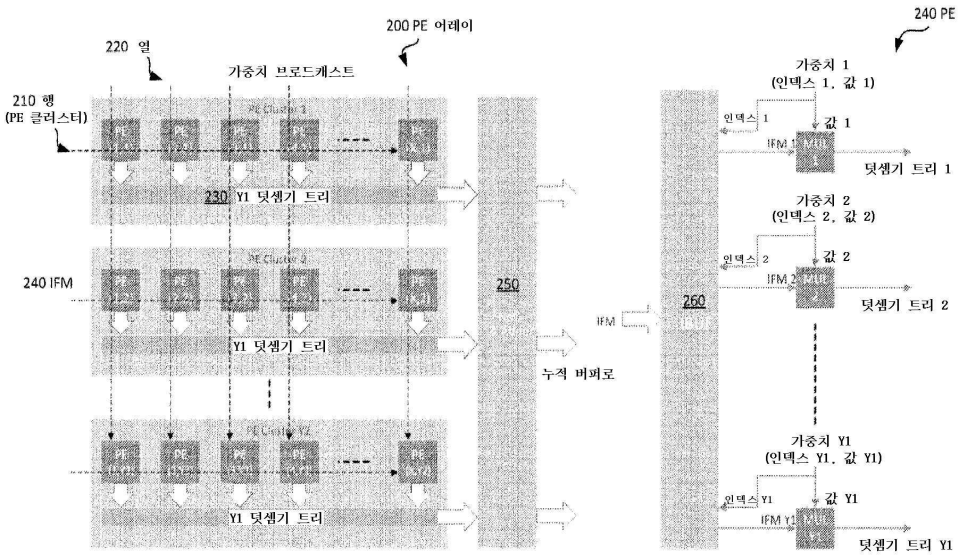
[0085] 본 명세서에 예시된 실시형태는 당업자라면 개시된 내용을 실시할 수 있을 만큼 충분히 상세히 설명되어 있다. 본 발명의 범위를 벗어나지 않고 구조적 및 논리적 대체 및 변경이 이루어질 수 있도록 본 실시형태로부터 다른 실시형태도 사용되고 유도될 수 있다. 따라서, 본 상세한 설명은 본 발명을 제한하는 의미로 받아들여져서는 안 되고, 다양한 실시형태의 범위는 본 청구범위에 주어지는 균등범위와 함께 첨부된 청구범위에 의해서만 한정된다.

도면

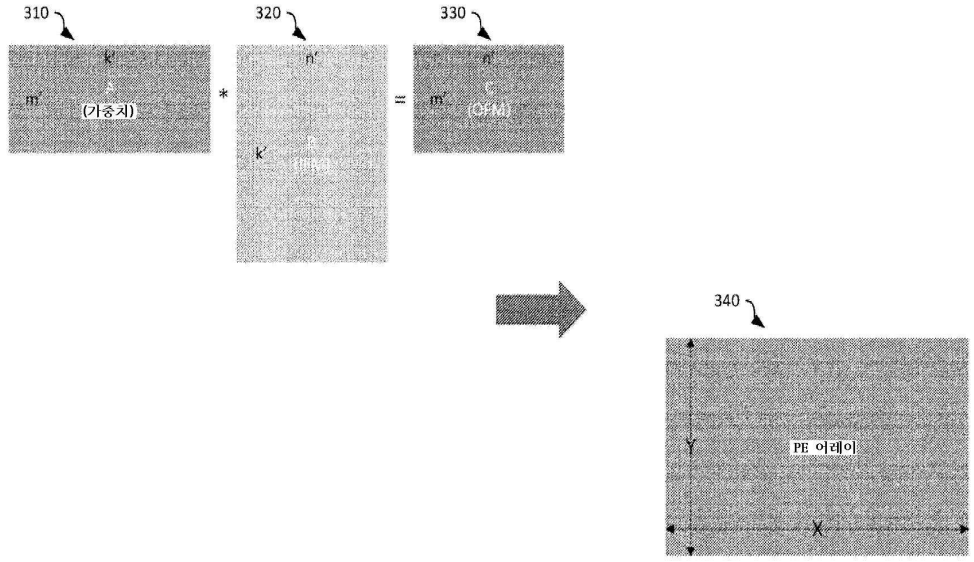
도면1


110 전역 가중치 센서
130 가중치 캐시
가중치
120
전역
입력
특징
맵
(IFM)
140 IFM
캐시
150
행렬 변환
IFM
160 PE 어레이
Psum
170
누적 버퍼
X
Y

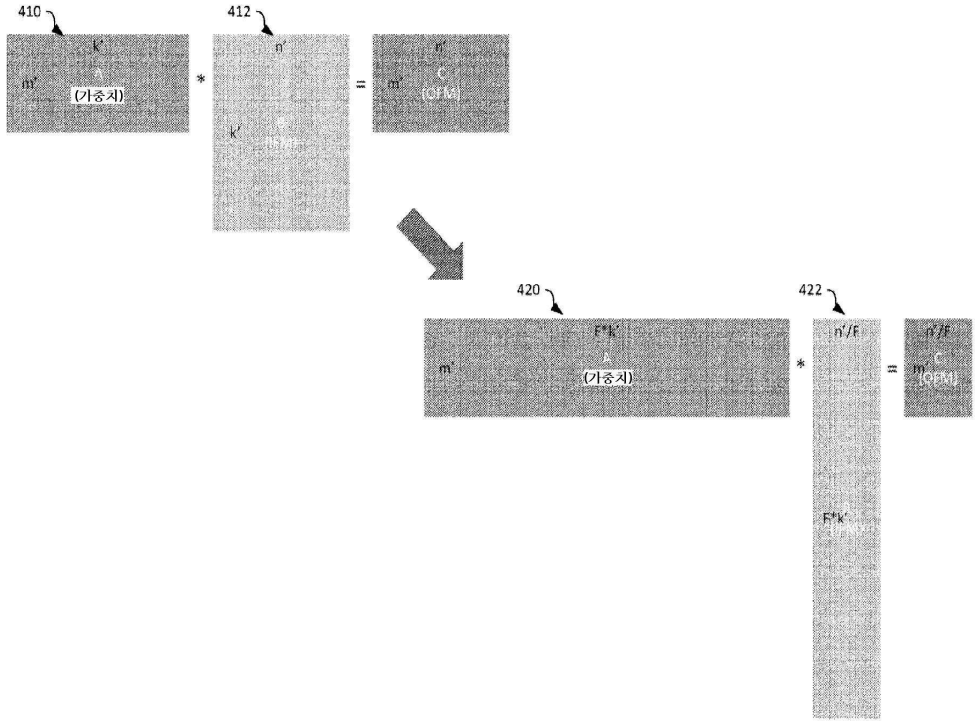
도면2


200 PE 어레이
240 PE
220 열
가중치 브로드캐스트
210 행
(PE 클러스터)
230 Y1 덧셈기 트리
240 IFM
Y1 덧셈기 트리
Y1 덧셈기 트리
250
누적 버퍼로
IFM
260
가중치 1
(인덱스 1, 값 1)
인덱스 1
값 1
IFM 1
덧셈기 트리 1
가중치 2
(인덱스 2, 값 2)
인덱스 2
값 2
IFM 2
덧셈기 트리 2
가중치 Y1
(인덱스 Y1, 값 Y1)
인덱스 Y1
값 Y1
IFM Y1
덧셈기 트리 Y1

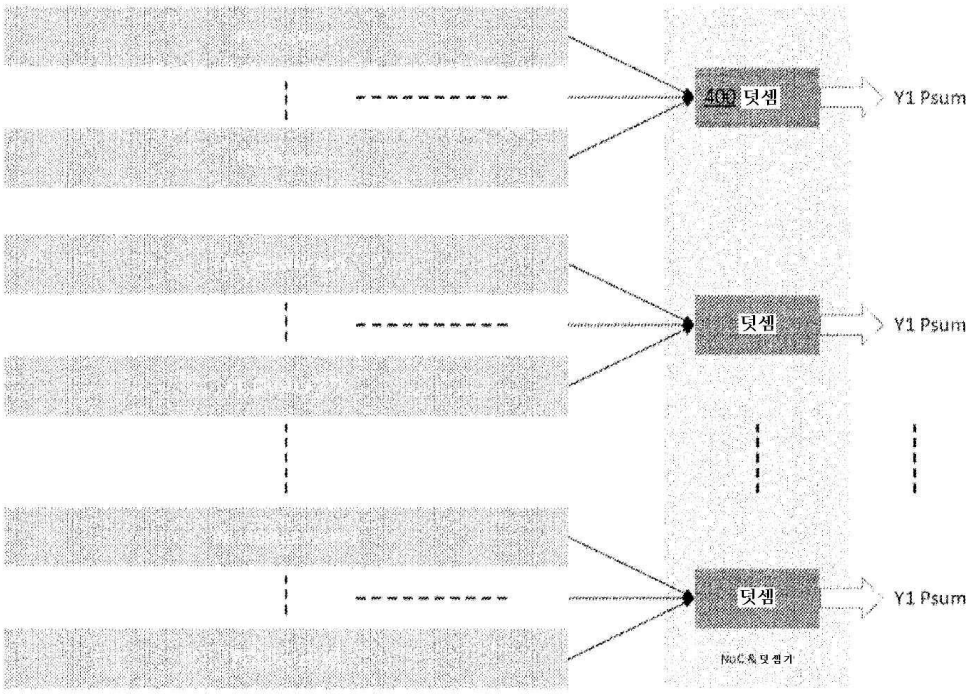
도면3



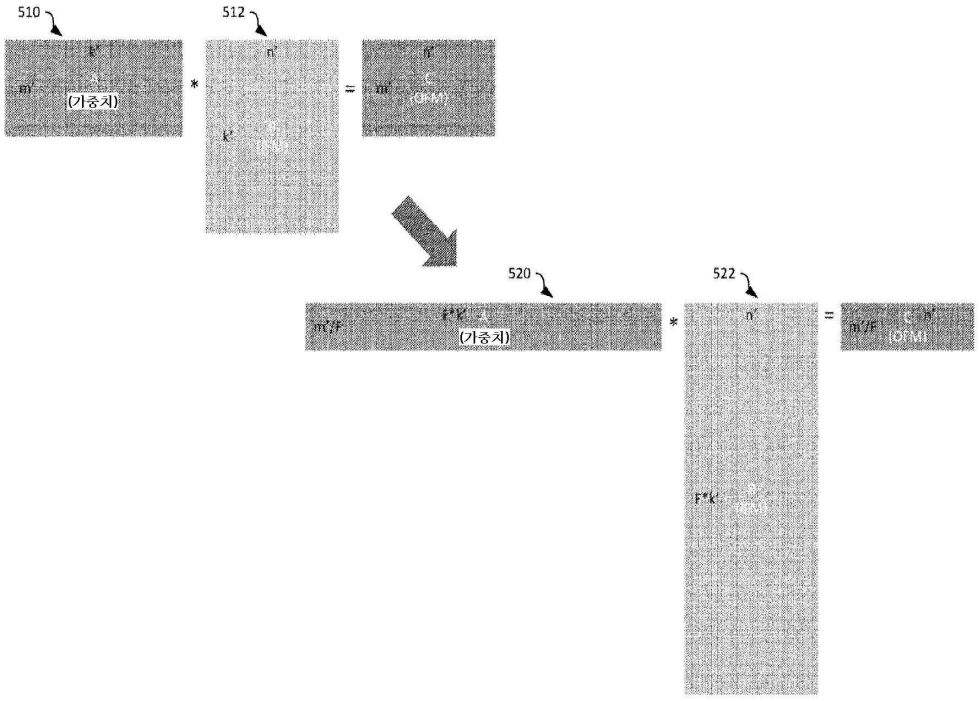
도면4a



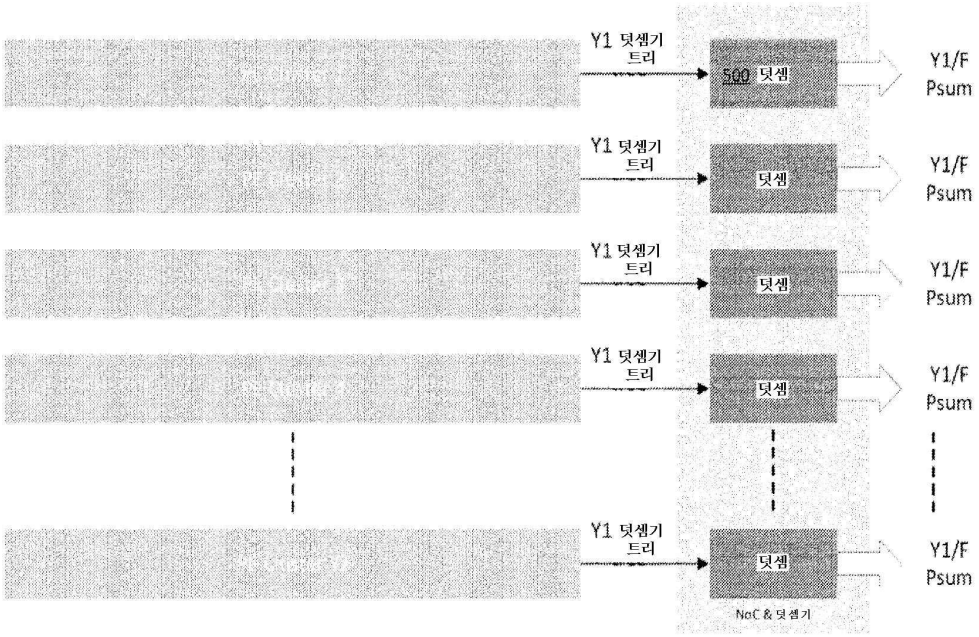
도면4b



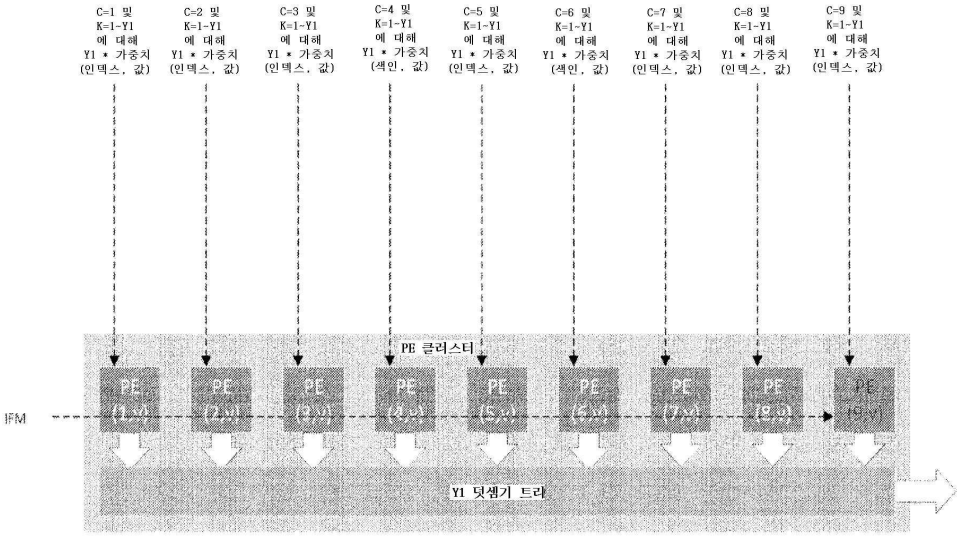
도면5a



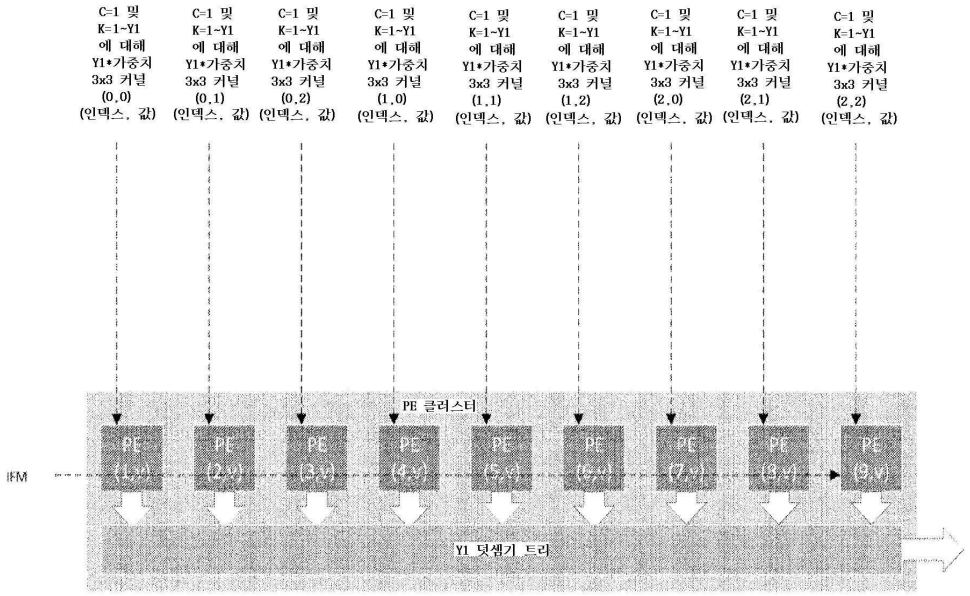
도면5b


Y1 덧셈기 트리
500 덧셈
Y1/F Psum
Y1 덧셈기 트리
덧셈
Y1/F Psum
Y1 덧셈기 트리
덧셈
Y1/F Psum
Y1 덧셈기 트리
덧셈
Y1/F Psum
Y1 덧셈기 트리
덧셈
Y1/F Psum
NoC & 덧셈기

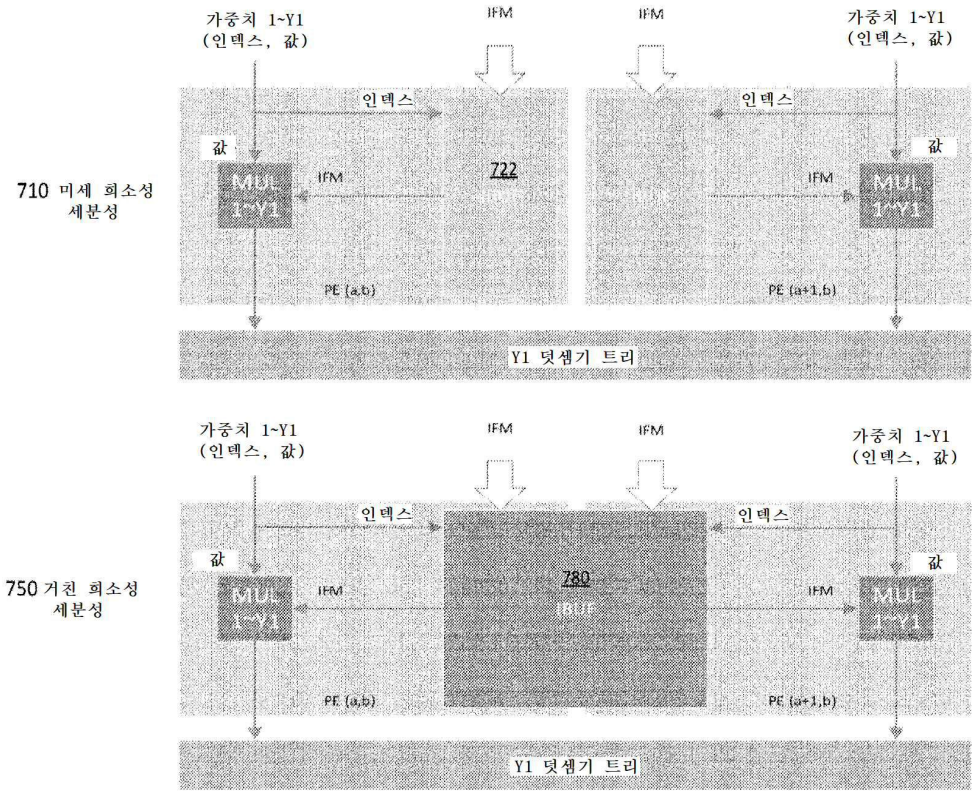
도면6a


C=1 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
C=2 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
C=3 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
C=4 및 K=1-Y1 에 대해 Y1 * 가중치 (색인, 값)
C=5 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
C=6 및 K=1-Y1 에 대해 Y1 * 가중치 (색인, 값)
C=7 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
C=8 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
C=9 및 K=1-Y1 에 대해 Y1 * 가중치 (인덱스, 값)
PE 클러스터
IFM
PE (1,0)
PE (2,0)
PE (3,0)
PE (4,0)
PE (5,0)
PE (6,0)
PE (7,0)
PE (8,0)
PE (9,0)
Y1 덧셈기 트리

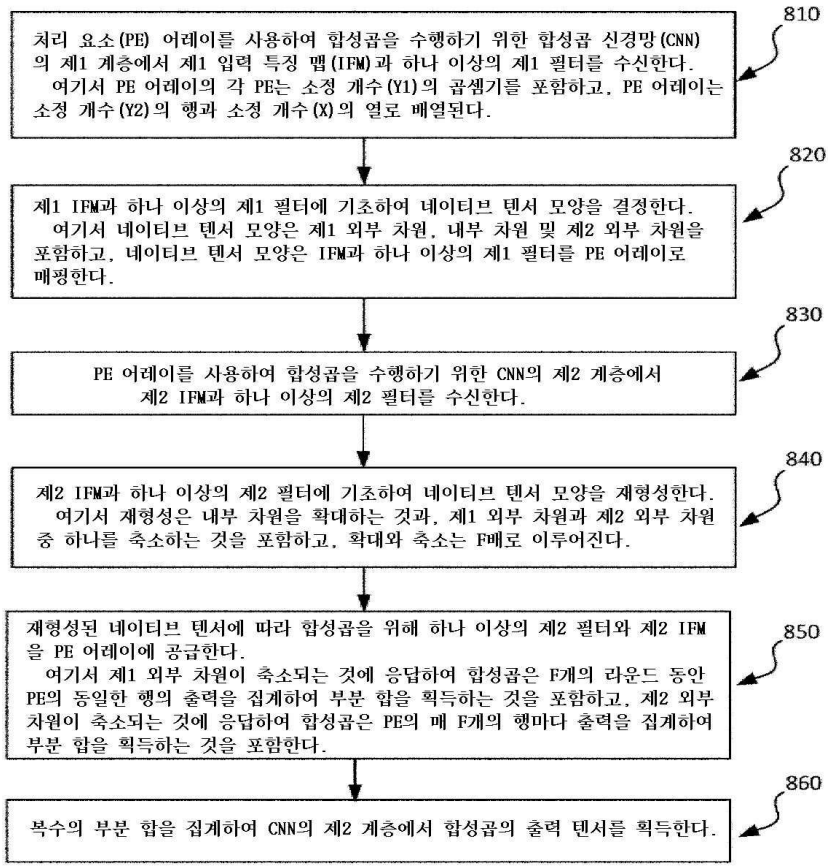
도면6b


C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (0,0) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (0,1) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (0,2) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (1,0) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (1,1) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (1,2) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (2,0) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (2,1) (인덱스, 값)
C=1 및 K=1~Y1 에 대해 Y1*가중치 3x3 커널 (2,2) (인덱스, 값)
PE 클러스터
IFM
Y1 덧셈기 트리

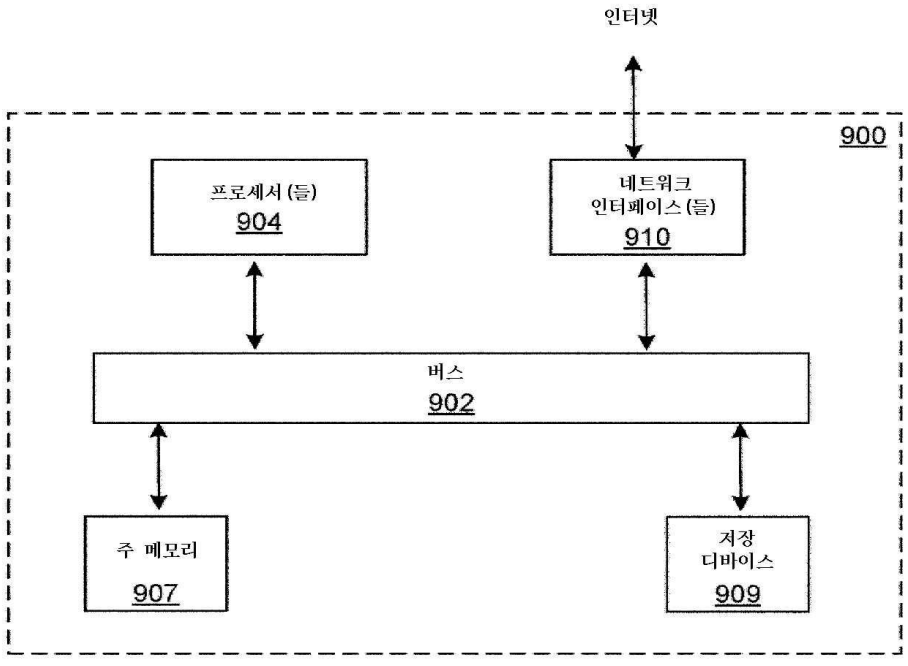
도면7


710 미세 희소성 세분성
가중치 1~Y1 (인덱스, 값)
IFM
인덱스
값
722
PE (a,b)
PE (a+1,b)
Y1 덧셈기 트리
750 거친 희소성 세분성
780

도면8


처리 요소(PE) 어레이를 사용하여 합성곱을 수행하기 위한 합성곱 신경망(CNN)의 제1 계층에서 제1 입력 특징 맵(IFM)과 하나 이상의 제1 필터를 수신한다.
여기서 PE 어레이의 각 PE는 소정 개수(Y1)의 곱셈기를 포함하고, PE 어레이는 소정 개수(Y2)의 행과 소정 개수(X)의 열로 배열된다.
810
제1 IFM과 하나 이상의 제1 필터에 기초하여 네이티브 텐서 모양을 결정한다.
여기서 네이티브 텐서 모양은 제1 외부 차원, 내부 차원 및 제2 외부 차원을 포함하고, 네이티브 텐서 모양은 IFM과 하나 이상의 제1 필터를 PE 어레이로 매핑한다.
820
PE 어레이를 사용하여 합성곱을 수행하기 위한 CNN의 제2 계층에서 제2 IFM과 하나 이상의 제2 필터를 수신한다.
830
제2 IFM과 하나 이상의 제2 필터에 기초하여 네이티브 텐서 모양을 재형성한다.
여기서 재형성은 내부 차원을 확대하는 것과, 제1 외부 차원과 제2 외부 차원 중 하나를 축소하는 것을 포함하고, 확대와 축소는 F배로 이루어진다.
840
재형성된 네이티브 텐서에 따라 합성곱을 위해 하나 이상의 제2 필터와 제2 IFM을 PE 어레이에 공급한다.
여기서 제1 외부 차원이 축소되는 것에 응답하여 합성곱은 F개의 라운드 동안 PE의 동일한 행의 출력을 집계하여 부분 합을 획득하는 것을 포함하고, 제2 외부 차원이 축소되는 것에 응답하여 합성곱은 PE의 매 F개의 행마다 출력을 집계하여 부분 합을 획득하는 것을 포함한다.
850
복수의 부분 합을 집계하여 CNN의 제2 계층에서 합성곱의 출력 텐서를 획득한다.
860

도면9


인터넷
900
프로세서(들)
904
네트워크 인터페이스(들)
910
버스
902
주 메모리
907
저장 디바이스
909